GENERALIZABILITY OF RANDOMIZED CONTROLLED TRIALS OF SUBSTANCE USE DISORDER TREATMENTS

by

Ryoko Susukida

A Dissertation submitted to Johns Hopkins University in conformity with the requirements for the Degree of Doctor of Philosophy

Baltimore, Maryland

May 2017

ABSTRACT

Randomized controlled trials (RCTs) are widely considered the gold standard to assess the effectiveness of new treatments. Decisions for clinical guidelines and health policies are often made based on findings of RCTs. While study designs of RCTs can mitigate threats to internal validity of the estimated treatment effectiveness, they do not assure external validity, which is how well findings from one particular sample can be applied to the target population of individuals for whom a treatment is intended. There is growing concern in the recent literature that the findings from RCTs may not be directly applicable to real world settings. Particularly in the context of RCTs of treatments for substance use disorders (SUD), there is a growing body of literature showing that strict eligibility criteria commonly used in RCTs of SUD treatment would exclude substantial proportions of individuals from the target population, which may adversely impact generalizability of the findings from SUD RCTs. However, very few past studies have assessed generalizability of findings of actual SUD RCTs to the intended target populations. The purpose of this dissertation was to assess generalizability of findings of SUD RCTs that were implemented in various settings, as compared with differently defined target populations. In Chapter 1, we provided an overview of the existing literature and described the data source and methodology used in this dissertation. In Chapter 2, we assessed generalizability of the findings from ten multi-site SUD RCTs to each target population of patients seeking SUD treatment in usual treatment settings in the United States. We weighted the RCT sample treatment effects on three outcomes, on retention, urine toxicology, and abstinence to make the RCT samples resemble the target populations, by using propensity scores representing the conditional probability of participating in RCTs. We found that weighting the samples changed the significance of estimated sample treatment effects. Most commonly, positive treatment effects of

RCTs became statistically insignificant after weighting. In Chapter 3, we assessed generalizability of the treatment effects on retention and abstinence from a multi-site web-based SUD intervention to two types of target populations: SUD treatment-seeking individuals and community-dwelling individuals with recent substance use, whether or not they sought treatment. The population effect on abstinence became insignificant after weighting the data by the generalizability weights of both target populations. In Chapter 4, we conducted a meta-analysis of generalized treatment effects on retention and abstinence from four RCTs of cocaine dependence treatments to the same two types of target population used in the previous chapter. We also conducted a network meta-analysis to examine comparative treatment efficacies across these four treatments while taking into account the generalizability of the findings. We found that the overall generalized treatment effect on retention was significantly larger than the unweighted effect. We also found that weighting changed the ranking of the effectiveness across treatments. Lastly, in Chapter 5, we provided a summary of the findings and discussed public health implications in light of strengths and limitations of this dissertation.

Thesis Advisory Committee:

Tamar Mendelson, PhD (Academic Advisor)

Rosa M. Crum, MD, MHS (Chair)

Ramin Mojtabai, MD, PhD, MPH

Holly C. Wilcox, PhD

Alternates:

Trang Q. Nguyen, PhD

Daniel O. Scharfstein, ScD

ACKNOWLEDGEMENTS

I have received excellent training and mentorship during my time at Johns Hopkins University. To my academic advisor, Dr. Tamar Mendelson, thank you for your great support and patient guidance throughout the PhD program. You taught me how to ask creative research questions and encouraged me to think about ideas with different perspectives. Most importantly, you have been a great role model of how to become a leading female researcher in this competitive academia. To my dissertation advisor, Dr. Ramin Mojtabai, thank you for your unconditional support throughout my dissertation writing process. You challenged me intellectually (always with a smile on your face) and pushed me to become a better researcher. I look forward to continuing to grow as a professional researcher and hope to become more like you who inspires others. To my other committee members, Dr. Holly Wilcox and Dr. Rosa Crum, thank you for being great mentors, role models and collaborators on various projects throughout my PhD program. To Dr. Elizabeth Stuart, it has been such a pleasure to collaborate with you in the past few years. Thank you for always being there for me when I encountered difficulties and being supportive in person as well as through emails. Last, but not least, I would like to thank my family for their unwavering support. To my parents, Koichi and Nobuko, thank you for supporting my decision to study overseas and always believing in me accomplishing this no matter how challenging it was. To my brother, Makoto, thank you for always making me laugh even during difficult times and making this entire PhD program more enjoyable. Lastly, to the love of my life, my best friend, and my great advocate, David, thank you for always being my “swim buddy” and your unconditional support throughout this process. I certainly could not have done this without you. Finally, I have been very fortunate to receive funding throughout this PhD program including Fulbright Japan Scholarship, Lucy Shum Memorial Scholarship Award, and the funding from the National Institute on Drug Abuse (NIDA) (R01 DA036520, PI: Dr. Ramin Mojtabai).

Table of Contents

LIST OF TABLES

Chapter 2

Chapter 3

Chapter 4

Appendices

LIST OF FIGURES

Chapter 2

Appendices

CHAPTER 1 Introduction

1.1. Background

Randomized controlled trials (RCTs) are generally considered the most reliable study design for evaluating the effectiveness of new treatments and interventions. Directions for treatment guidelines and health policies are often made based on findings of RCTs. Moreover, RCTs provide the most reliable causal inferences of the effects of new treatments and interventions by mitigating threats to internal validity. However, the study design of RCTs does not assure external validity, which is how well findings from one particular sample can be applied to the target population of individuals for whom a treatment or an intervention is intended.

There is growing concern in the recent literature that the findings from RCTs may not be directly applicable to real world settings.[1–6] The findings of an intervention with strong treatment effects in one particular setting often cannot be replicated or produce smaller effects in different settings.[7,8] Particularly in the context of RCTs of treatments for substance use disorders (SUD), there is a growing body of study findings showing that the characteristics of RCT samples differ substantially from those of target populations.[2,9,10] It has been found that women, especially pregnant women, African-Americans, individuals with lower income, and individuals with more severe substance use or psychiatric problems are under-represented in the RCTs for SUD treatments.[9,10] Additionally, it has been found that eligibility criteria commonly used in RCTs of SUD treatment exclude substantial proportions of individuals from the target population. According to Humphrey et al.,[9] 20% to 33% of individuals with alcohol use disorders would be excluded by the commonly used eligibility criteria in RCTs of alcohol use disorders. Another study by Okuda et al.[2] found that 80% of individuals with cannabis dependence would be excluded by the commonly used

eligibility criteria for cannabis treatment RCTs. Moreover, a recent review study by Moberg and Humphreys[11] estimated that commonly used exclusion criteria in SUD trials would exclude between 64% and 95% of potential participants. A more recent study by Susukida et al.[12] found that the participants of the SUD RCTs were more likely to have higher level of educational attainment and have full time jobs as compared with the individuals seeking SUD treatment in usual treatment setting.

While using less stringent eligibility criteria to improve the representativeness of RCTs may be a straightforward solution, concerns for non-adherence with treatment, and patient safety often prevent researchers from expanding eligibility criteria of RCTs. Furthermore, there is some evidence that the eligibility criteria of SUD treatment RCTs have become more stringent over the recent years.[13] Particularly, SUD RCTs funded by government tend to use such restrictive eligibility criteria.[13] In order to assess generalizability of the findings of RCTs, it is important to examine how representative the RCT samples are of potential target populations, and whether and how lack of representativeness might have affected the findings of RCTs.[14]

While most previous studies have examined what proportion of a putative target population would be hypothetically excluded from RCTs, very few studies with a few recent exceptions have compared the characteristics of actual RCT participants and the target populations to assess representativeness of the RCT samples. Also, few studies have examined whether and how representativeness of the RCT sample may affect the findings of the RCTs when generalized to a target population. Furthermore, there is little understanding of how generalizability of the findings of RCTs differ depending on the definitions of the target populations. For instance, the target population for the RCT for alcohol use disorder could be defined as individuals who are seeking treatment for alcohol

use disorder or could be defined as individuals with alcohol use disorder regardless of treatment seeking status. Particularly in the context of SUD, there is a large proportion of individuals with SUD who do not receive treatment despite their treatment needs. Hence, it is important to assess whether treatment effects estimated through SUD RCTs can be applicable to those with SUD regardless of treatment seeking status. Finally, there is increasing interest in assessing comparative treatment efficacies by comparing different treatments from different RCTs; however, no previous studies assessed comparative treatment efficacies taking into account generalizability of the RCTs to the target populations. This dissertation aimed to address these issues by using the data from SUD RCTs that were actually implemented in various settings.

1.2. Significance

Concerns for limited generalizability of the findings from RCTs

RCTs are widely considered the gold standard to assess efficacy of new interventions since the first introduction of this study design in the early 20th century. RCTs provide confidence that the estimated treatment effects are actually caused by new interventions. Despite its strength in assuring internal validity, the study design of RCTs does not necessarily assure external validity, which is how well findings from one particular setting can be applied to the target population for whom an intervention is intended.

In social science and medical fields, assuring generalizability of the findings from RCTs is more critical than in physical sciences where typically humans are not involved because humans react to new interventions or treatments differently based on their genetic predispositions and environmental factors including socio-economic status and cultures. Decisions regarding implementation and dissemination of new interventions and treatments should be made based on not

only the observed effects from RCTs but also the external validity of the observed effects to the intended target populations.

While many efforts to recruit RCT participants from real-world settings have been made to improve generalizability of the findings from RCTs,[15] there are certain obstacles that often make these efforts unsuccessful. For example, many people who are recruited to the RCTs decide not to participate. Those who participate in RCTs willingly may differ in terms of attitude toward treatments or interventions and socio-demographic status from those who refuse to participate in RCTs. The treatment effects estimated with only those who agreed to participate in RCTs may not be necessarily generalizable to those who refused to participate in RCTs, who may have responded to treatment differently. Refusal to participate in RCTs is particularly concerning in the context of SUD treatments because a large proportion of patients are referred to treatment through legal authorities such as criminal justice and are not seeking treatment voluntarily.

Stringent eligibility criteria of many RCTs often limits ability to include representative samples from broad target populations.[2,10,16] A review study of 41 NIH sponsored RCTs in various fields demonstrated that approximately 73% of a representative sample was excluded from these studies due to commonly used eligibility criteria.[17] In the literature on RCTs for SUD treatments, use of restrictive eligibility criteria is one of the major concerns for limited generalizability of the findings from RCTs to target populations. For instance, one study estimated that common eligibility criteria in cannabis treatment RCTs would exclude almost 80% of patients with cannabis treatment.[2]

While relaxing eligibility criteria may seem to be the most straightforward solution to make RCT samples more representative, it may not be always feasible for all RCTs, especially when there are safety concerns for patients. For example, co-existing medical conditions may make participation in RCTs of a

new treatment difficult, especially if the treatment includes use of novel medications which could interact with the medications that the potential RCT participant is already taking. In such cases where relaxing eligibility criteria is challenging, it is important for researchers to have useful tools to examine to what extent the findings of RCTs are applicable to target populations. This dissertation applies a novel weighting-based method to various SUD RCTs to assess the sample representativeness and the generalizability of the RCT findings to intended target populations.

Limited sample representativeness of SUD RCTs

Many RCTs of SUD treatments tend to exclude those with co-occurring medical and psychiatric disorders.[11] While some eligibility criteria are reasonable to ensure safety of RCT participants, some criteria are used merely for logistic convenience such as excluding those without stable housing, and some criteria are not based on a clear rationale.[18] There is a growing interest in whether and how the eligibility criteria for SUD RCTs impact the sample representativeness and external validity of the findings from RCTs.[11]

A review by Humphreys et al.[13] identified 14 eligibility criteria that are most commonly used in alcohol treatment research (683 studies): alcohol problems (39.1%), psychiatric problems (37.8%), prior alcohol treatment (31.8%), medical conditions (31.6%), compliance/motivation (31.5%), demographic (26.2%), neurocognitive problems (23.0%), illicit drug use (22.7%), social instability (14.9%), distance from treatment (10.1%), residential stability (8.6%), education/literacy (4.4%), legal problems (3.5%), and financial situation (1.3%).

Blanco et al.[19] found that the set of criteria identified by Humphreys et al.[13] excluded 50.5% of representative individuals with alcohol dependence in the US

and 79.4% of those who actually sought treatment for alcohol dependence, by using the National Epidemiologic Survey on Alcohol and Related Conditions (NESARC) data. Using the NESARC data of those with alcohol dependence, Hoertel et al.[20] found that 64.3% of those with co-occurring mood disorder were excluded with Humphreys-identified criteria.[13] Similarly, Storbjork[21] found that at least one of the 14 criteria by Humphreys et al.[13] excluded 96% of representative individuals with alcohol dependence who were seeking treatment in Stockholm County, Sweden. Okuda et al.[22] conducted a similar analysis in the context of cannabis dependence treatment. They found that as many as 80% of community-dwelling individuals with cannabis dependence in the NESARC data were excluded by Humphreys-identified criteria[13] from cannabis treatment RCTs.

Velasquez et al.[23] found that 52.9% (N=317) of the 599 individuals who were screened for eligibility for a multisite RCT of alcoholism treatment (Project Match) were excluded from the RCT. Similarly, Sofuoglu et al.[24] found that 70.8% (N=608) out of the 859 individuals who were screened by telephone for eligibility to participate in the inpatient cocaine treatment study were excluded from the study. Frewen et al.[25] used the data of patients from publically funded drug and alcohol center in Sydney, Australia, and found that of the 169 patients that were screened for eligibility, 52.1% (N=88) did not meet eligibility criteria and were excluded from RCT for cannabis treatment. A recent review[11] of these studies on impacts of eligibility criteria on the RCT sample representativeness estimated that between 64 and 96% of potential study participants can be excluded from SUD treatment RCTs with commonly used eligibility criteria.

Not only exclusion from the study but also refusal to participate in RCTs impacts the sample representativeness of the RCTs. Melberg and Humphreys[26] reviewed 98 illicit drug use treatment RCTs and found that an average of 29% of potential RCT participants were ineligible and an additional 29% of the eligible

participants refused to participate. An average of 36% of potential RCT participants were ineligible and an additional 32% of the eligible participants refused to participate when each study was weighted by sample size. The authors suggested that this indicates that RCTs with a larger number of participants do not necessarily include more representative samples.

While past studies assessed the hypothetical impact of eligibility criteria on the RCT sample representative, very few studies with some recent exceptions assessed the sample representativeness of the RCTs as compared with the intended target populations. A study by Susukida et al.[12] 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 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. Another recent study by Blanco et al.[27] also directly compared the RCT sample of the web-based intervention with the target population of individuals with SUD drawn from the Wave 1 of the NESARC. They found that there were substantial differences between the RCT sample and the target population in terms of race, educational attainment, marital status, and types of primary substance use problems.

Limited generalizability of SUD RCTs

Very few studies have examined the impact of sample representativeness on RCT findings. Humphreys et al.[14] assessed how applications of commonly used eligibility criteria impacted the outcomes of patients by using the data of real-world SUD treatment-seeking patients whose outcomes were known. They compared the outcomes between the samples with and without the application of

five widely used treatment research eligibility criteria, which are psychiatric problems, medical problems, social-residential instability, low motivation/noncompliance, and drug problems. It was shown that while eligibility criteria of psychiatric and medical problems created only a moderate bias (10% or less change) in outcome estimates, eligibility criteria of social-residential instability, low motivation/noncompliance, and drug use created a larger (up to a 51% change) bias in outcome estimates. More recently, Blanco et al.[27] applied a weighting-based method similar to the one used in this dissertation to the sample of a web-based SUD RCT and reweighted the outcomes of the RCT to the target population drawn from the NESARC data. They found that reweighting the RCT sample with the target population weight made the significant treatment effect of the web-based SUD RCT statistically insignificant.

Generalizing the findings of RCTs to target populations

Stuart and colleagues[28] proposed a statistical method to assess the sample representativeness of RCT samples using propensity score techniques. This method is to compute conditional probability, p, (similar to a propensity score) for being included into the RCT sample based on a number of covariates which are commonly observable in both the RCT sample and the target population. The difference in average propensity scores (Δp) between the RCT sample and its target population indicates how similar or different the distributions of characteristics between RCT sample and the target population are. Larger values of Δp indicate that the RCT sample and the target population tend to differ from each other, while smaller values of Δp indicate that the RCT sample and the target population tend to share more similarities. Stuart and colleagues[28] applied this method to a school-based RCT intervention, and they estimated a standardized Δp of 0.73. This suggests “substantial difference” between the two

samples. In observational studies, it is generally considered that values of Δp larger than 0.25 indicate substantial difference between two samples.[29–31] Mamdani et al.[32] suggested more conservative cutoff of Δp larger than 0.10 as indicating a meaningful difference between two samples.

The estimated propensity score can be used to generalize the findings from RCTs to intended target populations. Cole and Stuart[33] applied this method to generalize the results from a AIDS Clinical Trial Group (ACTG) study using data from a representative target population of a HIV-infected individuals in the US. The authors computed the inverse probability of selection weight, $(1-p)/p$, where p was the estimated conditional propensity score based on characteristics of study participants in the ACTG trial and the target population. The authors found that men, White patients, Hispanic patients and patients older than 30 years old had higher probability of being included in the RCT sample and these characteristics also moderated the treatment effect in the ACTG trial. This dissertation applies this propensity-score based method in the context of the SUD RCTs to assess the sample representativeness and the generalizability of the findings from multiple SUD RCTs.

1.3. Overview of specific aims

The study has the following specific aims and hypotheses:

Aim 1: To assess generalizability of the outcomes from ten SUD RCTs to target populations in usual treatment settings.

Chapter 2 covered Aim 1 and compared RCT sample treatment effects and the population effects of SUD treatment. The population effects were estimated through statistical weighting, which re-computes the effects in RCTs such that the participants in the RCTs had similar characteristics to individuals in the target populations. Chapter 2 used multi-site ten RCTs (five trials of

Buprenorphine/Naloxone detoxification for opioid dependence, three trials of motivational enhancement/interviewing on SUD, and two trials of motivational incentives for cocaine, methamphetamine or amphetamine use) drawn from National Institute of Drug Abuse (NIDA) Clinical Trials Network (CTN) and the target population of individuals seeking treatment in usual SUD treatment settings drawn from the Treatment Episodes Data Set-Admissions (TEDS-A). A total of 3,592 patients in ten RCTs and 1,602,226 patients from usual SUD treatment settings between 2001 and 2009 were included in the analyses of Chapter 2. Generalizability of treatment effects on three types of outcomes were examined: retention, urine toxicology, and abstinence. The RCT sample treatment effects were weighted to resemble target populations with propensity scores representing the conditional probability of participating in RCTs.

Aim 2: To assess generalizability of the outcomes from a web-based SUD intervention (Therapeutic Education System) to target populations in usual treatment settings as well as in community-dwelling settings.

Chapter 3 covered Aim 2 and assessed the generalizability of the findings from a multi-site web-based SUD intervention. We compared the sample of a web-based SUD intervention (Therapeutic Education System vs. Treatment-as-usual) (n=507) with two types of target populations: SUD treatment-seeking individuals drawn from the Treatment Episodes Data Set-Admissions (TEDS-A) and community-dwelling individuals with recent substance use, whether or not they sought treatment, drawn from the National Survey on Drug Use and Health (NSDUH). With propensity scores of RCT participation, we weighted the treatment effects on retention and abstinence to make the trial sample resemble these target populations.

Aim 3: To compare the effectiveness of four different pharmacological treatments for cocaine use disorders while taking into account generalizability of the treatment effects estimated through RCTs to two types of target populations in usual treatment settings as well as in community-dwelling settings.

Chapter 4 covered Aim 3 and conducted a meta-analysis to synthesize the treatment effectiveness of multiple medications for cocaine dependence and to assess comparative treatment effectiveness while incorporating the generalizability of the RCT findings to the target populations. We drew Individual-level data from four RCTs (Reserpine, Modafinil, Buspirone, and Ondansetron vs. placebo) from the National Institute of Drug Abuse Clinical Trials Network (n=456). The treatment effects on retention and abstinence from each RCT were weighted to make the distribution of the characteristics of the RCT sample similar to those of target population of treatment-seeking patients (Treatment Episodes Data Set-Admissions; TEDS-A) as well as target population of individuals with cocaine dependence in the general population (National Survey on Drug Use and Health; NSDUH). We used a one-step meta-analytic approach to synthesize the generalized outcomes from four RCTs using individual-level data. We also conducted a network meta-analysis to assess comparative effectiveness across these four treatments with study-level data while accounting for the generalizability of the RCT findings.

1.4. Public health significance

The findings from this dissertation will provide insight into differences between participants of SUD treatment RCTs and target patient populations in various settings based on direct comparisons of these groups. The results of this dissertation will also provide a better understanding of whether and how the

differences in the characteristics between the RCT samples and the target populations can influence the findings of the RCTs. Particularly, for interventions with potential scalability to large target populations like web-based SUD interventions,[34] the findings of this dissertation will have implications for careful consideration of the representativeness of the RCT sample with regard to target population of potential users of these types of intervention. The findings from this dissertation will also have implications for other trial networks, such as the National Cancer Institute (NCI) Clinical Trial Network Program, which intends to disseminate treatments on a large scale.[35] As attention to large-scale dissemination and implementation of evidence-based treatments and interventions increases,[36] it becomes increasingly important to understand the applicability of the findings of RCTs in different populations with varying characteristics, contexts, and locations. As the movement towards “practical clinical trials” to assess treatment effect in real-world settings increases, a growing number of RCTs with less stringent eligibility have been conducted[37]. However, relaxing eligibility criteria may not be always feasible especially when there are safety concerns for patients such as allergic reactions to certain medications. In these cases, the weighting-based method that this dissertation employs might provide useful solutions to examine to what extent the findings of RCTs are applicable to target populations.

CHAPTER 2 Generalizability of Findings from Randomized Controlled Trials: Application to the National Institute of Drug Abuse Clinical Trials Network

2.1. Abstract

Aims: To compare randomized trial (RCT) sample treatment effects and 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 United States.

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.

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 RCTs 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.

2.2. Introduction

There is growing concern that the results from randomized controlled trials (RCTs) may not generalize to real world settings.[2–6,38] 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 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.[1,2,39,40] 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.[9,40] 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.[9] 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[11] estimated

that commonly used exclusion criteria in SUD trials would exclude between 64% and 95% of potential participants.

A study by Susukida et al.[12] 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 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.

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.[13] Government-funded SUD treatment trials are particularly likely to use such restrictive exclusion criteria.[13] 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 for evaluating the real-world relevance of RCTs.[14] While previous studies have examined how well RCT samples represent target populations,[2,9,12,40] few studies have assessed how representativeness of the RCT sample may affect the findings of the RCTs when generalized to a target population.[28] 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.[41]

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,[28,33] and is similar to inverse probability weighting for non-experimental studies.[42] This method was used by Stuart et al.[41] 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.[12], 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 would 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.

2.3. Methods

Data source

The RCTs used in this study were the same RCTs used in our prior analyses.[12] 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.[43] For each RCT sample, we drew a

separate 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 sets 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).[44] 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.[12] describes the definitions of the target populations for each RCT.

Table 2.1. describes characteristics of each CTN trial. Five trials (CTN0001[45], CTN0002[45], CTN0003[46], CTN0010[47], CTN0030[48]) examined the effectiveness of Buprenorphine/Naloxone detoxification (Bup/Nx-Detox) for opioid

dependence. Three trials (CTN0004[49], CTN0005[50], CTN0013[51]) examined the effectiveness of motivational enhancement/interviewing (MEI) on SUD, and two trials (CTN0006[52], CTN0007[53]) 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.

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 study

similar with respect to observed characteristics.[42] In this study, we weighed both arms of the RCT samples to resemble the target populations.[28,33] 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.[12] 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"[54] package in R,[55] was used to calculate the propensity scores for each patient.[56,57] 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.[58]

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.

2.4. Results

Comparison of unweighted outcomes and outcomes weighted by propensity scores

Table 2.2. 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 2.3. 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 2.4. 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[12] indicated, the composition of the CTN samples deviated significantly from the composition of the target populations with regard to the socio-demographic characteristics on which these samples were compared. Appendix Table 2.1. 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 Appendix Table 2.2.).

Results of subgroup analysis of treatment effects are presented in Appendix Table 2.3. 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.

2.5. 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.

A recent study by Stuart et al.[41] 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;[2,9,12,40] however, those studies did not attempt to estimate the population effects from trial results.

Our study findings have implications for the external validity of results from SUD RCTs. Susukida et al.[12] 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.,[12] 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.[29] 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 do not include 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. Results may differ if future studies use broader 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.[45–53] For example, some original outcome studies published by primary investigators reported treatment effects by trial sites;[50] 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 the 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 [37]. 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,[36] 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,[59] which may affect profiles of patient groups who seek and access treatments.

Table 2.1. 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

Table 2.2. 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
CTN30					
Unweighted	0.91	0.67	1.24	.55	
Weighted	0.95	0.30	2.99	.93	F(1, 1305)=.00, p=.95

Table 2.3. 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
CTN30					
Unweighted	0.72	0.41	1.25	.24	
Weighted	1.38	0.36	5.21	.64	F(1, 1305)=0.00, p=.97

Table 2.4. 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	
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

CHAPTER 3 Generalizability of the Findings from a Randomized Controlled Trial of a Web-based Substance Use Disorder Intervention

3.1. Abstract

Background: There is a growing concern for generalizability of the findings from randomized controlled trials (RCT) of interventions for substance use disorders (SUD).

Objectives: This study assessed the generalizability of the findings from a multi-site web-based SUD intervention.

Methods: The sample of a web-based SUD intervention (Therapeutic Education System vs. Treatment-as-usual) (n=507) was compared with the characteristics of the two types of target populations: SUD treatment-seeking individuals from the Treatment Episodes Data Set-Admissions (TEDS-A) and community-dwelling individuals with recent substance use, whether or not they sought treatment, from the National Survey on Drug Use and Health (NSDUH). Using propensity scores of RCT participation, we weighted the treatment effects on retention and abstinence to make the trial sample resemble these target populations.

Results: Substantial differences between the RCT sample and the target populations were reflected in significant differences in the mean propensity scores (1.62 and 1.14 standard deviations for the TEDS-A and NSDUH, respectively, at $P < 0.001$). The population effect on abstinence (12 weeks and 6 months) was insignificant after weighting the data by TEDS-A and NSDUH generalizability weights. There was no significant difference between the population effect and unweighted effect on retention. Suggestive evidence of treatment effect heterogeneity was found across subgroups, some of which were consistent with the differences between weighted and unweighted treatment effects.

Conclusions: Generalizability of the findings of the RCT appears to be limited when the RCT sample does not well-represent the target populations and there is treatment effect heterogeneity across subgroups of RCT participants.

3.2. Introduction

Substance use disorders (SUD) impose significant societal and economic burdens. Tobacco use is the leading cause and alcohol use the third leading cause of preventable deaths in the United States.[60] Overdose death rates of illicit drugs as well as prescription drugs are steadily increasing over time. As of 2015, 52,404 individuals died of drug overdose, which is almost double the number of overdose deaths in 2004.[61] Annual costs of lost work productivity, crime and health care associated with abuse of tobacco, alcohol and illicit drugs is estimated to be more than $700 billion in the United States, which is more than 4.1% of the annual GDP.[62–64]

In spite of high prevalence of SUD and its various negative health consequences,[65] many individuals with SUD do not receive treatment.[66–68] Using the Wave 2 data of the National Epidemiologic Survey on Alcohol and Related Conditions, Blanco et al.[68] found 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. Although there are a number of effective evidence-based interventions for SUD,[69] strong stigma toward these conditions[70] and the limited access to SUD specialty treatment[71] 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.[34] A web-based SUD treatment can offer various potential benefits including lower implementation cost, greater scalability, greater accessibility in remote or rural areas with 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.[34,72] A growing number of randomized controlled trials (RCTs) demonstrated that web-based SUD interventions had higher treatment retention and resulted in increased level motivation to change, decreased substance use, and greater knowledge about SUD as compared with treatment-as-usual.[34]

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. 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. A recent review by Moberg and Humphreys[11] showed that commonly used exclusion criteria in RCTs of SUD treatments would exclude between 64% and 95% of potential participants. In addition to exclusion criteria, refusal to participate in the RCTs is another critical factor that may impact the representativeness of the RCT samples. In the context of RCTs of SUD treatments, refusal to participate in the RCTs is especially concerning because many individuals with SUD do not receive treatment voluntarily and are referred to treatment through the criminal justice system. A study by Susukida et al.[12] directly compared characteristics between actual participants in ten SUD RCTs and 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 target populations.

Another study by Susukida et al.[73] examined how lack of representativeness of the SUD RCT samples affected the findings of the RCTs.

Susukida et al.[73] used statistical weighting techniques to make the SUD RCT samples resemble the target populations and showed that significant sample treatment effects often became insignificant after weighting. This same study also demonstrated suggestive evidence of treatment effect heterogeneity across under- or over-represented subgroups of the RCT participants, some of which could potentially explain why weighted and unweighted treatment effects were different.

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 participants of RCTs 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.[27] directly compared the RCT sample of the web-based intervention, the Therapeutic Education System (TES), with two types of target populations: SUD treatment seeking population and those 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), which was conducted between 2001 and 2002. They found substantial differences in characteristics between the RCT sample and the two target populations and also demonstrated that the significant treatment effect of TES became insignificant after the sample was statistically 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.[27] may not appropriately represent the current population with SUD. According to the report by the National Institute of Drug Abuse (NIDA),[74] approximately 9.4 percent of the US population used an illicit drug during the past month in 2013 while the rate was 8.3 percent in 2002.[74] Given these

changes in nationwide trends in substance use behaviors in recent years, it is important assess the representativeness of the SUD RCT trials compared with the target populations drawn from more recent years in the US. Furthermore, in Blanco et al.,[27] both the treatment seeking target population and the general target population were drawn from the NESARC, which was a general household population survey and did not capture marginalized population groups such as homeless and/or incarcerated individuals.

The aims of this study were (1) to assess the sample representativeness of a large multi-site web-based SUD intervention RCT with two target populations: individuals with recent drug use who are admitted into SUD treatment in usual care settings, and individuals with recent drug use in the general population regardless of treatment seeking status, and (2) to estimate sample treatment effects and the population effects of the web-based intervention. Generalizing treatment effects to these two diverse target samples addresses two distinct policy questions: 1) the efficacy of the treatment for individuals who seek treatment in usual care settings and 2) the efficacy of treatment if treatment is disseminated to the much wider population group who are not currently seeking any treatment, but could potentially benefit from it. Unlike a recent study by Blanco et al.,[27] this study drew target populations from data of more recent years. Moreover, while one of the target populations in Blanco et al.[27] consisted of a group of individuals self-selected into SUD treatment, the target population of those receiving SUD treatment in this study was not entirely self-selected into treatment because it included patients referred to treatment through legal authorities such as criminal justice.

We first conducted a pairwise comparison of characteristics of the participants of the web-based SUD RCT and the two target populations. Next, we summarized differences between the RCT and the target populations by

computing propensity scores which represent the probabilities of participating in the RCT based on the eight commonly observed characteristics between the web-based SUD RCT and two target populations. We then used the estimated propensity scores to weight the RCT sample to resemble the target populations,[28,33] which is a similar approach to inverse probability weighting for non-experimental studies.[42] Finally, subgroup analyses were conducted to examine the treatment effect heterogeneity by under- or over-represented subgroups of RCT participants to help explain the differences between weighted and unweighted models. The findings have implications for assessing the representativeness of the samples and generalizability of the results of web-based interventions for SUD.

3.3. Methods

Data sources

Data for the "Web-delivery of Evidence-Based, Psychosocial Treatment for Substance Use Disorders" were drawn from the National Institute of Drug Abuse Clinical Trials Network (NIDA) Clinical Trials Network (CTN) Data share Website.[43] This trial sought to examine the effectiveness of an intervention called Therapeutic Education System (TES) plus motivational incentives for treatment of substance disorders.[75] The TES is a computerized psychosocial intervention, which includes skill building modules and incentives that are provided upon completion of the modules and abstinence from substance use.[76] The trial recruited study participants between June 2010 and August 2011 and randomized individuals seeking outpatient treatment for SUD into two arms. One arm received treatment as usual (TAU), comprised of standard SUD outpatient treatment and the other arm TAU plus TES. The trial lasted 12 weeks and enrolled adults in outpatient SUD treatment who reported any illicit drug use in

the past 30 days. A total of 507 patients were included in this study. Illicit drugs included cocaine, opiates (morphine, codeine, and heroin), amphetamines, cannabinoids (THC), methamphetamines, benzodiazepines, oxycodone, methadone, barbiturates, and MDMA.

The RCT sample was compared to two target populations. The first target population was 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 old or older to substance abuse treatment facilities across the US that are publically funded. 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 to make the sample comparable to the RCT sample. Since only a small portion of the TEDS-A was missing (n = 14,712, 2.4% out of 610,766), we conducted statistical analysis with complete cases (n = 596,054).

The second target population was drawn from the 2013 and 2014 National Survey on Drug Use and Health (NSDUH) –the most recent NSDUH data available at the time of this writing. The NSDUH is an annual cross-sectional national survey administered by the CBHSQ, SAMHSA. The NSDUH interviews household residents aged 12 years old or older randomly drawn from the fifty US states and Washington D.C. and collects information about their patterns of substance use and mental health problems. We limited the sample to those who were 18 years or older, and reported illicit drug use in the past 30 days. There were no missing data in the NSDUH and a total of 5,717 NSDUH participants were included in this study. Since some demographic were oversampled in the

NSDUH surveys (e.g., young adults), the sampling weight was taken into account in all the statistical analyses in this study.

Measures

There were eight common variables assessed in the RCT sample and the target populations: 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 populations: 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), 3-month follow-up, and 6-month follow-up. Study participants were considered "abstinent" if they submitted negative urine toxicology sample and they reported no drug use or heavy alcohol drinking in the last 4 days of each assessment point.

Statistical analysis

We first compared the RCT sample with the two target populations 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 populations.

Next, we estimated propensity scores, which is the conditional probability of participating in the RCT based on the eight variables, for every individual both for the RCT sample and the target populations. The propensity score was estimated

using the non-parametric random forests method[56,57] as implemented in the R package 'randomForest'.[54] The random forests method has been shown to have a higher predictive accuracy than parametric methods.[77] Another advantage of the random forests method is that it reduces misclassification errors through bootstrap resampling.[78] The bootstrap resampling method draws the same number of observations from the larger group to match the number of the observations from the smaller group in the data. In the case of this study, for example, the number of observations of the RCT sample was substantially smaller than that of the TEDS-A target population and the NSDUH sample. Especially when estimating propensity score in such "class-imbalanced data", this down-sampling method performs well to decrease misclassification error.[79]

After estimating the propensity score for each individual, we calculated a difference of mean propensity scores between the RCT sample as a group and the target populations. This mean propensity score difference, Δp, was introduced by Stuart et al.[80] as a measure to evaluate the representativeness of the RCT sample as compared with the target population. Standardized Δp, which is Δp 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, it is generally considered that values of Δp larger than 0.25 indicate substantial difference between two samples.[29–31] Mamdani et al.[32] suggested more conservative cutoff of Δp larger than 0.10 as indicating a meaningful difference between two samples.

We calculated weight for trial participation as $(1-p)/p$ for each individual, with which weighted regression analyses were conducted with the STATA *pweights* command (version 13). We not only compared the statistical significance of the treatment effects from unweighted and weighted models, but

also statistically compared the effect sizes of unweighted and weighted models, with the STATA *suest* (seemingly unrelated estimation) command.[58]

Lastly, subgroup analyses were conducted to examine the treatment effect heterogeneity by subgroups of RCT participants to explore potential reasons why the treatment effects from weighted and unweighted models might differ. For instance, if the significant unweighted treatment effect of the RCT became insignificant after weighting, and the RCT had a significantly larger proportion of married participants, we may observe stronger treatment effect among married individuals than non-married individuals. We conducted stratified analyses of treatment effects by subgroups based on variables used to estimate propensity scores and performed *chi-squared* tests to compare treatment effects in different subgroups.

3.4. Results

Comparison of characteristics of RCT sample and target populations

Table 3.1. presents the characteristics of the RCT participants and the two target populations. As compared with the TEDS-A target population, the RCT sample had significantly lower proportions of Hispanic individuals, those 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 those with 12 years or longer educational attainment and those with full time jobs, as compared with the TEDS-A target population.

As compared to the NSDUH target population, the RCT sample had significantly lower proportions of those between age 18-34, those with 12 years or longer educational attainment, individuals with fulltime jobs, married individuals, and those with a history of prior SUD treatment. In contrast, the RCT sample had significantly higher proportions of Black individuals, those between

ages 35 and 49, and those with intravenous drug use as compared with the NSDUH target population.

Comparison of propensity scores of RCT sample and target populations

Table 3.2. presents the comparison of the mean propensity scores for participating in the RCT obtained from comparison of the RCT and the target population characteristics. The estimated mean propensity score for the RCT sample was significantly larger than for the target populations. The standardized Δp for both target populations was substantially larger than the 0.25 standardized Δp cut-off suggested in the literature.[29–31]

Figure 3.1. shows the density plots of propensity scores for the RCT sample and the target populations. The more limited overlap the two density plots have, the less similar the RCT sample and the target population are. The RCT sample had a smaller overlapping area with the TEDS-A target population, which is consistent with the larger Δp between the RCT sample and the TEDS-A target population. The RCT sample shared a larger overlapping area of density plots with the NSDUH target population, which is consistent with the smaller Δp between the RCT sample and the NSDUH target population.

Comparison of unweighted outcomes and outcomes weighted by propensity scores

Table 3.3. presents the unweighted and weighted treatment effects on trial retention and abstinence. We presented odds ratios (ORs) from both unweighted and weighted logistic regression analyses with the 95% confidence intervals. 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 populations. 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 both TEDS-A and NSDUH generalizability weights. Treatment effect on retention was insignificant both in the unweighted and weighted analyses. However, for both abstinence and retention outcomes, comparison of treatment effect sizes did not reveal significant differences between unweighted and weighted models with either TEDS-A or NSDUH generalizability weights.

Subgroup analysis for treatment effect heterogeneity

We presented results of subgroup analyses of treatment effects in Appendix Table 3.1. There were some consistent patterns of treatment effect heterogeneity by subgroups of RCT participants with the differences in outcomes between unweighted and weighted models. For example, for abstinence at 12 weeks, some subgroups that were overrepresented in the RCT sample (e.g, non-married individuals, those without IV drug use, and those without history of prior SUD treatment) as compared with the TEDS-A target population also showed evidence of larger treatment effects on abstinence as compared with underrepresented subgroups. As compared with the NSDUH target population, some subgroups that were overrepresented in the RCT sample (e.g, Black patients, those with less than 12 years of educational attainment, non-married individuals, and those without history of prior SUD treatment) also showed evidence of larger treatment effects on retention as compared with underrepresented subgroups. Weighting this RCT sample to resemble the distribution for the target populations decreased the weights for over-represented subsamples with larger effect sizes, which may potentially explain insignificant population treatment effect estimates.

3.5. Discussion

This study demonstrated significant differences between the RCT sample of a web-based SUD intervention, the Therapeutic Education System (TES), and target populations of potential recipients of the intervention. Whether the target population consisted of treatment-seeking individuals with recent drug use (TEDS-A) or individuals with recent drug use regardless of treatment seeking status (NSDUH), the composition of the RCT sample substantially differed from those of the target populations.

Furthermore, the summary index of these differences, Δp, far exceeded the standardized Δp cutoff 0.25[29–31] or 0.10[32] for both target populations (1.62 and 1.14 standard deviations for the TEDS-A and NSDUH, respectively), indicating substantial differences between the RCT sample and the target populations. Standardized Δp can be interpreted similarly to *Cohen's d* effect size.[81] Cohen's d values of 1.62 and 1.14 are equivalent to a 42% and 57% probability that the distributions of the RCT sample and the target population will overlap, respectively. The density plots of estimated propensity scores confirmed this by showing a relatively narrow overlapping area between the RCT sample and each target population.

This study also demonstrated that the observed promising findings of the TES intervention may not be directly applicable to potential target populations. We showed that significant treatment effects on abstinence at 12 weeks and 6-month follow-up became insignificant after weighting. We showed some suggestive evidence that these differences between the unweighted sample treatment effects and the estimated population effects could be partially explained by the treatment effect heterogeneity across under- and over-represented subgroups of RCT participants.

The findings of this study are consistent with the previous studies by Susukida et al.,[12] which found statistically significant differences between the samples from ten SUD RCTs and the corresponding target populations. The findings of this study are also consistent with Susukida et al.,[73] 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.,[27] which found the significant treatment effect of TES estimated through RCT became insignificant after weighting the sample to resemble the target populations drawn from 2000-2001 NESARC data. Our study confirms that their findings hold when the generalizability of the TES RCT was assessed with the target populations from recent years. Furthermore, unlike treatment-seeking population drawn from general population in by Blanco et al.,[27] the TEDS-A target population in this study was not entirely self-selected into treatment, which included some marginalized population such as those admitted treatment through criminal justice system.

Several limitations should be taken into consideration when interpreting this study's findings. First, only a limited number of characteristics were assessed in the RCT sample and the target populations. Therefore, the estimated propensity scores did not reflect other characteristics that may have differed between the RCT and the target populations. Second, the RCT sample that this study used came from the sample collected in the clinical (outpatient) settings. Since web-based intervention could be implemented in non-clinical settings such as school or community settings, the applicability of the findings of this study might be

limited to the context of the web-based SUD RCTs in clinical settings. Third, although the TEDS-A is one of the largest administrative data sources that cover the data on most patients with SUD in the US, some states exclude patients whose treatment is not covered by the state substance use agency funds such as Federal Block Grant funds.[44] Patients who received treatment at private hospitals are usually excluded from the TEDS-A unless they have licenses from the state substance abuse treatment agency. Therefore, the TEDS-A might have not necessarily represented the entire population of patients with SUD in the US. Fourth, the original TES RCT did not intend to examine treatment effect heterogeneity. Therefore, the subgroup analyses we conducted were not sufficiently powered and we could only show suggestive evidence of treatment effect heterogeneity across subgroups of the trial participants.

In the context of these limitations, findings from this study provide insight into differences between the RCT participants of the web-based SUD intervention and two types of target populations from recent years. The results of this study also indicate how poor sample representativeness of the RCT compared with target populations impacted the observed findings of the web-based SUD intervention. Given the great potential for scalability of web-based SUD interventions,[34] the representativeness of the sample with regard to target population of potential users of this intervention should be carefully considered.

Table 3.1. Comparison of baseline characteristics (%) of the samples in the Therapeutic Education System RCT and two target populations (TEDS-A and NSDUH)

	Therapeutic Education System RCT (n=507)	**TEDS-A (2012) Target 1 (n=596,054)**	**NSDUH (2013 - 2014) Target 2 (n=5,717)**
	Percent	**Percent**	**Percent**
Sex			
Female	37.9	36.6	38.8
Male	62.1	63.4	61.2
Race			
White	58.2	59.0	**66.4**
Black	21.9	22.3	**14.8**
Hispanic	10.8	**14.1**	12.9
Other	9.1	**4.7**	**5.9**
Age			
18-34	51.1	58.3	54.0
35-49	34.3	29.4	**23.4**
50 and over	14.6	12.3	**22.6**
Education			
<12 years	23.3	**32.2**	**14.0**
≥12 years	76.7	**67.8**	**86.0**
Employment			
Full-time	40.0	**16.4**	**53.2**
Part-time	23.1	**9.8**	**17.5**
Out of labor	10.7	**31.2**	**21.1**
Unemployed	26.2	**42.7**	**8.1**
Marital Status			
Married	14.2	15.0	**27.0**
Never married	60.8	64.7	**55.2**
Separated/divorced/widowed	25.0	**20.3**	**17.9**
IV drug use			
Yes	7.5	**19.0**	**1.0**
No	92.5	**81.0**	**99.0**
Prior treatments			
Yes	7.7	**58.3**	**18.8**
No	92.3	**41.7**	**81.2**

Notes: Pearson's χ^2 test was conducted. Numbers shown in bold type indicate statistically significant differences between RCT and target populations at $P < 0.05$. The NSDUH sample was weighted with the sampling weight .

Table 3.2. Comparison of propensity scores between the Therapeutic Education System RCT and target samples from the Treatment Episodes Data-Admission (TEDS-A) and the National Survey of Drug Use and Health (NSDUH).

Target population = TEDS-A						
CTN0044	TEDS-A	Δp^a	Pooled standard deviation	Standard Δp^b	*t*-Test	P-value
0.77	0.26	0.51	0.31	1.62	36.35	<0.001
Target population = NSDUH						
CTN0044	NSDUH	Δp^a	Pooled standard deviation	Standard Δp^b	*t*-Test	P-value
0.69	0.34	0.36	0.32	1.14	25.22	<0.001

[a] Δp is difference between propensity scores of the RCT sample and the target population.
[b] Standardized Δp is computed as Δp divided by pooled standard deviation.

Table 3.3. Comparison of unweighted (RCT sample effect) and weighted (population effect) odds ratios of treatment effect on retention and abstinence

	Unweighted			Weighted (with TEDS-A)				Weighted (with NSDUH)			
	OR	95%CI	p	OR	95%CI	p	Comparison (unweighted vs. weighted)	OR	95%CI	p	Comparison (unweighted vs. weighted)
Retention 12 weeks	0.92	0.55, 1.52	0.74	0.80	0.30, 2.18	0.67	χ^2 =0.05, p= 0.82	1.79	0.63, 5.12	0.28	χ^2 =1.26, p= 0.26
Retention 3 months	0.60	0.33, 1.08	0.09	1.09	0.35, 3.35	0.88	χ^2 =0.85, p= 0.36	0.68	0.19, 2.42	0.55	χ^2 =0.03, p= 0.87
Retention 6 months	1.02	0.72, 1.45	0.90	1.16	0.60, 2.25	0.66	χ^2 =0.11, p= 0.74	0.76	0.35, 1.64	0.48	χ^2 =0.48, p= 0.49
Abstinence 12 weeks	1.67**	1.15, 2.44	<0.01	1.62	0.80, 3.26	0.18	χ^2 =0.01, p= 0.94	2.01	0.87, 4.65	0.10	χ^2 =0.15, p= 0.70
Abstinence 3 months	1.18	0.82, 1.71	0.37	0.74	0.37, 1.48	0.39	χ^2 =1.38, p= 0.24	1.64	0.73, 3.69	0.23	χ^2 =0.53, p= 0.47
Abstinence 6 months	2.04*	1.09, 3.85	0.03	1.65	0.50, 5.49	0.41	χ^2 =0.09, p= 0.76	2.80	0.75, 10.40	0.13	χ^2 =0.18, p= 0.67

Notes:
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 in the last 4 days of the assessment.

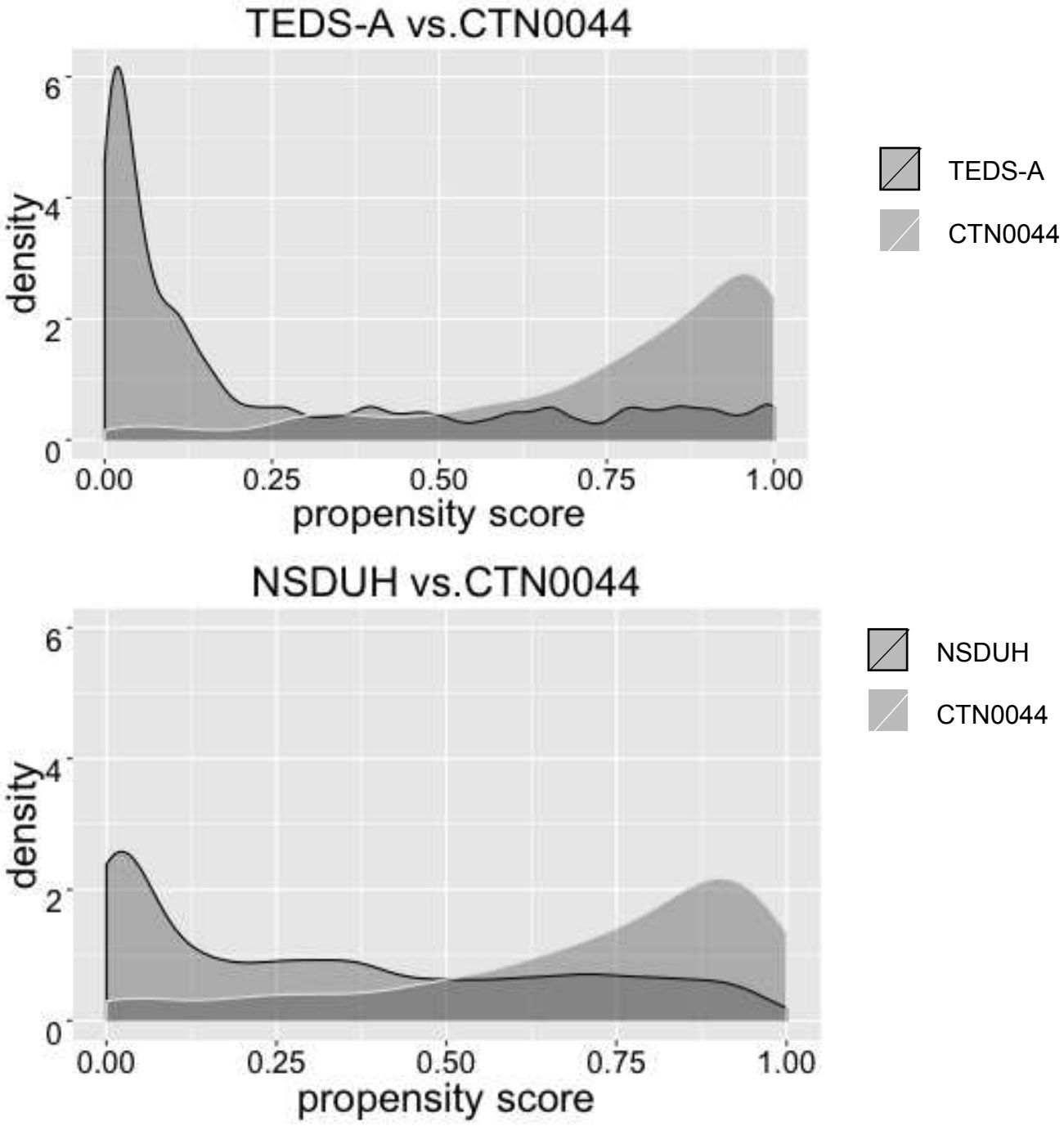


Figure 2.1. Density plots of propensity scores in CTN0044 and target samples from the Treatment Episodes Data-Admission (TEDS-A) and the National Survey of Drug Use and Health (NSDUH)

CHAPTER 4 Comparing pharmacological treatments for cocaine dependence: Addressing generalizability in meta-analysis

4.1. Abstract

Background: There are few head-to-head comparisons of cocaine dependence medications, and combining data from different studies is fraught with methodological challenges. Furthermore, randomized controlled trials (RCTs) often incorporate selective samples of patients, thus limiting generalizability of findings. We addressed these limitations by applying a novel meta-analytic approach to data on the efficacy of medications for cocaine dependence.
Methods: Individual-level data from four RCTs (Reserpine, Modafinil, Buspirone, and Ondansetron vs. placebo) were obtained from the National Institute of Drug Abuse Clinical Trials Network (n=456). The treatment effects on retention and abstinence from these trials were weighted to make the trial sample resemble treatment-seeking patients (Treatment Episodes Data Set-Admissions; TEDS-A) and individuals with cocaine dependence in the general population (National Survey on Drug Use and Health; NSDUH). We synthesized the generalized outcomes using a one-step meta-analytic approach with individual-level data and network meta-analysis with study-level data.
Results: After weighting the data by TEDS-A and NSDUH generalizability weights, the overall population effect on retention was significantly larger than the unweighted effect. However, there was no significant difference between the population effect and unweighted effect on abstinence. Weighting also changed the ranking of the effectiveness across treatments. For retention, the second most efficacious treatment, Ondansetron, became the most efficacious after applying NSDUH generalizability weights. For abstinence, the least efficacious

treatment, Reserpine, became the second most efficacious after weighting for both target populations.

Conclusions: Application of generalizability weights to meta-analysis is feasible and provides a useful tool for assessing comparative effectiveness of treatments for substance use disorders, with potential utility for comparative assessments in other fields as well.

4.2. Introduction

In 2014, the United States had approximately 913,000 individuals (0.29% of the US population) who met the Diagnostic and Statistical Manual of Mental Disorders criteria for cocaine dependence or abuse during the past 12 months.[82] According to the 2011 Drug Abuse Warning Network (DAWN) report,[83] nearly 40% of drug misuse or abuse-related emergency room visits (505,224 out of 1.3 million visits) involved cocaine. Chronic cocaine use is associated with a number of adverse outcomes including psychotic symptoms,[84] cardiovascular complications,[85] intracerebral hemorrhage,[86] and movement disorders such as Parkinson's disease.[87] It is also known that those with regular lifetime use of cocaine have significantly higher likelihood of premature deaths as compared with non-cocaine using peers.[88]

There is clearly a need for evidence-based interventions for cocaine use disorders. A number of behavioral interventions have been shown to be effective for treating cocaine use disorders including contingency management,[89] cognitive-behavioral therapy,[90] and therapeutic communities.[91] However, there are no pharmacological treatments for cocaine use disorder currently approved by the U.S. Food and Drug Administration.[92] A number of potential medications for cocaine use disorders have been examined in randomized clinical trials (RCTs).[89,90,93–95] Existing studies targeted several neurobiological agents with putative effects on receptors considered to be involved in cocaine use disorder, such as dopamine, serotonin, gamma-aminobutyric acid (GABA), glutamate, and norepinephrine.[96] The list of medications tried in past studies is long and includes the glutaminergic medication modafinil,[97] GABAergic medications such as baclofen, tiagabine, and topiramate,[94] disulfiram,[98] antidepressants such as desipramine[99] and cocaine vaccination that produces antibodies against cocaine.[100]

Despite some promising evidence of effectiveness in individual RCTs, meta-analyses of medications for cocaine use disorders have failed to produce evidence of overall treatment effectiveness of these medications or to identify clear advantages for one pharmacological agent.[93,95] These previous meta-analyses of medications for cocaine use disorders used a traditional approach of synthesizing study-level data typically obtained from publications. This approach makes it difficult to take into account the differences in the composition of RCT samples and to reliably compare different treatments.

Advances in meta-analytic methodology now make it possible to synthesize individual-level data from different RCTs.[101] Furthermore, the newly introduced method of network meta-analysis, also referred to as mixed treatment meta-analysis or multiple treatment comparison meta-analysis, now makes it possible to estimate comparative effectiveness across multiple interventions that are not evaluated against each other in any one study.[102]

Another major limitation of past RCTs for treatment of cocaine use disorder is the selective nature of the RCT samples[24] which limits the external validity, or generalizability of the findings of the RCTs to the target population. This concern is not limited to cocaine treatment RCTs as there is a growing concern that the findings from RCTs in a number of health fields may not be generalizable to real world settings.[2–6,38] However, the concerns may be amplified with regard to RCTs for treatments of substance use disorders (SUD) because of the stigma associated with such treatment and the specialized setting where treatments are offered.

There is a growing body of research showing that individuals participating in RCTs are substantially different from the target populations.[1,2,39,40] According to a recent review by Moberg and Humphreys[11] which synthesized 15 studies examining the impact of SUD trial exclusion criteria on distributions of

participants' characteristics, commonly used exclusion criteria in SUD trials would exclude between 64% and 95% of potential participants. A study by Susukida et al.[12] found substantial differences in distributions of characteristics between RCT samples and the target populations by comparing the characteristics of participants in ten RCTs from the National Institute of Drug Abuse Clinical Trials Network (NIDA-CTN) and the intended target populations consisting of people who seek treatment for SUD in usual care settings. A more recent study by Susukida et al.[73] found that 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 statistical weighting techniques. Most commonly, positive effects of trials in unweighted RCTs became statistically non-significant after weighting. To the best of our knowledge, however, no past studies of SUD treatments have synthesized data from individual RCTs with a view to the generalizability of results for the target population or have attempted to improve generalizability using statistical adjustments.

In this study, we embarked on a meta-analysis of individual-level data from four RCTs of medication used for treatment of cocaine dependence. We used the techniques of network meta-analysis to compare the effects of these four treatments while considering generalizability of the findings of the RCTs to the target populations and adjusting the results to make them more generalizable. The original RCTs had each compared the efficacy of one medication with placebo. Two target populations were selected to investigate and enhance generalizability of the findings from meta-analyses: individuals seeking treatment for cocaine dependence at usual care settings and individuals with cocaine dependence in the general population, regardless of their treatment seeking behavior. Generalizing to these two diverse target samples addresses two

distinct policy questions: 1) the efficacy of the treatment for individuals who seek treatment in usual care settings and 2) the efficacy of treatment if treatment is disseminated to the much wider population group who are not currently seeking any treatment, but could potentially benefit from it.

4.3. Methods

Data sources

RCT data were drawn from the NIDA-CTN Data share Website.[43] The NIDA CTN studies are nationwide multi-site clinical trial studies to assess the effectiveness of SUD treatments. At the time of this writing, four data sets of RCTs of cocaine dependence medications were available (CTO0001, MDS0004, CTN00052 and CTO0005). CTO0001 (n=119) examined the effectiveness of Reserpine, a dopamine depletory medication,[51] MDS0004 (n=210) examined the effectiveness of Modafinil, a non-amphetamine psychostimulant,[103] CTN00052 (n=62) examined the effectiveness of Buspirone, an anxiolytic drug,[104] and CTO0005 (n=65) examined the effectiveness of Ondansetron, a medicine mainly used for prevention of nausea.[105] All four RCTs included a placebo arm and involved adults who met the Diagnostic and Statistical Manual of Mental Disorders-Fourth Edition (DSM-IV) [106] for cocaine dependence. A total of 456 patients from four RCTs were included in this study.

We selected two different target populations for generalizability weighting of the RCT samples. The first target population was drawn from the TEDS-A in 2012 (the most recent wave of TEDS-A data available at the time of this writing). The TEDS-A is a part of the Behavioral Health Services Information System (BHSIS), which is maintained by the Center for Behavioral Health Statistics and Quality (CBHSQ), Substance Abuse and Mental Health Services Administration (SAMHSA). More than 1.5 million admissions aged 12 years old or older are

included in the TEDS-A every year. We limited the TEDS-A sample to patients with DSM-IV cocaine dependence who were 18 years old or older (n=36,997) for generalizability to the treatment-seeking population.

The second target population was drawn from the NSDUH 2013 and 2014 (the most recent waves of the survey available at the time of this writing). The NSDUH is an annual cross-sectional national survey also administered by the CBHSQ, SAMHSA. Every year NSDUH interviews a nationally representative sample of household residents 12 years old or older about their patterns of substance use and mental health problems. We limited the sample to those who met DSM-IV cocaine dependence criteria in the past year (n=235) for generalizability to the cocaine dependent individuals in the general population. The sampling weight was taken into account in every statistical analysis in this study because some demographic groups were oversampled in the NSDUH surveys (e.g., young adults).

Measures

We identified eight comparable variables between the RCTs and the target populations with which to compute statistical weights for generalizing RCT outcomes to the two target populations: sex, race-ethnicity, age, educational attainment, employment status, marital status, intravenous drug use, and the number of past treatments for SUD.

We generalized the following two outcomes from the RCTs to the target populations: successful retention in the study, and days of abstinence in the past 30 days. Successful retention in the study was defined as participation in the study until the end of the trial. Days of abstinence were calculated as the numbers of days without self-reported cocaine use in the past 30 days.

We used the same outcomes across the four RCTs to allow us to compare how statistical weighting impacts the outcomes across the studies. The original investigators of the four RCTs reported on different outcomes. For example, while the original investigators of CTN00052[104] (the Buspirone trial) used maximum days of continuous cocaine abstinence as the primary outcome, the investigators of CTO0005[105] (the Ondansetron trial) used percentage of study participants with a cocaine-free week as their primary outcome. Therefore, the observed estimates of RCT results in this study were not necessarily the same as the findings in published reports by original investigators.[103–105,107]

Statistical analysis

Analyses were conducted in four stages. First, we compared the eight characteristics noted above between the RCT samples and the target populations. Pearson's chi-squared tests were conducted to examine if there were significant differences in the distributions of the eight variables between the RCT samples and the target populations.

As the TEDS-A includes a significant amount of missing data (12.7 %), similar to our previous generalizability studies using the TEDS-A target population data,[12,73] we used multiple imputation with the STATA ice command (version 13) and created 50 imputed data sets. A detailed description of the missing data is presented in Appendix Table 4.1.

Second, to generalize the results from the RCTs to the target populations, we used a weighting-based method, which weights RCT samples to resemble the target populations.[28,33] This approach is similar to the inverse probability weighting method, which is often used for non-experimental studies.[42] Stuart et al.[41] used this approach to assess the generalizability of the findings of an RCT of behavioral intervention trial in school settings. To assess generalizability of the

findings of each RCT, we computed trial participation weights for each trial as $(1-p)/p$, where p was the propensity score, defined as the conditional probability of an individual participating in the RCTs based on the eight variables described above. The mean propensity score across 50 imputed data sets was used for TEDS-A to take into account the missing data. There were no missing data in the NSDUH sample. To calculate the propensity scores, we used a non-parametric random forests approach, using the "randomForest"[54] package in R.[55] Although the random forests approach has some advantages over a parametric approach such as a higher predictive accuracy and the ability to reduce extreme propensity scores,[77] we still encountered some outlying values for propensity scores and trial participation weights. In order to improve the performance of the propensity score based weighting, 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 al.[108] We conducted weighted regression analyses with the weights for each trial, $(1-p)/p,$ using the STATA *pweights* command (version 13).

Third, we conducted unweighted and weighted meta-analyses of four RCTs to estimate the overall treatment effect of cocaine dependence medications. Unweighted and weighted analyses were conducted for two outcomes: retention and abstinence. The unweighted analyses estimated the sample treatment effects while the weighted analyses estimated population-generalized effects. For the binary outcome of retention, we used logistic regression models; whereas, for the continuous measure of days of abstinence in the past 30 days, we used linear regression models. Baseline variables in the trial samples were not adjusted, assuming that randomization was successful in each trial. We estimated multi-level mixed effects models, which allow random intercepts and coefficient on treatment assignment variable. The standard errors were clustered

at trial level. In addition to comparing the statistical significance of the treatment effects from unweighted and weighted regression models, we also statistically compared the treatment effect sizes of unweighted and weighted models, using the STATA suest (seemingly unrelated estimation) command.[58] Furthermore, to explore potential reasons for differences in sample and generalized treatment effects, we conducted a series of subgroup analyses in which we estimated treatment effects in subgroups that were over- or under-represented in the RCT samples compared to the target populations.

Fourth, to directly compare the effects of the medications, we conducted an unweighted and weighted network meta-analysis of the four RCTs to estimate the comparative treatment effects across the four medications. Past research has shown that statistical precision of estimated treatment effects from network meta-analyses are often better than that of estimated effects from pairwise comparisons in meta-analysis.[109,110] Network meta-analysis also allows for determination of relative rankings of multiple treatments.[111] We estimated fixed-effect network meta-analysis models.

4.4. Results

Comparison of characteristics of four CTN trials and target populations

Table 4.1. presents the comparison of characteristics between four RCTs and the TEDS-A target population. Overall, the RCT samples had significantly lower proportions of women, non-Hispanic White individuals, and patients younger than 35 years old than the TEDS-A; whereas, the RCT samples had significantly higher proportions of individuals with 12 or more years of education, individuals with fulltime jobs, married individuals, and individuals who used intravenous drugs than the TEDS-A. For all the RCTs, the proportions of patients with fulltime jobs were significantly higher than the TEDS-A.

Table 4.2. presents the comparison of characteristics between four RCTs and the NSDUH target population. Overall, the RCT samples had significantly lower proportions of women, individuals from non-Hispanic White racial-ethnic groups, and those younger than 35 years old than the NSDUH; whereas, the trial samples had significantly higher proportions of married individuals than the NSDUH.

Meta-analysis

Table 4.3. presents the results of the analyses for the overall treatment effect on trial retention and abstinence. Odds ratios (ORs) and regression coefficients (βs) from both unweighted and weighted logistic and linear regression models, respectively, are presented with the 95% confidence intervals. The unweighted models estimated the overall treatment effect in the RCTs while the weighted models estimated the effects that would be expected if the distributions of characteristics in RCTs were similar to those in the target populations. In addition, the results of comparisons of regression coefficients are presented in Table 4.3.

For retention, the overall population treatment effect was significantly larger for the analyses weighted by TEDS-A as well as NSDUH than the sample treatment effect (Table 4.3). An odds ratio of 1.76 suggests that individuals receiving the active pharmaceutical agents have 76% higher odds of being retained in the follow-up compared to those treated with placebo. In contrast, there was no statistically significant difference between the weighted effect and the unweighted effect on self-reported abstinence (Table 4.3).

The results of subgroup analysis of treatment effects are presented in Appendix Table 4.2. We found some consistent patterns in the directions of change in outcomes through weighting and by subgroup analysis. For example,

the non-married individuals that were slightly underrepresented in the overall RCTs (p=0.10) showed evidence of larger treatment effects on retention than the overrepresented group of married individuals. Weighting the RCT samples to resemble the target populations increased the weights for the subsample of non-married individuals which had larger treatment effect sizes, leading to a significantly larger treatment effect on retention after weighting of the data.

Network meta-analysis

Table 4.4. presents the results of network meta-analysis comparing the effect of the four medications on cocaine dependence. Odds ratios (ORs) and regression coefficients (βs) from both unweighted and weighted logistic or linear regression models, respectively, are presented with the 95% confidence intervals. We also present the relative rankings of the treatments, computed as the probabilities of each treatment being the best among all the treatments in the network meta-analysis.

In unweighted model for each medication, there was no significant treatment effect. Although the 95% confidence intervals for the estimated treatment effects for each medication overlapped with each other, Buspirone was the most efficacious medication for retention and Modafinil was the most efficacious medication for abstinence. Weighting altered the relative ranking of the treatments. For retention, weighting by TEDS-A did not change the ranking while weighting by NSDUH made the second most efficacious treatment, Ondansetron, the most efficacious. For abstinence, the least efficacious treatment, Reserpine, became the second most efficacious after weighting for both target populations. Moreover, Ondansetron was the second most efficacious treatment for retention before weighting; however, it became the second least efficacious treatment after weighting.

Data from subgroup analyses presented in Appendix Table 4.2. suggest possible reasons for the changes in ranking of treatments. The married individuals that were overrepresented in the CTO0005 (Ondansetron trial) sample showed evidence of larger treatment effects on abstinence than overrepresented group of married individuals. Weighting the CTO0005 sample to resemble the target populations decreased the weights for subsample of married individuals with larger treatment effect sizes, leading to a smaller treatment effect on abstinence.

4.5. Discussions

This study showed that the findings from meta-analyses of cocaine dependence medications may not be directly applicable to potential target populations. The estimated overall target population-weighted treatment effect of four cocaine dependence medications on retention was significantly larger than the treatment effect from the RCTs whether the effect was generalized by the TEDS-A, representing the treatment-seeking target population of individuals with cocaine dependence, or by NSDUH, representing the target population of individuals with cocaine dependence in the community. Weighting the RCT samples to resemble target populations also altered the relative ranking of the efficacy across different medications. The results from the subgroup analysis of treatment effects partially explained these differences in effect estimates between unweighted meta-analysis and weighted meta-analyses.

A study by Stuart et al.[41] which used the same weighting-based approach to generalize the results of a school-based behavioral intervention trial found that the estimated population effect of intervention was slightly attenuated compared to the estimated sample effect from the RCT. In the context of SUD RCTs, a study by Susukida et al.[73] demonstrated a statistically significant difference

between the sample treatment effect and the population treatment effect by applying the statistical weighing-based method to the ten CTN studies, which assessed efficacy of SUD treatments. To our knowledge, the present study is the first to perform a meta-analysis and network meta-analysis of multiple SUD treatments while using a weighting approach to enhance generalizability of the findings to the target populations.

The findings from this study have implications for future meta-analyses of SUD treatments. As shown in this study, the overall treatment effect size and comparative effects changed when the deviations of each RCT sample from the target population were taken into account. Unlike the previous study by Susukida et al.[73] that found a decrease in treatment effect sizes after weighting of data from 10 SUD RCT samples, the effect size associated with the cocaine dependence medications on retention in the present study became significantly larger after weighting by population weights. This implies that the effect of weighting-based methods may vary depending on how and to what extent the composition of the RCT samples and target populations vary. Furthermore, differences in efficacies among different treatments for the same condition may be impacted by the compositions of the RCT samples for each treatment. Target population weighting of the RCTs changed the relative ranking of treatments for cocaine dependence in this study. This study also showed some suggestive evidence that the mechanisms through which the population treatment effects were different from the sample effects could be partially explained by treatment effect heterogeneity among under- or over-represented subgroups of individuals in the RCTs.

Several limitations should be taken into account when interpreting the findings of this study. First, the number of characteristics recorded in both the clinical trial samples and the target populations was relatively small, which likely

limited our ability to account for potentially important treatment effect modifiers that may have been different between the RCTs and the target populations, such as motivation for treatment and severity of SUD. Second, the weighting-based approach might not have made the distributions of the clinical trial samples sufficiently close to the distributions of the target populations to estimate the population treatment effects because of the substantial differences between the clinical trial samples and the corresponding target populations. The weighting-based method is more suitable to estimate the population treatment effect when the distributions of characteristics in RCTs overlap with those of the target populations, as they did in this study. Third, the present study could only show suggestive evidence of treatment effect heterogeneity across subgroups of individuals in the clinical trial samples because the CTN studies did not originally intend to assess the treatment effect heterogeneity and the subgroup analyses conducted here were not sufficiently powered. Fourth, there were substantial missing data in the TEDS-A which may have biased the results. Fifth, the number of trials included in this study was limited, which likely limited the reliability of the network meta-analysis [112] and our ability to conclude which medication is the most promising for treating cocaine dependence.

Limitations notwithstanding, the findings of this study provide insight into the generalizability of meta-analysis of cocaine dependence medications. The overall population weighted effect on trial retention appears promising for four cocaine dependence medications (Reserpine, Modafinil, Buspirone, and Ondansetron). The relative ranking of effectiveness among the four treatments was altered when we considered generalizability of the findings to the target populations. Modafinil appears to be the most promising treatment among these four medications, although both the sample treatment effect and the target population-weighted effects were statistically nonsignificant. With the growing

number of RCTs for cocaine dependence medications, future meta-analytic studies should assess overall treatment effects as well as comparative effectiveness while considering generalizability to target populations. The weighting-based approach used in this study is applicable to meta-analyses of clinical trials of other SUD treatments, as well as other health interventions, especially when generalizability of the findings is a concern. Although an increasing number of clinical trials for SUD treatments use less stringent inclusion and exclusion criteria, reducing concerns about generalizability of findings,[37] it may not always be possible to recruit samples that are fully representative of the target populations, e.g., when there are safety concerns for certain population subgroups. In these circumstances, the weighting-based method used in this study could be useful to assess applicability of the findings to treatment seeking target populations or to all individuals with the health condition of interest in the general community.

Table 4.1. Comparison of baseline characteristics (%) of four CTN trials and TEDS-A

	Target population	Randomized controlled trials				
	TEDS-A, 2012 N=36,997	Overall N=456	CTO0001 (1) N=119	MDS0004 (2) N=210	CTN0052 (3) N=62	CTO0005 (4) N=65
	%	%	%	%	%	%
Sex						
Female	42.3	**27.9**	**29.4**	**28.1**	37.1	**15.4**
Race						
White	41.2	**27.4**	**18.8**	**29.2**	**25.8**	38.5
Age						
<35	33.8	**19.1**	**22.7**	**14.3**	**6.5**	40.0
Education						
≥12 years	66.4	**82.8**	**84.9**	**89.5**	64.5	**81.5**
Employment						
Full-time	9.2	**33.8**	**31.1**	**32.4**	**21.0**	**55.4**
Marital Status						
Married	12.3	**23.3**	**19.5**	**26.2**	12.9	**30.8**
IV drug use						
Yes	5.7	**9.4**	6.7	8.6	4.8	**21.5**
Prior treatments						
Yes	64.3	61.0	57.6	58.6	**100.0**	**37.5**

a. Pearson's $\chi 2$ test was conducted. Numbers shown in bold type indicate statistically significant differences between RCT and TEDS-A samples at $P < 0.05$

Table 4.2. Comparison of baseline characteristics (%) of four CTN trials and NSDUH

	Target population	Randomized controlled trials				
	NSDUH, 2013-2014 N= 235	Overall N=456	CTO0001 (1) N=119	MDS0004 (2) N=210	CTN0052 (3) N=62	CTO0005 (4) N=65
	%	%	%	%	%	%
Sex						
Female	37.4	**27.9**	29.4	**28.1**	37.1	**15.4**
Race						
White	52.8	**27.4**	**18.8**	**29.2**	**25.8**	**38.5**
Age						
<35	49.9	**19.1**	**22.7**	**14.3**	**6.5**	**35.3**
Education						
≥12 years	79.4	82.8	84.9	**89.5**	**64.5**	81.5
Employment						
Full-time	33.7	33.8	31.1	32.4	**21.0**	**55.4**
Marital Status						
Married	15.5	23.3	19.5	**26.2**	12.9	**30.8**
IV drug use						
Yes	6.3	9.4	6.7	8.6	4.8	**21.5**
Prior treatments						
Yes	65.5	61.0	57.6	58.6	**100.0**	**37.5**

a. Pearson's $\chi 2$ test was conducted. Numbers shown in bold type indicate statistically significant differences between RCT and NSDUH samples at $P < 0.05$.
b. The sampling weight was taken into consideration for the NSDUH target population.

Table 4.3. Unweighted and weighted meta-analysis of pharmacological cocaine dependence treatments with two target populations

Target population = TEDS-A					
Outcome	**Unweighted OR**	**95%CI**	**Weighted OR**	**95%CI**	**Statistics for difference**
Retention	1.36**	1.14, 1.62	1.76**	1.30, 2.38	χ^2 =5.48, p= 0.02
Outcome	**Unweighted β**	**95%CI**	**Weighted β**	**95%CI**	**Statistics for difference**
Abstinence	-0.81	-3.10, 1.48	-1.27	-2.76, 0.23	χ^2 =1.01, p=0.31
Target population = NSDUH					
Outcome	**Unweighted OR**	**95%CI**	**Weighted OR**	**95%CI**	**Statistics for difference**
Retention	1.36**	1.14, 1.62	2.74**	1.54, 4.88	χ^2 =5.99, p= 0.01
Outcome	**Unweighted β**	**95%CI**	**Weighted β**	**95%CI**	**Statistics for difference**
Abstinence	-0.81	-3.10, 1.48	-0.84	-3.12, 1.44	χ^2 =0.00, p=0.98

Note: Weighted results were weighted by the TEDS-A and NSDUH target populations.
We estimated multi-level mixed effect models, which allow random intercepts and a coefficient on treatment assignment variable. The standard errors were clustered at the trial level. The weight was truncated at 95 percentiles to eliminate extreme weights.
** p<0.01, * p<0.05

Table 4.4. Unweighted and weighted network meta-analysis of four pharmacological treatments for cocaine dependence.

Retention									
	Unweighted			Weighted (with TEDS-A)			Weighted (with NSDUH)		
	OR	95%CI	Rank (%)	OR	95%CI	Rank (%)	OR	95%CI	Rank (%)
Reserpine vs. Placebo	1.10	0.27, 1.94	5.4%	1.28	-0.31, 2.85	2.6%	1.97	0.48, 8.17	2.9%
Modafinil vs. Placebo	1.47	0.58, 2.36	17.7%	1.66	0.15, 3.18	4.7%	2.19	0.78, 6.17	1.9%
Buspirone vs. Placebo	1.84	-1.11, 4.79	47.2%	6.22	-6.46, 18.9	56.3%	7.80	0.70, 87.36	46.3%
Ondansetron vs. Placebo	1.58	-0.21, 3.37	29.8%	4.45	-3.78, 12.7	36.5%	8.45**	1.81, 39.25	48.9%
Abstinence									
	Unweighted			Weighted (with TEDS-A)			Weighted (with NSDUH)		
	Coefficient	95%CI	Rank (%)	Coefficient	95%CI	Rank (%)	Coefficient	95%CI	Rank (%)
Reserpine vs. Placebo	-2.63	-6.21, 0.94	2.3%	-2.13	-7.42, 3.16	15.0%	-1.43	-8.16, 5.34	23.0%
Modafinil vs. Placebo	1.14	-1.63, 3.90	61.9%	-0.09	-3.95, 3.78	40.0%	0.75	-3.44, 4.92	50.3%
Buspirone vs. Placebo	-2.15	-5.50, 1.21	3.2%	-2.62	-5.57, 0.32	2.0%	-3.66	-8.61, 1.32	3.2%
Ondansetron vs. Placebo	-1.82	-9.12, 5.46	20.6%	-5.05	-13.41, 3.24	8.5%	-4.07	-9.24, 1.12	2.6%

Note: Weighted results were weighted by the TEDS-A and NSDUH target populations. The weight was truncated at 95 percentiles to eliminate extreme weights. Rank (%) represents the estimated probability of each medication being the most efficacious medication among four medications. ** p<0.01, * p<0.05

CHAPTER 5 Conclusions and Policy Implications

5.1. Summary of main findings

The purpose of this dissertation was to assess generalizability of findings of various SUD RCTs to intended target populations. The Chapter 2 aimed to generalize the treatment effects estimated through SUD RCTs to intended target populations by comparing RCT sample treatment effects and the population effects of SUD treatment. In Chapter 2, we generalized three outcomes (retention, urine toxicology and abstinence) from ten RCTs from the NIDA CTN studies (five trials of Buprenorphine/Naloxone detoxification for opioid dependence, three trials of motivational enhancement/interviewing on SUD, and two trials of motivational incentives for cocaine, methamphetamine or amphetamine use) to the target populations of treatment-seeking individuals drawn from the TEDS-A. We demonstrated that the observed outcomes of some RCTs may not be directly applicable to potential target populations. Statistically significant treatment effects estimated in the RCT samples became insignificant when they were weighted to the target population in most cases (three trials for retention and urine toxicology, and one trial for abstinence); but also we found that insignificant effects became significantly positive (in one trial for abstinence as an outcome), and significantly negative effects became insignificant (in two trials for abstinence as an outcome). We also presented suggestive evidence that these differences in effect estimates between the unweighted and weighted models could be partially explained by treatment effect heterogeneity across over- and under- represented subgroups of the RCT participants as compared with the target populations.

The Chapter 3 aimed to assess the generalizability of the findings from a multi-site web-based SUD intervention to two different types of target populations. In that chapter, we generalized two outcomes (retention and

abstinence) of the RCT sample of a web-based SUD intervention, the Therapeutic Education System (TES), to two types of target populations: SUD treatment-seeking individuals drawn from the TEDS-A and community-dwelling individuals with recent substance use, whether or not they sought treatment, drawn from the NSDUH. We first demonstrated significant differences between the RCT sample and target populations of potential recipients of the intervention. As compared with the TEDS-A target population, the RCT sample had significantly lower proportions of Hispanic individuals, those with a history of intravenous drug use, and those with history of prior SUD treatment while the RCT sample had significantly higher proportions of those with 12 years or longer of education and those with full time jobs. As compared with the NSDUH target population, the RCT sample had significantly lower proportions of individuals between age 18-34 years, those with 12 years or longer of education, individuals with fulltime jobs, married individuals, and those with a history of prior SUD treatment while the RCT sample had significantly higher proportions of Blacks, those between age 35-49 years, and those with intravenous drug use. Moreover, we showed that the summary index of these differences, Δp, was much larger than the standardized Δp cutoff 0.25[29–31] or 0.10[32] for both target populations (1.62 and 1.14 standard deviations for the TEDS-A and NSDUH, respectively), indicating substantial differences between the RCT sample and the target populations. Finally, these analyses showed that the observed promising findings of the TES intervention may not be directly generalizable to potential target populations. We found that significant treatment effects on abstinence at 12 weeks and 6-month follow-up became insignificant after weighting. We also showed some suggestive evidence that these differences between the unweighted sample treatment effects and the estimated population effects could

be partially explained by the treatment effect heterogeneity across under- and over-represented subgroups of RCT participants.

The Chapter 4 aimed to conduct a meta-analysis of RCTs of cocaine dependence medications as well as a network meta-analysis to compare the effects of multiple treatments of cocaine dependence while considering generalizability of the findings of the RCTs to the target populations and adjusting the results to make them more generalizable. In Chapter 4, we generalized two outcomes (retention and abstinence) of four RCTs (Reserpine, Modafinil, Buspirone, and Ondansetron vs. placebo) drawn from the NIDA CTN to two types of target populations: treatment-seeking individuals with cocaine dependence drawn from the TEDS-A and community-dwelling individuals with cocaine dependence, regardless of their treatment seeking behaviors, drawn from the NSDUH. We found that the results from meta-analyses of cocaine dependence medications may not directly carry over to potential target populations. The estimated overall target population-weighted treatment effect of four cocaine dependence medications on retention was significantly larger than the treatment effect from the RCTs whether the effect was generalized by the TEDS-A or by NSDUH. Weighting the RCT samples to resemble target populations also changed the relative ranking of the treatment effectiveness across different medications. For retention, the second most efficacious treatment, Ondansetron, became the most efficacious after applying NSDUH generalizability weights. For abstinence, the least efficacious treatment, Reserpine, became the second most efficacious after weighting for both target populations. The results from the subgroup analysis of treatment effects partially explained these differences in effect estimates between unweighted meta-analysis and weighted meta-analyses.

5.2. Synthesis of findings

This dissertation applied the propensity score based weighting method to the context of SUD RCTs. Stuart et al.[41] generalized the results of a school-based behavioral intervention trial with propensity score-based weighting and found that the estimated population effect of intervention was slightly attenuated compared to the estimated sample effect from the RCT. The findings of this dissertation have implications for the external validity of results from SUD RCTs. Previous studies suggested that stringent eligibility criteria of SUD RCTs would create a substantial selection bias in RCT samples.[11] Furthermore, a recent study by Susukida et al.[12] showed substantial variability in the likelihood of being in RCT samples across patient subgroups by using the actual samples of RCTs and indicated that poor representation of target populations might impact the generalizability of findings from RCTs. The results of Chapter 2 confirm this prediction by showing differences in the statistical significance between the sample treatment effects and the population treatment effects with the data of ten RCT samples (five trials of Buprenorphine/Naloxone detoxification for opioid dependence, three trials of motivational enhancement/interviewing on SUD, and two trials of motivational incentives for cocaine, methamphetamine or amphetamine use). Chapter 2 also demonstrated 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.

Similar findings were shown in the context of a web-based SUD intervention (the Therapeutic Education System; TES) in Chapter 3. Whether the target population consisted of treatment-seeking individuals with recent drug use (TEDS-A) or individuals with recent drug use regardless of treatment seeking status (NSDUH), the composition of the RCT sample of the TES substantially

differed from those of the target populations. The findings of Chapter 3 indicated that the promising treatment effects of the TES intervention estimated through RCTs may not be directly generalizable to both types of target populations. Chapter 3 also demonstrated some suggestive evidence that these differences between the unweighted sample treatment effects and the estimated population effects could be partially explained by the treatment effect heterogeneity across under- and over-represented subgroups of RCT participants. Regarding 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.,[27] which found the significant treatment effect of TES estimated through RCT became insignificant after weighting the sample to resemble the target populations drawn from 2000-2001 NESARC data. Our study confirms that their findings hold when the generalizability of the TES RCT was assessed with the target populations from recent years.

To our knowledge, Chapter 4 was the first study to conduct a meta-analysis and network meta-analysis of multiple SUD treatments while using a weighting approach to enhance generalizability of the findings to the target populations. The findings from Chapter 4 have implications for future meta-analyses of SUD treatments. As shown in this study, the overall treatment effect size and comparative effects changed when the deviations of each RCT sample from the target population were taken into account. Unlike Chapter 2, which found a decrease in treatment effect sizes after weighting of data from ten SUD RCT samples, the effect size associated with the cocaine dependence medications on retention became significantly larger after weighting by population weights. This implies that the effect of weighting-based methods may vary depending on how and to what extent the composition of the RCT samples and target populations vary. Furthermore, differences in efficacies among different treatments for the

same condition may be impacted by the compositions of the RCT samples for each treatment. Target population weighting of the RCTs changed the relative ranking of treatments for cocaine dependence in this study. This study also showed some suggestive evidence that the mechanisms through which the population treatment effects were different from the sample effects could be partially explained by treatment effect heterogeneity among under- or over-represented subgroups of individuals in the RCTs.

5.3. Strengths and limitations of these findings

The results of this dissertation should be interpreted in light of certain strengths as well as limitations. Although a growing number of studies suggest that stringent eligibility criteria commonly used in RCTs of SUD treatment would create a substantial selection bias in the RCT samples, which may adversely impact generalizability of the findings from SUD RCTs, very few previous studies have examined generalizability of findings of actual SUD RCTs to the intended target populations. A major strength of this dissertation was its use of actual RCT samples of various trials of SUD interventions to assess generalizability of the findings of these RCTs to differently defined target populations by applying a novel approach with propensity score-based weighting.

The following are several limitations of this dissertation. 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 which may have differed between the RCT samples and target populations and also contributed to treatment heterogeneity. Second, due to the significant differences between the RCT samples and their target populations, the weighting-based method may not have made the RCT samples adequately resemble the target populations to estimate the population

treatment effects. Weighting the RCT samples to estimate the population treatment effects is more suitable when the RCT samples and the target populations are more similar to start with. Third, the primary goal of the original RCTs 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. Fourth, one of the data sources of our target population, TEDS-A data missed 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. Although the TEDS-A is one of the largest administrative data sources that cover the data on most patients with SUD in the US, some states exclude patients whose treatment is not covered by the state substance use agency funds such as Federal Block Grant funds.[44] Patients who received treatment at private hospitals are usually excluded from the TEDS-A unless they have licenses from the state substance abuse treatment agency. Therefore, the TEDS-A might not have represented the entire population of patients with SUD in the US. Fifth, the RCT sample of a web-based SUD intervention in Chapter 3 was collected in clinical (outpatient) settings. Due to its potential scalability, a web-based intervention could be implemented in non-clinical settings such as school or community settings; therefore, the applicability of the findings of Chapter 3 might be limited to the context of the web-based SUD RCTs in clinical settings only. Sixth, the number of trials included in Chapter 4 was limited, which may have limited the reliability of the network meta-analysis[112] to draw conclusions regarding which cocaine dependence treatment was more promising as compared with the alternative treatments.

5.4. Conclusions

Acknowledging these limitations, results from this dissertation provide insight into whether and how deviations in RCT sample representativeness from target populations influence the observed outcomes of various SUD RCTs. 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.[37] 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,[36] it becomes increasingly important to understand the applicability of the findings of RCTs in different populations with varying characteristics, contexts, and locations. Particularly given the potential scalability of web-based SUD interventions,[34] representativeness of the sample with regard to target populations of potential users of this intervention should be carefully considered. As a growing number of RCTs of SUD treatments are implemented in various settings, future meta-analytic studies should estimate treatment effects as well as comparative effectiveness while accounting for generalizability to intended target populations. The weighting-based approach used in this study is applicable to meta-analyses of clinical trials of other SUD treatments, as well as other health interventions, especially when generalizability of the findings is a concern. 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,[59] which may affect profiles of patient groups who seek and access treatments.

REFERENCES

1. Blanco C, Olfson M, Goodwin RD, 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):1276-1280. doi:ej07m03694 [pii].
2. Okuda M, Hasin DS, Olfson M, et al. Generalizability of clinical trials for cannabis dependence to community samples. *Drug Alcohol Depend*. 2010;111(1-2):177-181. doi:10.1016/j.drugalcdep.2010.04.009.
3. 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):265-270. doi:10.4088/JCP.12m07935.
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):81-87. doi:10.1037/per0000091.
5. 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*. 201AD;29(7):614-620. doi:10.1002/da.21937.
6. Rothwell PM. External validity of randomised controlled trials: “to whom do the results of this trial apply?” *Lancet*. 2005;365(9453):82-93. doi:10.1016/S0140-6736(04)17670-8.
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):729-730.
8. Ioannidis JP. Evolution and translation of research findings: from bench to where? *PLoS Clin Trials*. 2006;1(7):e36. doi:10.1371/journal.pctr.0010036.
9. 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):988-995. doi:10.1111/j.1530-0277.2007.00391.x.

10. 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):588-594. doi:10.1176/appi.ajp.157.4.588.
11. 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.
12. 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.
13. 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):1249-1257. doi:10.1111/j.1360-0443.2005.01175.x.
14. 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):757-764.
15. Insel TR. Beyond efficacy: The STAR*D trial. *Am J Psychiatry*. 2006;163(1):5-7. doi:10.1176/appi.ajp.163.1.5.
16. Blanco C, Olfson M, Goodwin RD, 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):1276-1280. doi:ej07m03694 [pii].
17. Charlson ME, Horwitz RI. Applying results of randomised trials to clinical practice: impact of losses before randomisation. *Br Med J (Clin Res Ed)*. 1984;289(6454):1281-1284. doi:10.1136/bmj.289.6454.1281.
18. Van Spall HGC, Toren A, Kiss A, Fowler RA. Eligibility criteria of randomized controlled trials published in high-impact general medical journals: a systematic sampling review. *JAMA*. 2007;297(11):1233-1240. doi:10.1001/jama.297.11.1233.

19. Blanco C, Olfsona M, Okuda M, Nunesa E V., Liua S-M, Hasin DS. Generalizability of clinical trials for alcohol dependence to community samples. *Drug Alcohol Depend*. 2008;1(98):123-128.
20. Hoertel N, Falissard B, Humphreys K, Gorwood P, Seigneurie A-S, Limosin F. Do Clinical Trials of Treatment of Alcohol Dependence Adequately Enroll Participants With Co-Occurring Independent Mood and Anxiety Disorders? *J Clin Psychiatry*. 2014;75(3):231-237. doi:10.4088/JCP.13m08424.
21. Storbjork J. Implications of enrolment eligibility criteria in alcohol treatment outcome research: Generalisability and potential bias in 1- and 6-year outcomes. *Drug Alcohol Rev*. 2014;33(6):604-611. doi:10.1111/dar.12211.
22. Okuda M, Hasin DS, Olfson M, et al. Generalizability of clinical trials for cannabis dependence to community samples. *Drug Alcohol Depend*. 2010;111(1-2):177-181. doi:10.1016/j.drugalcdep.2010.04.009.
23. Velasquez MM, Diclemente CC, Addy RC. Generalizability of Project Match: A comparison of clients enrolled to those not enrolled in the study at one aftercare site. *Drug Alcohol Depend*. 2000;59(2):177-182. doi:10.1016/S0376-8716(99)00118-0.
24. Sofuoglu M, Dudish-Poulsen S, Nicodemus KK, Babb DA, Hatsukami DK. Characteristics of research volunteers for inpatient cocaine studies: Focus on selection bias. *Addict Behav*. 2000;25(5):785-790. doi:10.1016/S0306-4603(00)00064-2.
25. Frewen AR, Baillie AJ, Montebello ME. Are cannabis users who participate in a randomized clinical trial different from other treatment seekers? *J Subst Abuse Treat*. 2009;36(3):339-344. doi:10.1016/j.jsat.2008.07.004.
26. Melberg HO, Humphreys K. Ineligibility and refusal to participate in randomised trials of treatments for drug dependence. *Drug Alcohol Rev*. 2010;29(2):193-201. doi:10.1111/j.1465-3362.2009.00096.x.
27. 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.

28. Stuart E a., 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):369-386. doi:10.1111/j.1467-985X.2010.00673.x.
29. 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.
30. Cochran WG, Rubin DB. Controlling bias in observational studies: A review. *Sankhyā Indian J Stat Ser A*. 1973;35(4):417-446.
31. Rubin DB. The use of matched sampling and regression adjustment to remove bias in observational studies. *Biometrics*. 1973;29(1):185-203.
32. 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.
33. Cole SR, Stuart EA. Generalizing evidence from randomized clinical trials to target populations: The ACTG 320 trial. *Am J Epidemiol*. 2010;172(1):107-115. doi:10.1093/aje/kwq084.
34. 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.
35. National Cancer Institute. An Overview of NCI's National Clinical Trials Network. http://www.cancer.gov/research/areas/clinical-trials/nctn.
36. Flay BR, Biglan A, Boruch RF, et al. Standards of evidence: Criteria for efficacy, effectiveness and dissemination. *Prev Sci*. 2005;6(3):151-175. doi:10.1007/s11121-005-5553-y.
37. Wang PS, Ulbricht CM, Schoenbaum M. Improving mental health treatments through comparative effectiveness research. *Heal Aff*. 2009;28(3):783-791. doi:10.1377/hlthaff.28.3.783.
38. Blanco C, Olfson M, Goodwin RD, 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):1276-1280.

39. 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):988-995.
40. 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):588-594.
41. Stuart EA, Bradshaw CP, Leaf PJ. Assessing the Generalizability of Randomized Trial Results to Target Populations. *Prev Sci*. 2015;16(3):475-485.
42. Mansournia MA, Altman DG. Inverse probability weighting. *BMJ Br Med J*. 2016.
43. 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.
44. Substance Abuse & Mental Health Services and Quality. Quick Statistics from the Drug and Alcohol Services Information System. http://wwwdasis.samhsa.gov/webt/information.htm. Accessed November 6, 2016.
45. Ling W, Amass L, Shoptaw S, 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):1090-1100. doi:10.1111/j.1360-0443.2005.01154.x.
46. Ling W, Hillhouse M, Domier C, et al. Buprenorphine tapering schedule and illicit opioid use. *Addiction*. 2009;104(2):256-265. doi:10.1111/j.1360-0443.2008.02455.x.
47. Woody GE, Poole SA, Subramaniam G, et al. Extended vs short-term buprenorphine-naloxone for treatment of opioid-addicted youth: a randomized trial. *JAMA*. 2008;300(17):2003-2011. doi:10.1001/jama.2008.574.
48. Weiss RD, Potter JS, Fiellin DA, 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):1238-1246. doi:10.1001/archgenpsychiatry.2011.121.

49. Ball SA, Martino S, Nich C, et al. Site matters: multisite randomized trial of motivational enhancement therapy in community drug abuse clinics. *J Consult Clin Psychol*. 2007;75(4):556-567. doi:10.1037/0022-006X.75.4.556.
50. Carroll KM, Ball SA, Nich 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):301-312. doi:10.1016/j.drugalcdep.2005.08.002.
51. Winhusen T, Kropp F, Babcock D, et al. Motivational enhancement therapy to improve treatment utilization and outcome in pregnant substance users. *J Subst Abus Treat*. 2008;35(2):161-173. doi:10.1016/j.jsat.2007.09.006.
52. Petry NM, Peirce JM, Stitzer ML, 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):1148-1156. doi:10.1001/archpsyc.62.10.1148.
53. Peirce JM, Petry NM, Stitzer ML, 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):201-208. doi:10.1001/archpsyc.63.2.201.
54. Liaw a, Wiener M. Classification and Regression by randomForest. *R news*. 2002;2(December):18-22. doi:10.1177/154405910408300516.
55. R Development Core Team R. R: A Language and Environment for Statistical Computing. *R Found Stat Comput*. 2016;1(2.11.1):409. doi:10.1007/978-3-540-74686-7.
56. Breiman L. Random Forests. *Mach Learn* . 2001;45(1):5-32.
57. Lee BK, Lessler J, Stuart EA. Improving propensity score weighting using machine learning. *Stat Med*. 2010;29(3):337-346. doi:10.1002/sim.3782.

58. Weesie J. Seemingly unrelated estimation and the cluster-adjusted sandwich estimator. *Stata Tech Bull*. 1999;52:34-47.
59. The Center for Medicaid and CHIP Services. *Affordable Care Act*.
60. Mokdad AH, Marks JS, Stroup DF, Gerberding JL. Actual causes of death in the United States, 2000. *JAMA*. 2004;291:1238-1245. doi:10.1001/jama.271.9.660c.
61. NIDA. Overdose Death Rates. *Natl Inst Drug Abus*. 2015;(December).
62. Bouchery EE, Harwood HJ, Sacks JJ, Simon CJ, Brewer RD. Economic costs of excessive alcohol consumption in the U.S., 2006. *Am J Prev Med*. 2011;41(5):516-524. doi:10.1016/j.amepre.2011.06.045.
63. National Drug Intelligence Center. National Drug Threat Assessment 2011. *Natl Drug Intell Cent*. 2011:1-72. doi:10.1037/e618352012-001.
64. U S Department of Health and Human Services. The Health Consequences of Smoking—50 Years of Progress: A Report of the Surgeon General. *U S Dep Heal Hum Serv Dis Control Prev Natl Cent Chronic Dis Prev Heal Promot Heal Off Smok Heal*. 2014:1-36. doi:NBK179276.
65. 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.
66. 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.
67. Kessler RC, Zhao S, Katz SJ, et al. Past-year use of outpatient services for psychiatric problems in the national comorbidity survey. *Am J Psychiatry*. 1999;156(1):115-123. doi:10.1176/ajp.156.1.115.
68. Blanco C, Iza M, Rodr??guez-Fern??ndez JM, Baca-Garc??a E, Wang S, Olfson M. Probability and predictors of treatment-seeking for substance use disorders in the U.S. *Drug Alcohol Depend*. 2015;149:136-144.

doi:10.1016/j.drugalcdep.2015.01.031.

69. Jhanjee S. Evidence based psychosocial interventions in substance use. *Indian J Psychol Med*. 2014;36(2):112-118. doi:10.4103/0253-7176.130960.

70. 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:10.1093/aje/kwq304.

71. Substance Abuse and Mental Health Services Administration. *Results from the 2012 National Survey on Drug Use and Health : Summary of National Findings*.; 2012.

72. 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.

73. 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;Forthcomin.

74. NIDA. Nationwide Trends. *Heal (San Fr*. 2015;(June):3-5. doi:10.1037/e302582003-001.

75. Campbell ANC, Nunes E V, 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.

76. Budney AJ, Higgins ST. *Therapy Manuals for Drug Addiction. Manual 2, a Community Reinforcement plus Vouchers Approach : Treating Cocaine Addiction*.; 1998.

77. 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.

78. Anaissi A, Kennedy PJ, Goyal M, Catchpoole DR. A balanced iterative

random forest for gene selection from microarray data. *BMC Bioinformatics*. 2013;14:261. doi:10.1186/1471-2105-14-261.

79. Chen C, Liaw A, Breiman L. *Using Random Forest to Learn Imbalanced Data*.; 2007.
80. 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.
81. Cohen J. *Statistical Power Analysis for the Behavioral Sciences* . 2nd Editio. Hillsdale, NJ: Lawrence Erlbaum Associates; 1988.
82. Center for Behavioral Health Statistics and Quality. *Behavioral Health Trends in the United States: Results from the 2014 National Survey on Drug Use and Health*.; 2015.
83. Substance Abuse and Mental Health Services Administration. Drug Abuse Warning Network, 2011: National Estimates of Drug-Related Emergency Department Visits. *HHS Publ No 13-4760, Daw Ser D-39*. 2013.
84. Morton WA. Cocaine and Psychiatric Symptoms. *Prim Care Companion J Clin Psychiatry*. 1999;1(4):109-113. doi:10.4088/PCC.v01n0403.
85. Lange RA, Hillis LD. Cardiovascular Complications of Cocaine Use. *N Engl J Med*. 2001;345(5):351-358. doi:10.1056/NEJM200108023450507.
86. Buttner a. Neuropathological alterations in cocaine abuse. *Curr Med Chem*. 2012;19:5597-5600. doi:10.2174/092986712803988947.
87. Riezzo I, Fiore C, De Carlo D, et al. Side effects of cocaine abuse: multiorgan toxicity and pathological consequences. *Curr Med Chem*. 2012;19(33):5624-5646. doi:10.2174/092986712803988893.
88. Qureshi AI, Chaudhry SA, Suri MFK. Cocaine use and the likelihood of cardiovascular and all-cause mortality: data from the Third National Health and Nutrition Examination Survey Mortality Follow-up Study. *J Vasc Interv Neurol*. 2014;7(1):76-82.
89. Kampman KM. What's new in the treatment of cocaine addiction? *Curr Psychiatry Rep*. 2010;12(5):441-447. doi:10.1007/s11920-010-0143-5.
90. Penberthy JK, Ait-Daoud N, Vaughan M, Fanning T. Review of treatment

for cocaine dependence. *Curr Drug Abuse Rev*. 2010;3(1):49-62. doi:10.2174/1874473711003010049.

91. Vanderwert RE, Marshall PJ, Nelson C a, Zeanah CH, Fox N a. Timing of intervention affects brain electrical activity in children exposed to severe psychosocial neglect. *PLoS One*. 2010;5(7):e11415. doi:10.1371/journal.pone.0011415.
92. National Institute on Drug Abuse. How is cocaine addiction treated?
93. de Lima MS, de Oliveira Soares BG, Reisser AAP, Farrell M. Pharmacological treatment of cocaine dependence: a systematic review. *Addiction*. 2002;97(8):931-949.
94. Kampman KM. New medications for the treatment of cocaine dependence. *Psychiatry (Edgmont)*. 2005;2(12):44-48.
95. Castells X, Casas M, Vidal X, et al. Efficacy of central nervous system stimulant treatment for cocaine dependence: a systematic review and meta-analysis of randomized controlled clinical trials. *Addiction*. 2007;102(12):1871-1887. doi:10.1111/j.1360-0443.2007.01943.x.
96. Shorter D, Domingo CB, Kosten TR. Emerging drugs for the treatment of cocaine use disorder: a review of neurobiological targets and pharmacotherapy. *Expert Opin Emerg Drugs*. 2015;20(1):15-29. doi:10.1517/14728214.2015.985203.
97. Dackis C, O'Brien C. Glutamatergic Agents for Cocaine Dependence. In: *Annals of the New York Academy of Sciences*. Vol 1003. ; 2003:328-345. doi:10.1196/annals.1300.021.
98. Carroll KM, Fenton LR, Ball SA, et al. Efficacy of disulfiram and cognitive behavior therapy in cocaine-dependent outpatients: a randomized placebo-controlled trial. *Arch Gen Psychiatry*. 2004;61(3):264-272. doi:10.1001/archpsyc.61.3.264.
99. Arndt IO, McLellan AT, Dorozynsky L, Woody GE, O'Brien CP. Desipramine treatment for cocaine dependence. Role of antisocial personality disorder. *J Nerv Ment Dis*. 1994;182(3):151-156.
100. Martell BA, Mitchell E, Poling J, Gonsai K, Kosten TR. Vaccine

pharmacotherapy for the treatment of cocaine dependence. *Biol Psychiatry*. 2005;58(2):158-164. doi:10.1016/j.biopsych.2005.04.032.

101. Riley RD, Lambert PC, Abo-Zaid G. Meta-analysis of individual participant data: rationale, conduct, and reporting. *BMJ*. 2010;340:c221. doi:10.1136/bmj.c221.
102. Mills EJ, Thorlund K, Ioannidis JP a. Demystifying trial networks and network meta-analysis. *Bmj*. 2013;2914(May):10-15. doi:10.1136/bmj.f2914.
103. Anderson AL, Reid MS, Li S-H, et al. Modafinil for the treatment of cocaine dependence. *Drug Alcohol Depend*. 2009;104(1-2):133-139. doi:10.1016/j.drugalcdep.2009.04.015.
104. Winhusen TM, Kropp F, Lindblad R, et al. A multi-site, double-blind, placebo-controlled pilot clinical trial to evaluate the efficacy of buspirone as a relapse-prevention treatment for cocaine dependence. *J Clin Psychiatry*. 2014;75(7):757-764. doi:10.4088/JCP.13m08862.
105. Johnson BA, Roache JD, Ait-Daoud N, et al. A preliminary randomized, double-blind, placebo-controlled study of the safety and efficacy of ondansetron in the treatment of cocaine dependence. *Drug Alcohol Depend*. 2006;84(3):256-263. doi:S0376-8716(06)00070-6 [pii]10.1016/j.drugalcdep.2006.02.011.
106. American Psychiatric Association. *Diagnostic and Statistical Manual of Mental Disorders, Fourth Edition, Text Revision*. Vol Washington.; 2000. doi:10.1002/jps.3080051129.
107. Winhusen T, Somoza E, Sarid-Segal O, et al. A double-blind, placebo-controlled trial of reserpine for the treatment of cocaine dependence. *Drug Alcohol Depend*. 2007;91(2-3):205-212. doi:10.1016/j.drugalcdep.2007.05.021.
108. Lee BK, Lessler J, Stuart EA. Weight trimming and propensity score weighting. *PLoS One*. 2011;6(3). doi:10.1371/journal.pone.0018174.
109. Salanti G. Indirect and mixed-treatment comparison, network, or multiple-treatments meta-analysis: many names, many benefits, many concerns for

the next generation evidence synthesis tool. *Res Synth Methods*. 2012;3(March):80-97. doi:10.1002/jrsm.1037.

110. Cipriani A, Higgins JPT, Geddes JR, Salanti G. Conceptual and technical challenges in network meta-analysis. *Ann Intern Med*. 2013;159(2):130-137. doi:10.7326/0003-4819-159-2-201307160-00008.
111. Salanti G, Ades AE, Ioannidis JPA. Graphical methods and numerical summaries for presenting results from multiple-treatment meta-analysis: An overview and tutorial. *J Clin Epidemiol*. 2011;64(2):163-171. doi:10.1016/j.jclinepi.2010.03.016.
112. Jansen JP, Fleurence R, Devine B, et al. Interpreting indirect treatment comparisons and network meta-analysis for health-care decision making: Report of the ISPOR task force on indirect treatment comparisons good research practices: Part 1. *Value Heal*. 2011;14(4):417-428. doi:10.1016/j.jval.2011.04.002.

APPENDICES

Appendix Table 2.1. Comparison of baseline characteristics (%) of the samples in ten National Institute of Drug Abuse Clinical Trial Network (CTN) studies and target samples from the Treatment Episodes Data-Admission (TEDS-A).

	Buprenorphine/Naloxone (Bup/Nx) Detoxification										**Motivational enhancement/interviewing**						**Motivational incentives**			
	CTN0001		**CTN0002**		**CTN0003**		**CTN0010**		**CTN0030**		**CTN0004**		**CTN0005**		**CTN0013**		**CTN0006**		**CTN0007**	
	RCT	**TEDS**	**RCT**	**TEDS**	**RCT**	**TEDS**	**RCT**	**TEDS**	**RCT**	**TEDS**	**RCT**	**TEDS**	**RCT**	**TEDS**	**RCT**	**TEDS**	**RCT**	**TEDS**	**RCT**	**TEDS**
Total number	113	3,111	230	57,959	516	157,619	154	22,588	653	260,754	461	520,636	423	258,887	200	57,526	454	213,869	388	49,277
Sex																				
Female	**39.8**[1]	25.9	**28.3**	38.2	**32.8**	39.4	**41.6**	32.4	**39.9**	44.3	**29.1**	35.6	**42.1**	35.4	100.0	100.0	**54.9**	38.7	45.1	40.8
Race																				
White	55.8	52.5	**40.0**	53.2	**71.1**	57.6	70.1	69.4	**91.0**	68.5	**42.0**	62.9	**71.9**	62.2	**37.2**	58.0	**35.7**	54.4	**24.9**	50.4
Age																				
≥ 35	55.8	48.0	**66.9**	51.2	**54.3**	48.0	0.0	0.0	**37.8**	46.2	**55.7**	45.2	44.7	45.5	**9.5**	15.4	**59.3**	48.7	**85.2**	56.7
Education																				
≥12 years	**87.6**	63.0	**73.5**	63.5	**84.3**	61.8	**51.9**	28.5	**84.4**	66.7	**80.5**	63.4	**76.4**	62.7	54.5	54.4	66.5	62.2	64.5	61.3
Employment																				
Full-time	**58.0**	5.6	**54.3**	22.5	**58.9**	18.6	--[2]	--[2]	**63.1**	16.1	**58.8**	24.3	**57.4**	25.5	**32.5**	8.6	**48.2**	20.6	**31.7**	14.6
Marital Status																				
Married	**33.6**	12.8	**24.3**	18.5	**30.2**	17.8	**14.3**	1.9	**28.7**	17.7	18.4	16.8	19.9	16.9	14.5	16.7	**23.3**	16.2	14.0	14.8
Criminal justice admission																				
Yes	**8.9**	3.9	**3.1**	31.6	**2.5**	28.5	--[2]	--[2]	**0.5**	18.7	32.2	35.8	**53.2**	35.6	**13.0**	31.4	35.9	30.9	**5.0**	17.7
IV drug use																				
Yes	**31.1**	18.5	26.7	26.0	26.3	29.7	--[2]	--[2]	**3.4**	36.8	**6.0**	20.0	16.3	20.5	**2.8**	16.3	**9.3**	22.2	**35.6**	52.5
# prior treatments																				
≥ 5 times	--[3]	--[3]	**15.4**	6.1	**20.4**	7.7	--[2]	--[2]	**4.1**	11.6	**24.1**	13.6	**9.0**	13.9	**11.4**	7.3	**19.1**	14.6	**42.6**	23.0

[1]Pearson chi-square test was conducted. Numbers written in bold letters indicate statistically significant differences between RCT and TEDS-A samples at p<.05.

[2] Not included in the analyses as these variables were not available for CTN0010.

[3] Not included in the analyses because of a large number of missing values for this variable in TEDS-A.

Appendix Table 2.2. Types of subgroup analyses

CTN trial	Outcome	Characteristics identifying subgroups
0001	Self-reported abstinence	Sex
		Education
		Employment
		Marital status
		Criminal justice
		IV drug use
0002	Retention	Sex
		Race
		Age
		Education
		Employment
		Marital status
		Criminal justice
		Prior treatment
	Self-reported abstinence	Sex
		Race
		Age
		Education
		Employment
		Marital status
		Criminal justice
		Prior treatment
0003	Retention	Sex
		Race
		Age
		Education
		Employment
		Marital status
		Criminal justice
		Prior treatment
	Urine toxicology	Sex
		Race
		Age
		Education
		Employment
		Marital status
		Criminal justice
		Prior treatment
0010	Retention	Sex
		Education
		Marital status
	Urine toxicology	Sex
		Education
		Marital status
	Self-reported abstinence	Sex
		Education
		Marital status
0030	Self-reported abstinence	Sex
		Race
		Age
		Education
		Employment
		Marital status
		Criminal justice
		IV drug use
		Prior treatment
0004	Self-reported abstinence	Sex
		Race
		Education
		Employment
		IV drug use
		Prior treatment
0006	Retention	Sex
		Race
		Age
		Employment
		Marital status
		IV drug use
		Prior treatment

Appendix Table 2.3. Results of the subgroup analysis of treatment effects

		Retention (end of trial)				Urine tox (end of trial)				Abstinence (follow-up 1)			
		T	C	Test	Interaction	T	C	Test	Interaction	T	C	Test	Interaction
Study	Variable												
CTN0001	**Sex**												
	Men	-	-	-	-	-	-	-	-	25.6	22.7	t=-0.9	--
	Women	-	-	-	-	-	-	-	-	25.3	15.4	**t=-2.6**	p=0.16
	Education												
	< 12 years	-	-	-	-	-	-	-	-	26.2	30.0	t=0.9	--
	≥ 12 years	-	-	-	-	-	-	-	-	25.4	17.8	**t=-2.9**	p=0.15
	Employment												
	Full-time	-	-	-	-	-	-	-	-	24.4	19.2	t=-1.5	--
	Other	-	-	-	-	-	-	-	-	27.3	21.3	t =-1.9	p=0.87
	Marital status												
	Married	-	-	-	-	-	-	-	-	25.8	15.0	**t=-2.1**	--
	Non-married	-	-	-	-	-	-	-	-	25.3	20.4	t=-1.8	p=0.29
	Criminal justice												
	Yes	-	-	-	-	-	-	-	-	29.2	30.0	t=1.2	--
	No	-	-	-	-	-	-	-	-	25.0	16.3	**t=-3.2**	p=0.15
	IV drug use												
	Yes	-	-	-	-	-	-	-	-	25.2	16.7	t=-1.7	--
	No	-	-	-	-	-	-	-	-	25.5	20.1	t =-1.9	p=0.57
CTN0002	**Sex**												
	Men	56.1	31.4	**χ_1^2=8.7**	--	-	-	-	-	18.9	17.4	t=-0.5	--
	Women	57.1	17.4	**χ_1^2=9.6**	p=0.26	-	-	-	-	16.9	10.9	t=-1.4	p=0.39
	Race												
	White	54.8	26.7	**χ_1^2=6.5**	--	-	-	-	-	18.3	19.3	t=0.2	--
	Non-White	57.5	27.3	**χ_1^2=10.9**	p=0.91	-	-	-	-	18.3	13.5	t=-1.5	p=0.28
	Age												
	< 35	46.3	13.6	**χ_1^2=7.2**	--	-	-	-	-	16.8	14.8	t=-0.3	--
	≥ 35	61.8	32.7	**χ_1^2=11.7**	p=0.52	-	-	-	-	19.0	15.3	t=-1.4	p=0.81
	Education												
	< 12 years	55.0	14.3	**χ_1^2=9.4**	--	-	-	-	-	16.4	12.0	t=-0.8	--
	≥ 12 years	56.9	32.1	**χ_1^2=9.0**	p=0.22	-	-	-	-	18.8	16.4	t=-0.9	p=0.74
	Employment												
	Full-time	59.5	34.2	**χ_1^2=7.1**	--	-	-	-	-	19.9	17.3	t=-0.9	--
	Other	52.8	18.2	**χ_1^2=11.1**	p=0.38	-	-	-	-	16.5	10.3	t=-1.4	p=0.48
	Marital status												
	Married	58.6	14.3	**χ_1^2=4.4**	--	-	-	-	-	17.8	16.0	--[1]	--
	Non-married	55.9	28.4	**χ_1^2=13.4**	p=0.41	-	-	-	-	18.4	15.2	t=-1.2	p=0.90
	Criminal justice												
	Yes	83.3	0.0	χ_1^2=2.9	--	-	-	-	-	16.7	--[3]	--[3]	--
	No	55.0	27.8	**χ_1^2=14.5**	--[2]	-	-	-	-	18.5	15.2	t=-1.3	--[3]
	Prior treatment												
	< 5	56.6	29.8	**χ_1^2=11.5**	--	-	-	-	-	18.5	15.6	t=-1.1	--
	≥ 5	55.0	17.7	**χ_1^2=5.5**	p=0.47	-	-	-	-	17.1	13.0	t=-0.6	p=0.87
CTN0003	**Sex**												
	Men	76.2	63.4	**χ_1^2=6.7**	--	42.4	29.7	**χ_1^2=6.1**	--	-	-	-	-
	Women	84.3	66.3	**χ_1^2=7.4**	p=0.38	47.0	30.2	**χ_1^2=5.0**	p=0.69	-	-	-	-
	Race												
	White	78.1	60.3	**χ_1^2=13.6**	--	42.3	26.3	**χ_1^2=10.4**	--	-	-	-	-
	Non-White	80.9	72.8	χ_1^2=1.3	p=0.39	48.5	37.0	χ_1^2=2.0	p=0.54	-	-	-	-

	Age												
	< 35	79.2	61.2	**χ_1^2=9.1**	--	40.0	27.6	**χ_1^2=4.1**	--	-	-	-	-
	≥ 35	78.5	66.9	χ_1^2=4.7	p=0.48	47.4	31.7	**χ_1^2=7.2**	p=0.78	-	-	-	-
	Education												
	< 12 years	75.6	60.0	χ_1^2=2.3	--	43.9	32.5	χ_1^2=1.1	--	-	-	-	-
	≥ 12 years	79.4	65.2	**χ_1^2=11.0**	p=0.99	43.9	29.1	**χ_1^2=9.9**	p=0.77	-	-	-	-
	Employment												
	Full-time	77.8	60.7	**χ_1^2=12.2**	--	47.7	27.7	**χ_1^2=12.8**	--	-	-	-	-
	Other	80.2	72.6	χ_1^2=1.7	p=0.24	38.7	33.0	χ_1^2=0.7	p=0.10	-	-	-	-
	Marital status												
	Married	79.8	63.9	**χ_1^2=4.9**	--	46.4	37.5	χ_1^2=1.3	--	-	-	-	-
	Non-married	78.4	64.6	**χ_1^2=8.3**	p=0.80	42.7	27.0	**χ_1^2=9.8**	p=0.40	-	-	-	-
	Criminal justice												
	Yes	100.0	62.5	χ_1^2=2.4	--	40.0	12.5	χ_1^2=1.3	--	-	-	-	-
	No	78.4	64.4	**χ_1^2=12.0**	p=0.98	44.0	30.4	**χ_1^2=9.9**	p=0.50	-	-	-	-
	Prior treatment												
	< 5	76.9	66.0	**χ_1^2=5.9**	--	44.7	31.1	**χ_1^2=8.1**	--	-	-	-	-
	≥ 5	85.7	57.1	**χ_1^2=10.7**	p=0.07	41.1	24.5	χ_1^2=3.2	p=0.70	-	-	-	-
CTN0010	**Sex**												
	Men	78.6	56.3	**χ_1^2=5.0**	p=0.83	62.5	15.6	**χ_1^2=14.8**	p=0.30	22.0	20.8	t=-0.4	p=0.24
	Women	78.1	59.4	χ_1^2=2.6	--	54.8	22.9	**χ_1^2=9.7**	--	18.1	22.6	t=1.2	--
	Education												
	< 12 years	69.9	55.3	χ_1^2=1.6	--	47.2	13.2	**χ_1^2=10.3**	--	20.5	19.7	t=-1.2	--
	≥ 12 years	86.8	59.5	**χ_1^2=7.5**	p=0.24	68.4	26.2	**χ_1^2=14.3**	p=0.97	20.5	22.7	t=0.8	p=0.50
	Marital status												
	Married	71.4	40.0	χ_1^2=1.9	--	28.6	0.0	**χ_1^2=4.7**	--	29.2	18.5	**t=-2.3**	--
	Non-married	79.1	61.5	**χ_1^2=4.9**	p=0.67	61.2	24.6	**χ_1^2=18.1**	p=0.99	19.2	22.3	t=1.2	**p=0.03**
CTN0030	**Sex**												
	Men	-	-	-	-	-	-	-	-	18.4	21.2	**t=2.5**	--
	Women	-	-	-	-	-	-	-	-	18.8	19.5	t=0.5	p=0.21
	Race												
	White	-	-	-	-	-	-	-	-	18.5	20.3	**t=2.2**	--
	Non-White	-	-	-	-	-	-	-	-	19.9	21.2	t=0.5	p=0.84
	Age												
	< 35	-	-	-	-	-	-	-	-	18.2	20.3	**t=2.2**	--
	≥ 35	-	-	-	-	-	-	-	-	19.3	20.5	t=0.8	p=0.55
	Education												
	< 12 years	-	-	-	-	-	-	-	-	18.2	20.9	t=1.6	--
	≥ 12 years	-	-	-	-	-	-	-	-	18.7	20.3	t=1.8	p=0.59
	Employment												
	Full-time	-	-	-	-	-	-	-	-	18.6	20.8	**t=2.0**	--
	Other	-	-	-	-	-	-	-	-	18.5	19.9	t=1.1	p=0.67
	Marital status												
	Married	-	-	-	-	-	-	-	-	17.7	19.8	t=1.3	--
	Non-married	-	-	-	-	-	-	-	-	19.0	20.7	t=1.9	p=0.84
	Criminal justice												
	Yes	-	-	-	-	-	-	-	-	25.0	11.0	--[4]	--
	No	-	-	-	-	-	-	-	-	18.7	20.4	**t=2.1**	p=0.18
	IV drug use												
	Yes	-	-	-	-	-	-	-	-	15.1	19.9	t=1.2	--
	No	-	-	-	-	-	-	-	-	18.8	20.4	**t=2.0**	p=0.44
	Prior treatment												
	< 5	-	-	-	-	-	-	-	-	18.9	20.2	t=1.6	--
	≥ 5	-	-	-	-	-	-	-	-	17.0	23.5	t=1.8	p=0.19

CTN0004	**Sex**												
	Men	-	-	-	-	-	-	-	-	25.3	27.2	**t=2.0**	--
	Women	-	-	-	-	-	-	-	-	22.6	26.3	t=1.8	p=0.36
	Race												
	White	-	-	-	-	-	-	-	-	24.2	26.3	t=1.5	--
	No-White	-	-	-	-	-	-	-	-	24.8	27.3	**t=2.3**	p=0.86
	Education												
	< 12 years	-	-	-	-	-	-	-	-	24.1	26.8	t=1.2	--
	≥ 12 years	-	-	-	-	-	-	-	-	24.5	27.0	**t=2.6**	p=0.93
	Employment												
	Full-time	-	-	-	-	-	-	-	-	25.9	27.3	t=1.3	--
	Other	-	-	-	-	-	-	-	-	22.2	26.5	**t=2.9**	p=0.10
	IV drug use												
	Yes	-	-	-	-	-	-	-	-	22.6	25.8	t=0.6	--
	No	-	-	-	-	-	-	-	-	24.3	26.8	**t=2.7**	p=0.89
	Prior treatment												
	< 5	-	-	-	-	-	-	-	-	24.2	27.2	**t=3.0**	--
	≥ 5	-	-	-	-	-	-	-	-	25.0	26.2	t=0.6	p=0.40
CTN0006	**Sex**												
	Men	41.5	34.3	χ_1^2=1.1	--	37.7	33.3	χ_1^2=0.4	--	-	-	-	-
	Women	46.8	31.7	**χ_1^2=6.0**	p=0.39	37.3	25.2	**χ_1^2=4.2**	p=0.35	-	-	-	-
	Race												
	White	50.0	30.7	**χ_1^2=6.2**	--	44.6	25.0	**χ_1^2=6.9**	--	-	-	-	-
	Non-White	41.8	34.3	χ_1^2=1.7	p=0.22	34.2	31.3	χ_1^2=0.3	p=0.07	-	-	-	-
	Age												
	< 35	52.7	30.4	**χ_1^2=9.4**	--	45.2	27.2	**χ_1^2=6.5**	--	-	-	-	-
	≥ 35	34.6	38.9	χ_1^2=0.5	p=0.06	32.4	30.0	χ_1^2=0.2	p=0.10	-	-	-	-
	Employment												
	Full-time	42.8	34.0	**χ_1^2=14.2**	--	41.8	27.5	**χ_1^2=4.9**	--	-	-	-	-
	Other	42.6	34.5	χ_1^2=1.6	p=0.24	33.6	30.1	χ_1^2=0.3	p=0.24	-	-	-	-
	Marital status												
	Married	52.3	29.0	**χ_1^2=5.9**	--	45.5	25.8	**χ_1^2=4.4**	--	-	-	-	-
	Non-married	42.6	34.4	χ_1^2=2.4	p=0.17	35.6	30.0	χ_1^2=1.2	p=0.20	-	-	-	-
	IV drug use												
	Yes	42.9	33.3	χ_1^2=0.4	--	37.1	29.4	χ_1^2=2.8	--	-	-	-	-
	No	44.8	32.8	**χ_1^2=6.1**	p=0.17	42.9	23.8	χ_1^2=1.7	p=0.46	-	-	-	-
	Prior treatment												
	< 5	42.8	31.6	**χ_1^2=4.8**	--	36.4	27.7	χ_1^2=3.1	--	-	-	-	-
	≥ 5	50.0	36.4	χ_1^2=1.6	p=0.19	45.2	31.8	χ_1^2=1.6	p=0.30	-	-	-	-

1: There was only one observation in the control arm and it was not possible to calculate t-statistics.

2: There was no observations in the control arm and it was not possible to estimate a coefficient for interaction term.

3: There was no observations in the control arm and it was not possible to calculate the mean, t-statistics p-value for interaction term.

4: There was only one observation each in the treatment arm and in the control arm and it was not possible to calculate t-statistics.

Appendix Table 3.1. Results of subgroup analysis of treatment effects

	Retention (12 weeks)				Abstinence (12 weeks)				Retention (3 months)				Abstinence (3 months)				Retention (6 months)				Abstinence (6 months)			
	T	C	Test	T*X[1]	T	C	Test	T*X	T	C	Test	T*X	T	C	Test	T*X	T	C	Test	T*X	T	C	Test	T*X
Variable																								
Sex																								
Men	85.4	84.8	χ_1^2=0.0	--	60.0	46.1	**χ_1^2=5.2**	--	87.8	90.1	χ_1^2=0.4	--	54.2	47.1	χ_1^2=1.4	--	50.0	48.3	χ_1^2=0.1	--	25.6	13.7	χ_1^2=3.4	--
Women	86.8	90.1	χ_1^2=0.5	p=0.50	54.4	44.0	χ_1^2=1.9	p=0.72	86.8	95.1	**χ_1^2=4.0**	p=0.20	48.1	49.0	χ_1^2=0.0	p=0.41	60.4	60.4	χ_1^2=0.0	p=0.86	21.8	13.1	χ_1^2=1.5	p=0.81
Race																								
White	81.3	85.6	χ_1^2=1.0	--	54.9	50.4	χ_1^2=0.5	--	84.2	90.2	χ_1^2=2.4	--	49.6	48.6	χ_1^2=0.0	--	58.3	51.6	χ_1^2=1.3	--	29.6	17.7	χ_1^2=3.1	--
Non-White	91.9	88.9	χ_1^2=0.5	p=0.24	59.8	37.5	**χ_1^2=9.4**	p=0.07	91.9	95.0	χ_1^2=0.8	p=0.95	54.9	46.8	χ_1^2=1.3	p=0.46	50.5	55.6	χ_1^2=0.5	p=0.19	16.1	7.3	χ_1^2=2.1	p=0.76
Age																								
< 35	81.8	85.0	χ_1^2=0.5	--	51.5	45.1	χ_1^2=0.9	--	88.9	92.5	χ_1^2=1.0	--	49.1	43.1	χ_1^2=0.9	--	55.6	58.7	χ_1^2=0.3	--	24.3	12.8	χ_1^2=3.3	--
≥ 35	89.9	89.1	χ_1^2=0.0	p=0.55	63.8	45.3	**χ_1^2=7.7**	p=0.20	86.1	91.6	χ_1^2=1.9	p=0.82	55.0	53.2	χ_1^2=0.1	p=0.65	51.9	47.1	χ_1^2=0.6	p=0.37	23.9	14.3	χ_1^2=1.8	p=0.82
Education																								
< 12 years	85.0	93.1	χ_1^2=2.0	--	62.8	38.9	**χ_1^2=6.0**	--	81.7	96.6	**χ_1^2=6.7**	--	42.9	41.1	χ_1^2=0.0	--	55.0	65.5	χ_1^2=1.4	--	24.2	13.2	χ_1^2=1.5	--
≥ 12 years	86.2	85.1	χ_1^2=0.1	p=0.17	56.6	47.3	χ_1^2=2.9	p=0.19	89.2	90.7	χ_1^2=0.2	p=0.05	54.6	50.0	χ_1^2=0.7	p=0.81	53.3	49.5	χ_1^2=0.6	p=0.17	24.0	13.5	χ_1^2=3.6	p=0.95
Employment																								
Full-time	89.3	84.0	χ_1^2=1.1	--	62.9	50.0	χ_1^2=2.6	--	92.2	93.0	χ_1^2=0.0	--	57.9	52.7	χ_1^2=0.5	--	50.5	43.0	χ_1^2=1.1	--	26.9	18.6	χ_1^2=0.9	--
Other	83.6	88.8	χ_1^2=1.8	p=0.10	55.1	42.2	**χ_1^2=4.4**	p=0.94	84.2	91.5	χ_1^2=3.7	p=0.37	47.7	44.6	χ_1^2=0.2	p=0.82	55.9	59.9	χ_1^2=0.5	p=0.21	22.4	11.0	**χ_1^2=4.1**	p=0.57
Marital status																								
Married	88.9	83.3	χ_1^2=0.5	--	62.5	40.0	χ_1^2=3.1	--	94.4	91.7	χ_1^2=0.2	--	55.9	57.6	χ_1^2=0.0	--	47.2	58.3	χ_1^2=0.9	--	35.3	19.1	χ_1^2=1.3	--
Non-married	85.4	87.5	χ_1^2=0.4	p=0.39	57.2	46.0	**χ_1^2=4.7**	p=0.41	86.3	92.1	χ_1^2=3.8	p=0.29	51.3	46.2	χ_1^2=1.0	p=0.61	54.8	52.3	χ_1^2=0.3	p=0.29	22.5	12.4	**χ_1^2=4.1**	p=0.88
IV drug use																								
Yes	75.0	88.9	χ_1^2=1.2	--	46.7	62.5	χ_1^2=0.8	--	70.0	100.0	**χ_1^2=6.4**	--	57.1	55.6	χ_1^2=0.0	--	40.0	50.0	χ_1^2=0.4	--	25.0	33.3	χ_1^2=0.1	--
No	86.8	86.8	χ_1^2=0.0	p=0.30	58.8	43.8	**χ_1^2=9.1**	p=0.10	88.9	91.5	χ_1^2=0.8	--[2]	51.7	47.2	χ_1^2=0.8	p=0.88	54.9	53.4	χ_1^2=0.1	p=0.50	24.0	12.0	χ_1^2=0.2	p=0.27
Prior treatment																								
Yes	90.0	84.2	χ_1^2=0.3	--	44.4	37.5	χ_1^2=0.2	--	95.0	79.0	χ_1^2=2.2	--	42.1	53.3	χ_1^2=0.4	--	55.0	52.6	χ_1^2=0.0	--	9.1	10.0	χ_1^2=0.0	--
No	85.5	87.1	χ_1^2=0.3	p=0.52	59.2	45.8	**χ_1^2=7.3**	p=0.73	86.8	93.1	**χ_1^2=5.2**	p=0.05	52.9	47.5	χ_1^2=1.3	p=0.35	53.6	53.2	χ_1^2=0.0	p=0.91	25.4	13.7	**χ_1^2=5.4**	p=0.57

1. Interaction term between treatment dummy and each variable.

2. All the individuals in the control arm successfully retained in the study and it was not possible to estimate the interaction term.

Appendix Table 4.1. Number and percentage of cases with missing values for each covariate in the four cocaine dependence clinical trials and the target populations.

	TEDS-A, 2012	NSDUH, 2013-14	CTO0001	MDS0004	CTN0052	CTO0005
	Target	Target	RCT	RCT	RCT	RCT
Total Valid N	32,287	235	115	209	62	65
Total Missing N	4,710	0	4	1	0	0
% Total Missing	12.7	0.0	3.3	0.5	0.0	0.0
Sex						
Valid N	36,919	235	119	210	62	65
Missing N	78	0	0	0	0	0
% Missing	0.2	0.0	0.0	0.0	0.0	0.0
Race						
Valid N	36,734	235	117	209	62	65
Missing N	263	0	2	1	0	0
% Missing	0.7	0.0	1.7	0.5	0.0	0.0
Age						
Valid N	36,997	235	119	210	62	65
Missing N	0	0	0	0	0	0
% Missing	0.0	0.0	0.0	0.0	0.0	0.0
Education						
Valid N	36,281	235	119	210	62	65
Missing N	716	0	0	0	0	0
% Missing	0.2	0.0	0.0	0.0	0.0	0.0
Employment						
Valid N	35,965	235	119	210	62	65
Missing N	1,032	0	0	0	0	0
% Missing	2.8	0.0	0.0	0.0	0.0	0.0
Marital Status						
Valid N	34,428	235	118	210	62	65
Missing N	2,569	0	1	0	0	0
% Missing	6.9	0.0	0.8	0.0	0.0	0.0
IV drug use						
Valid N	36,328	235	119	210	62	65
Missing N	669	0	0	0	0	0
% Missing	0.2	0.0	0.0	0.0	0.0	0.0
Number of prior treatments						
Valid N	35,800	235	118	210	62	65
Missing N	1,197	0	1	0	0	0
% Missing	0.3	0.0	0.8	0.0	0.0	0.0

Appendix Table 4.2. Results of subgroup analysis of treatment effects

		Retention (end of trial)				Abstinence (end of trial)			
		T	C	Test	Interaction	T	C	Test	Interaction
Study (Treatment)	Variable								
Overall	**Sex**								
	Men	69.2	63.6	χ_1^2=1.1	--	22.3	23.2	t=0.8	--
	Women	71.6	67.9	χ_1^2=0.2	p=0.87	21.3	22.0	t=0.4	p=0.91
	Race								
	White	65.9	54.8	χ_1^2=1.5	--	22.6	25.0	t=1.3	--
	Non-White	71.4	68.5	χ_1^2=0.3	p=0.48	21.8	22.3	t=0.4	p=0.37
	Age								
	< 35	59.7	43.3	χ_1^2=2.1	--	20.2	23.2	t=1.1	--
	≥ 35	72.4	69.4	χ_1^2=0.4	p=0.32	22.5	22.8	t=0.3	p=0.33
	Education								
	< 12 years	75.0	73.1	χ_1^2=0.0	--	21.2	24.6	t=1.5	--
	≥ 12 years	68.8	63.5	χ_1^2=1.1	p=0.82	22.3	22.5	t=0.2	p=0.19
	Employment								
	Full-time	58.8	48.1	χ_1^2=1.6	--	22.7	22.7	t=-0.0	--
	Other	76.1	72.1	χ_1^2=0.6	p=0.60	21.8	22.9	t=1.0	p=0.61
	Marital status								
	Married	56.3	69.1	χ_1^2=1.8	--	22.4	22.4	t=-0.0	--
	Non-married	73.7	63.6	**χ_1^2=4.0**	**p=0.03**	22.1	23.0	t=0.9	p=0.66
	IV drug use								
	Yes	76.5	66.7	χ_1^2=0.4	--	18.7	19.8	t=0.3	--
	No	69.0	64.9	χ_1^2=0.8	p=0.72	22.6	23.1	t=0.5	p=0.86
	Prior treatment								
	Yes	76.7	68.2	χ_1^2=2.4	--	21.7	23.1	t=1.1	--
	No	60.5	58.7	χ_1^2=0.1	p=0.40	22.7	22.5	t=-0.1	p=0.46
CTO0001 (Reserpine)									
	Sex								
	Men	67.4	68.3	χ_1^2=0.0	--	20.6	23.2	t=1.3	--
	Women	64.7	55.6	χ_1^2=0.3	p=0.61	16.8	19.1	t=0.6	p=0.92
	Race								
	White	61.1	44.4	χ_1^2=0.7	--	20.6	29.5	t=1.3	--
	Non-White	69.3	69.4	χ_1^2=0.0	p=0.47	18.9	21.8	t=1.4	p=0.36
	Age								
	< 35	66.7	46.7	χ_1^2=1.1	--	11.4	20.7	t=2.1	--
	≥ 35	66.7	70.5	χ_1^2=0.2	p=0.28	21.6	22.5	t=0.5	p=0.05
	Education								
	< 12 years	62.5	70.0	χ_1^2=0.1	--	22.2	22.1	t=0.0	--
	≥ 12 years	67.3	63.3	χ_1^2=0.2	p=0.64	19.2	22.2	t=1.5	p=0.56
	Employment								
	Full-time	54.6	66.7	χ_1^2=0.5	--	18.2	25.0	**t=2.3**	--
	Other	73.7	63.6	χ_1^2=1.0	p=0.25	20.1	21.2	t=0.5	p=0.15
	Marital status								
	Married	46.2	70.0	χ_1^2=1.3	--	17.5	22.3	t=1.1	--
	Non-married	71.7	63.3	χ_1^2=0.8	p=0.16	20.1	22.2	t=1.0	p=0.59
	IV drug use								
	Yes	100.0	80.0	χ_1^2=0.7	--	25.0	17.8	t = -1.2	--
	No	64.9	63.0	χ_1^2=0.0	--[1]	19.1	22.7	t=1.9	p=0.10
	Prior treatment								
	Yes	67.7	67.6	χ_1^2=0.0	--	19.5	23.3	t=1.7	--
	No	65.5	57.1	χ_1^2=0.4	p=0.67	19.6	20.2	t=0.2	p=0.39
MDS0004 (Modafinil)									
	Sex								
	Men	69.0	58.8	χ_1^2=1.5	--	22.9	22.0	t = -0.6	--
	Women	76.3	71.4	χ_1^2=0.2	p=0.79	21.2	19.8	t = -0.5	p=0.89
	Race								
	White	73.0	62.5	χ_1^2=0.7	--	23.7	23.2	t = -0.2	--
	Non-White	70.3	61.7	χ_1^2=1.1	p=0.88	21.9	20.2	t = -1.0	p=0.70
	Age								

	< 35	59.1	50.0	χ_1^2=0.2	--	20.3	24.8	t=0.9	--
	≥ 35	73.3	64.1	χ_1^2=1.7	p=0.94	22.7	20.9	t = -1.2	p=0.19
	Education								
	< 12 years	76.5	60.0	χ_1^2=0.5	--	20.6	25.7	t=1.1	--
	≥ 12 years	70.3	62.7	χ_1^2=1.1	p=0.70	22.7	21.0	t = -1.2	p=0.20
	Employment								
	Full-time	62.2	47.8	χ_1^2=1.3	--	22.1	21.3	t = -0.5	--
	Other	75.3	69.4	χ_1^2=0.6	p=0.66	23.3	21.2	t = -0.7	p=0.68
	Marital status								
	Married	70.6	76.2	χ_1^2=0.2	--	23.3	23.7	t=0.2	--
	Non-married	71.2	56.9	χ_1^2=3.1	p=0.21	22.1	19.9	t = -1.2	p=0.40
	IV drug use								
	Yes	80.0	33.3	χ_1^2=2.7	--	15.2	18.0	--[9]	--
	No	69.9	63.8	χ_1^2=0.8	p=0.21	23.4	21.3	t = -1.5	p=0.54
	Prior treatment								
	Yes	79.3	63.4	χ_1^2=3.6	--	21.7	19.2	t = -1.3	--
	No	58.9	61.3	χ_1^2=0.0	p=0.15	23.7	24.1	t=0.2	p=0.32
CTN0052 (Buspirone)									
	Sex								
	Men	95.8	80.0	χ_1^2=2.5	--	24.0	26.3	t=0.9	--
	Women	81.8	91.7	χ_1^2=0.5	p=0.14	26.7	27.6	t=0.5	p=0.72
	Race								
	White	88.9	71.4	χ_1^2=0.8	--	21.4	28.2	t=1.5	--
	Non-White	92.3	90.0	χ_1^2=0.1	p=0.61	25.9	26.6	t=0.4	p=0.13
	Age								
	< 35	100.0	0.0	**χ_1^2=4.0**	--	29.3	--[10]	--[10]	--
	≥ 35	90.6	88.5	χ_1^2=0.1	--[2]	24.3	26.9	t=1.5	--[10]
	Education								
	< 12 years	92.9	100.0	χ_1^2=0.6	--	23.5	29.0	**t=2.4**	--
	≥ 12 years	90.5	79.0	χ_1^2=1.0	--[3]	25.6	25.8	t=0.1	p=0.13
	Employment								
	Full-time	85.7	33.3	χ_1^2=3.7	--	29.0	26.5	t = -1.3	--
	Other	92.8	100.0	χ_1^2=1.6	--[4]	23.8	27.0	t=1.7	p=0.29
	Marital status								
	Married	66.7	60.0	χ_1^2=0.0	--	16.5	20.7	t=1.0	--
	Non-married	93.8	90.9	χ_1^2=0.2	p=0.95	25.3	27.9	t=1.5	p=0.78
	IV drug use								
	Yes	100.0	100.0	--[5]	--	27.5	30.0	--[9]	--
	No	90.9	84.6	χ_1^2=0.6	--[5]	24.6	26.8	t=1.2	p=0.97
	Prior treatment								
	Yes	91.4	85.2	χ_1^2=0.6	--	24.8	26.9	t=1.3	--
	No	--[6]	--[6]	--[6]	--[6]	--[6]	--[6]	--[6]	--[6]
CTO0005 (Ondansetron)									
	Sex								
	Men	56.1	50.0	χ_1^2=0.2	--	21.2	23.3	t=0.5	--
	Women	50.0	0.0	χ_1^2=1.7	--[7]	23.0	--[10]	--[10]	--[10]
	Race								
	White	47.6	25.0	χ_1^2=0.7	--	22.3	28.0	--[9]	--
	Non-White	60.7	50.0	χ_1^2=0.4	p=0.69	20.9	22.5	t=0.4	p=0.69
	Age								
	< 35	58.6	50.0	χ_1^2=0.2	--	24.4	29.0	t=1.4	--
	≥ 35	50.0	33.3	χ_1^2=0.5	p=0.78	19.7	21.0	t=0.2	p=0.68
	Education								
	< 12 years	55.0	46.2	χ_1^2=0.3	--	15.8	4.0	--[9]	--
	≥ 12 years	55.6	33.3	χ_1^2=0.4	p=0.72	22.7	26.5	t=1.1	p=0.10
	Employment								
	Full-time	50.0	25.0	χ_1^2=1.6	--	19.9	26.3	t=1.4	--
	Other	61.9	62.5	χ_1^2=0.0	p=0.72	22.9	16.0	t = -1.0	p=0.10
	Marital status								

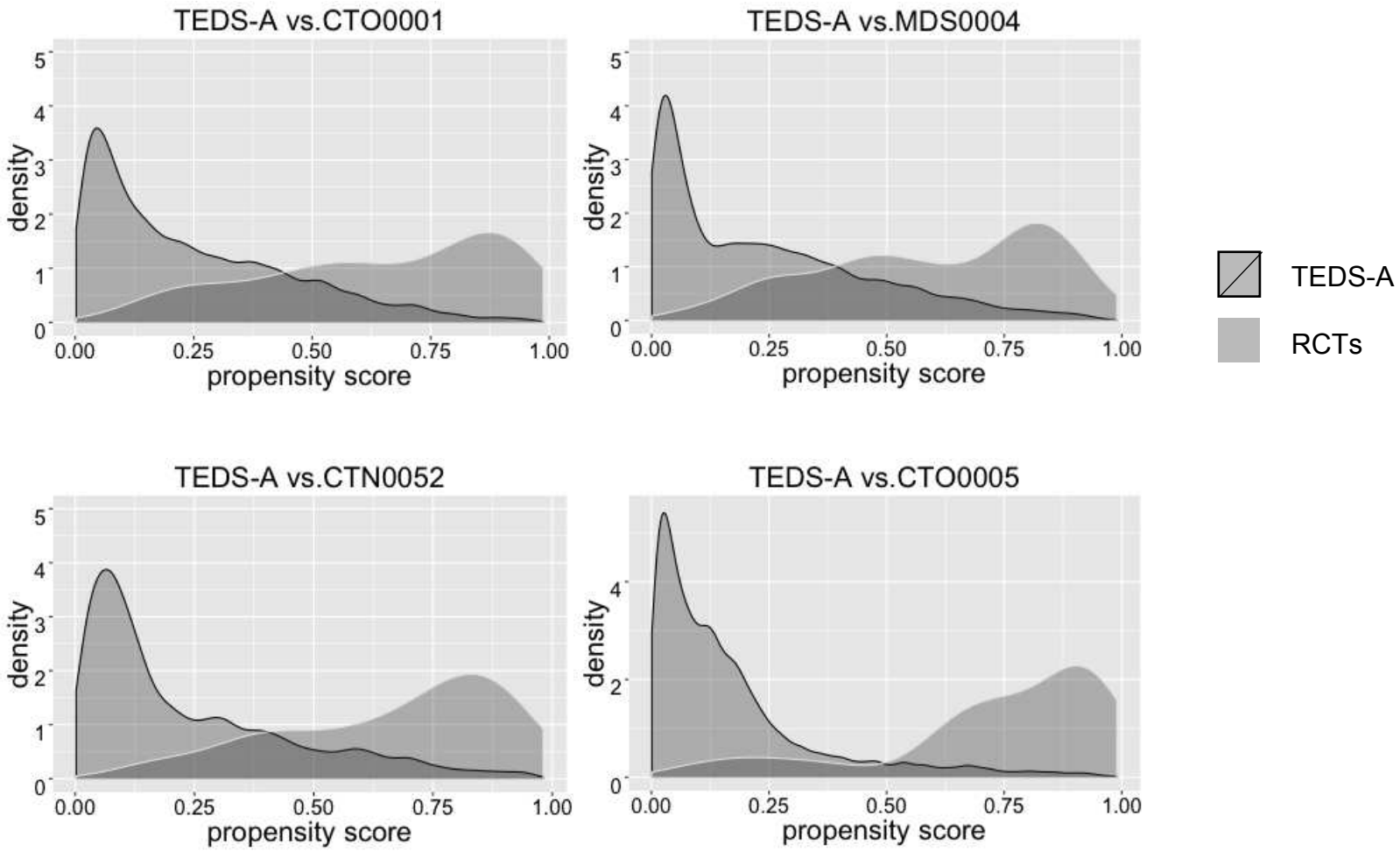
	Married	28.6	50.0	χ_1^2=0.8	--	27.5	17.3	t = -1.7	--
	Non-married	65.7	40.0	χ_1^2=2.1	p=0.11	20.4	27.8	t=1.6	**p=0.03**
	IV drug use								
	Yes	--[8]	64.3	--[8]	--	19.4	--[10]	--[10]	--
	No	51.3	43.8	χ_1^2=0.3	--[8]	22.4	23.3	t=0.2	--[10]
	Prior treatment								
	Yes	52.6	20.0	χ_1^2=1.7	--	16.9	28.0	--[9]	--
	No	58.6	54.6	χ_1^2=0.1	p=0.35	24.1	22.5	t = -0.5	p=0.19

1: All the participants who had IV drug use in the treatment arm successfully retained in the study and it was not possible to estimate a coefficient for interaction term.
2: All the participants aged <35 years old in the treatment arm successfully retained in the study and it was not possible to estimate a coefficient for interaction term.
3: All the participants with less than 12 years of education in the control arm successfully retained in the study and it was not possible to estimate a coefficient for interaction term.
4: All the participants without fulltime job in the control arm successfully retained in the study and it was not possible to estimate a coefficient for interaction term.
5: All the participants who had IV drug use successfully retained in the study and it was not possible to calculate chi-squared statistics and to estimate a coefficient for interaction term.
6: There was no observation of those who had no prior treatments.
7: No female participants in the control arm successfully retained in the study and it was not possible to estimate a coefficient for interaction term.
8: All the participants who had IV drug use were in the treatment arm and it was not possible to calculate chi-squared statistics and to estimate a coefficient for interaction term.
9: There was only one observation in the control arm and it was not possible to calculate t-statistics.
10: There was no observation in the control arm and it was not possible to calculate t-statistics and to estimate a coefficient for interaction term.

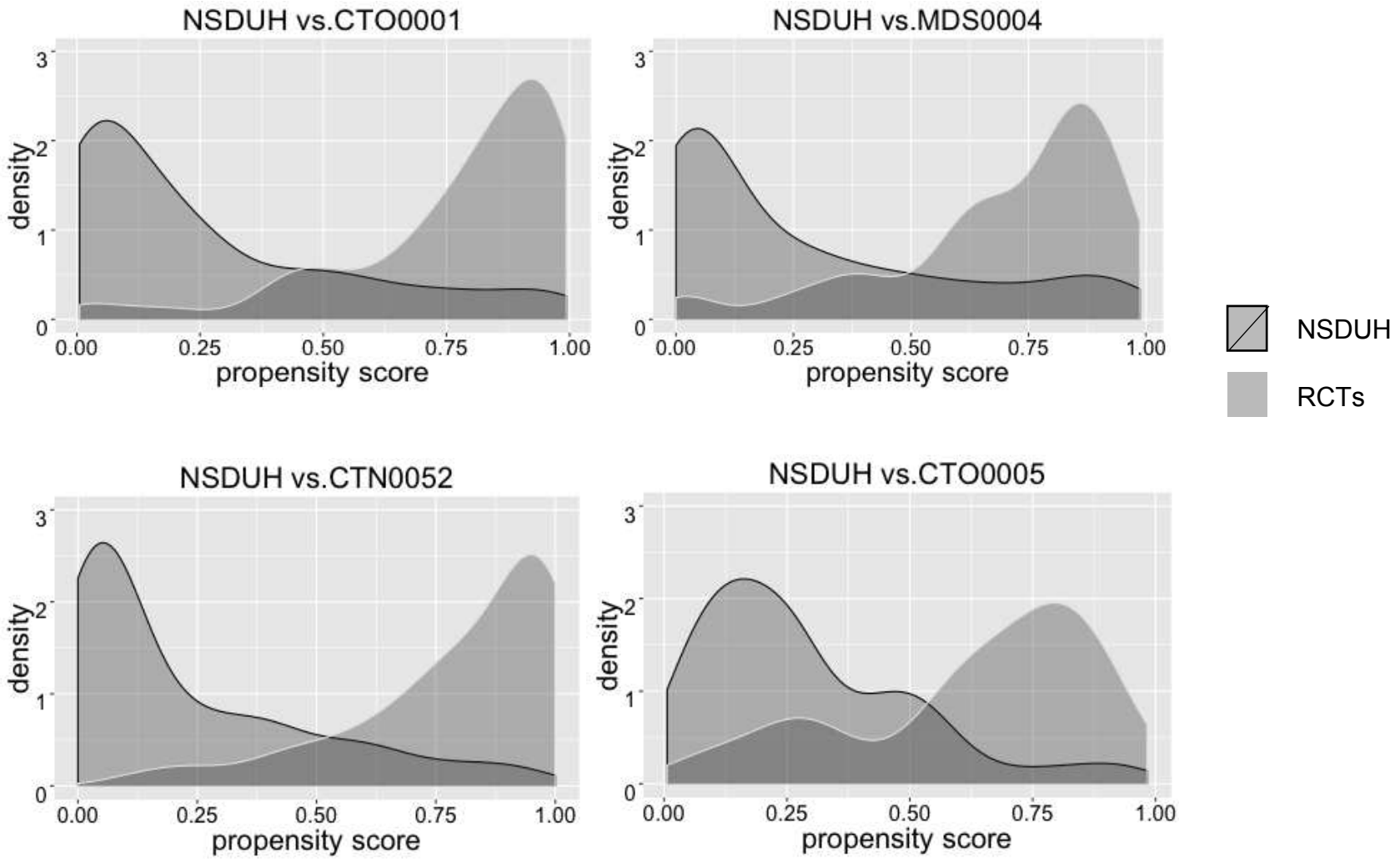
Appendix Table 4.3. Comparison of propensity scores between the RCT samples and target samples from the Treatment Episodes Data-Admission (TEDS-A) and the National Survey of Drug Use and Health (NSDUH).

Target population = TEDS-A								
Study number	Intervention	RCT	TEDS-A	Δp[a]	Pooled standard deviation	Standard Δp[b]	*t*-Test	P-value
CTO0001	Reserpine	0.64	0.25	0.39	0.21	1.86	20.41	<0.001
MDS0004	Modafinil	0.60	0.24	0.36	0.22	1.61	23.44	<0.001
CTN0052	Buspirone	0.66	0.24	0.42	0.22	1.94	15.30	<0.001
CTO0005	Ondansetron	0.73	0.17	0.56	0.19	3.01	24.44	<0.001
Target population = NSDUH								
Study number	Intervention	RCT	NSDUH	Δp	Pooled standard deviation	Standard Δp	*t*-Test	P-value
CTO0001	Reserpine	0.73	0.24	0.49	0.35	1.40	16.61	<0.001
MDS0004	Modafinil	0.74	0.31	0.43	0.36	1.22	16.25	<0.001
CTN0052	Buspirone	0.80	0.23	0.57	0.34	1.68	16.05	<0.001
CTO0005	Ondansetron	0.61	0.31	0.30	0.25	1.20	9.86	<0.001

[a] Δp is difference between propensity scores of the RCT sample and the target population.
[b] Standardized Δp is computed as Δp divided by pooled standard deviation.



Appendix Figure 4.1. Density plots of propensity scores in RCT samples and target sample from the Treatment Episodes Data-Admission (TEDS-A)



Appendix Figure 4.2. Density plots of propensity scores in RCT samples and target sample from the National Survey of Drug Use and Health (NSDUH)

CURRICULUM VITAE

Ryoko Susukida

May 2017

Johns Hopkins Bloomberg School of Public Health

624 N. Broadway, Room 897, Baltimore MD 21205

Email: rsusuki1@jhu.edu

Date of Birth: December 15, 1987 Okayama, JAPAN

EDUCATION

Doctor of Philosophy

- Public Economics, National Graduate Institute for Policy Studies, Tokyo, JAPAN, September 2013
- Dissertation title: An Inquiry into Mental Health and Help Seeking Behaviors in Japan

Master of Arts

- Public Economics, National Graduate Institute for Policy Studies, Tokyo, JAPAN, 2012

Bachelor of Arts

- Economics, the University of Tokyo, Tokyo, JAPAN, 2010

WORK EXPERIENCE

Research Fellow December 2016-Present (Supervisor: Dr. Daisuke Nishi)

National Center of Neurology and Psychiatry, National Institute of Mental Health, Tokyo, Japan

- Assess generalizability of the findings of randomized clinical trials of omega-3 supplements for preventing/treating depressing among pregnant women in Japan

Research Analyst August 2015-Present (Supervisor: Dr. Holly C. Wilcox)

Department of Health, Behavior & Society, Johns Hopkins Bloomberg School of Public Health, Baltimore, United States

- An objective description of the state of the science on data linkage strategies and analytic approaches in suicide prevention research in adolescents
- Provide a systematic summary of ongoing research and research needs to serve as the foundation for an NIH Pathway to Prevention workshop

Research Analyst June 2014-Present (Supervisor: Dr. Ramin Mojtabai)

Department of Mental Health, Johns Hopkins Bloomberg School of Public Health, Baltimore, United States

- Generalizability study of multi-site clinical studies in the United States
- Data organization, statistical analyses and manuscript preparation

Research Intern August 2011-October 2011 (Supervisor: Dr. Yumiko Aratani)

National Center for Children in Poverty (NCCP), the Mailman School of Public Health, Columbia University, New York, USA

- Literature review about post-disaster mental health
- Descriptive data analysis about inter-ethnic differences in mental health service utilization among Asian children and adolescents in California State

Graduate Teaching Assistant Spring 2011, 2012 (Supervisor: Prof. Yosuke Yasuda)

Policy Analysis Program, National Graduate Institute for Policy Studies, Tokyo, JAPAN

- Supplementary classes of Advanced Microeconomics I, II (Ph.D. core courses) to approximately fifteen students per course

RESEARCH SKILLS

Cross sectional and panel data analysis using STATA

Propensity score matching using R

SCHOLARSHIPS AND GRANTS

Global Health Leadership Program Scholarship ($15000), August 2011-October 2011

Global Health Policy, Graduate School of Medicine, the University of Tokyo, JAPAN

- Given to the selected doctoral students (ten or less) in order to complete internship in the field of global health, which is a requirement for the completion of Global Health Leadership Program

GRIPS Tuition Scholarship (Tuition waiver and $1000 monthly), 2010-2013

National Graduate Institute for Policy Studies, Tokyo, JAPAN

- This fellowship is given to the selected Ph.D. candidates in recognition of outstanding academic performance

Fulbright Scholarship (Tuition and monthly stipend), August 2013- August 2016

The Lucy Shum Memorial Scholarship Award (Tuition support, $5545), 2015

- This award is made annually to a doctoral student who has demonstrated excellence in public health for mental health issues

R01 DA036520 Mojtabai (PI) (Total award: $722,925), March 2014 **–**

- **Research Analyst** - Generalizing RCT Efficacy Evidence: Application to NIDA Clinical Trials Network

PRESENTATIONS

"The association of lifetime suicidal ideation with perceived parental love and family structure in childhood in a nationally representative adult sample," IASR/AFSP International Summit on Suicide Research, October 13th 2015, New York, USA

"Lifetime suicidal ideation and perceived parental love in childhood: a cross-sectional study in the National Comorbidity Survey Replication," 15th International Mental Health Conference, August 27th 2014, Gold Coast, Australia

"Relationship between frequent family relocation and risk of depression and substance use among children in the National Survey on Drug Use and Health," Psychiatry and Behavioral Sciences Annual Research Potpourri, May 27th 2014, Baltimore, United States

"Counter-cyclicality of suicide and its potential sources: Evidence from Japanese prefecture data," CAMPUS Asia Ph.D. Seminar, March 14th 2012, KDI School, Seoul, the Republic of Korea

PUBLICATION

Kaufmann, C. N., **Susukida, R.**, Depp, C. A. (2017). "Sleep apnea, psychopathology, and mental health care" *Sleep Health*, Forthcoming.

Susukida, R., Crum, R.M., Stuart, E. A., Ebnesajjad, C., Mojtabai, R. (2017). "Generalizability of Findings from Randomized Controlled Trials: Application to the National Institute of Drug Abuse Clinical Trials Network" Addiction, Forthcoming.

Wilcox, H. C., Kharrazi, H., Wilson, R. F., Musci, R. J., **Susukida, R.**, Gharghabi, F., Zhang, A., Wissow, L., Robinson, K. A. (2016). "Data Linkage Strategies to Advance Youth Suicide Prevention: A Systematic Review for a National Institutes of Health Pathways to Prevention Workshop" Annals of Internal Medicine, 165(11), 779-785.

Susukida, R., Crum, R.M., Stuart, E. A., Ebnesajjad, C., Mojtabai, R. (2016). "Assessing Sample Representativeness in Randomized Control Trials: Application to the National Institute of Drug Abuse Clinical Trials Network" Addiction, 111(7), 1226-1234.

Susukida, R., Wilcox, H. C., Mendelson, T. (2016). "The association of lifetime suicidal ideation with perceived parental love and family structure in childhood in a nationally representative adult sample" Psychiatry Research, 237, 246-51.

Susukida, R., Mojtabai, R., Guillermo M., Mendelson, T. (2015). "Residential mobility and risk of major depressive episode among adolescents in the National Survey on Drug Use and Health" Journal of Public Health, 38(3), 432-440.

Mojtabai R., Stuart E. A., Hwang, I., **Susukida, R.**, Eaton, W. W., Sampson, N., Kessler, R. C. (2015). "Long-term effects of mental disorders on employment in the National Comorbidity Survey ten-year follow-up" Social psychiatry and psychiatric epidemiology, 50(11), 1657-1668.

Susukida, R., Mojtabai, R. Mendelson, T. (2015). "Sex differences in help seeking for mood and anxiety disorders in the National Comorbidity Survey-Replication" Depression and Anxiety, 32(11), 853-60.

Susukida, R. "A Prefecture-Level Panel Data Analysis of Mental Health Problems and Suicide" (in Japanese) 2014. Nihon Keizai Kenkyu (Japan Economic Journal).

PEER REVIEWER EXPERIENCE

Stigma and Health (American Psychological Association)

Psychology Research and Behavior Management

Drug and Alcohol Dependence