Published in final edited form as:

Addict Behav. 2019 July ; 94: 124–132. doi:10.1016/j.addbeh.2018.10.033.

Implementing Statistical Methods for Generalizing Randomized Trial Findings to a Target Population

Benjamin Ackerman^{1,*}, Ian Schmid², Kara E. Rudolph³, Marissa J. Seamans², Ryoko Susukida⁴, Ramin Mojtabai², and Elizabeth A. Stuart^{1,2}

¹Department of Biostatistics, The Johns Hopkins Bloomberg School of Public Health

²Department of Mental Health, The Johns Hopkins Bloomberg School of Public Health

³Emergency Medicine, School of Medicine, University of California, Davis

⁴Department of Mental Health Policy, National Center of Neurology and Psychiatry, Japan

Abstract

Randomized trials are considered the gold standard for assessing the causal effects of a drug or intervention in a study population, and their results are often utilized in the formulation of health policy. However, there is growing concern that results from trials do not necessarily generalize well to their respective target populations, in which policies are enacted, due to substantial demographic differences between study and target populations. In trials related to substance use disorders (SUDs), especially, strict exclusion criteria make it challenging to obtain study samples that are fully “representative” of the populations that policymakers may wish to generalize their results to. In this paper, we provide an overview of post-trial statistical methods for assessing and improving upon the generalizability of a randomized trial to a well-defined target population. We then illustrate the different methods using a randomized trial related to methamphetamine dependence and a target population of substance abuse treatment seekers, and provide software to implement the methods in R using the “generalize” package. We discuss several practical considerations for researchers who wish to utilize these tools, such as the importance of acquiring population-level data to represent the target population of interest, and the challenges of data harmonization.

*Corresponding Author: backer10@jhu.edu, 615 N. Wolfe Street, E3031, Baltimore, MD 21205, USA.

Contributors: BA, EAS, RS and RM contributed to the original conceptualization of the project. BA and EAS conducted a literature search. RS, BA and EAS acquired the trial and target population data for the data example. BA, KR, IS and MRS contributed to the statistical analysis of the data example. BA and IS contributed to the development of the R package for implementing the statistical methods. BA wrote the initial draft of the manuscript. All authors have contributed to and approved of the final manuscript.

Conflict of Interest: All authors declare that they have no conflicts of interest.

Data Statement

Due to the terms of the NIDA-CTN Data Use Agreement, which state that “the recipient of the data agrees... to retain control over the received data, and not to transfer any portion of the received data, with or without charge, to any other entity or individual”, the authors are unable to share the data used in this manuscript.

Publisher's Disclaimer: This is a PDF file of an unedited manuscript that has been accepted for publication. As a service to our customers we are providing this early version of the manuscript. The manuscript will undergo copyediting, typesetting, and review of the resulting proof before it is published in its final citable form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

Keywords

generalizability; external validity; randomized trials; statistics

1. Introduction

Randomized controlled trials (RCTs) are considered the gold standard for estimating the average causal effect of a drug or intervention in a study sample. Experimental study designs allow researchers to study the treatment of interest under highly controlled and ideal circumstances, and the randomization of treatment assignment removes confounding, providing strong internal validity. RCTs often have great influence on evidence-based decisions, particularly in the presence of conflicting study results (Weisberget al., 2009). However, while RCTs have strong internal validity, they often have weaker external validity, making it difficult to generalize trial results from a “non-representative” study sample to a broader population (Imai et al., 2008; Shadish et al., 2002). In particular, when the distribution of a factor that modifies treatment effects in the trial differs from the distribution of that factor in the population, the sample average treatment effect (SATE) will not equal the target population average treatment effect (TATE) (Cole & Stuart, 2010; Lesko et al., 2017). This makes it challenging for policymakers to accurately draw population-level conclusions from trial evidence.

Differences between the sample and population may be particularly pronounced in studies of substance abuse treatment. Susukida et al. (2016) documented prominent differences between substance use disorder (SUD) treatment-related trial participants and a population of SUD treatment seekers across ten trials supported by the National Drug Abuse Treatment Clinical Trials Network (NIDA-CTN). Most of those 10 trials studied the effectiveness of buprenorphine/naloxone (Bup/Nx-Detox) detoxification for opioid dependence, and Susukida et al. (2016) found that the SUD trial participants were more likely to have more than 12 years of education, be employed full time, and to have had a greater number of prior treatments than the general population of SUD treatment seekers. Some of these factors have been associated with more positive attitudes towards SUD treatment (Moradveisi et al., 2014), which may lead to different levels of adherence and thus different effectiveness of the interventions. Therefore, differences in these covariates between the trial samples and populations could lead to limited generalizability. When generalized to the target population, Susukida et al. (2017) found that most significant trial results became statistically insignificant, a shift that could be attributed largely to treatment effect heterogeneity. The issue of generalizability has been discussed across many other disciplines as well, such as medicine (Rubin, 2008), social work (Stuart et al., 2017; Zhai et al., 2010), and child development (Dababnah & Parish, 2016), reinforcing the importance of developing guidelines and methods for handling the poor external validity of RCTs.

Given increasing concern about potential lack of generalizability of RCT findings, statistical methods have recently been proposed to estimate population average treatment effects using RCT and population data. While thinking about generalizability is important throughout the study design and implementation processes (Flay, 1986; Insel, 2006; Kern et al., 2016; Peto

et al., 1995), these methods are meant to be implemented after the study is already conducted. In this paper, we provide an introductory overview of several posttrial statistical methods to generalize average treatment effects to a well-defined target population. These methods rely on the existence of individual-level data for the target population, or a representative sample of it (Stuart et al., 2011). The paper proceeds as follows: Section 2 describes the notation and assumptions. Section 3 describes methods for assessing and improving upon the generalizability of RCT findings. Section 4 provides guidance for preparing data and implementing the described methods using our R package, "generalize." We illustrate the use of "generalize" in Section 5 using data from an RCT related to methamphetamine dependence and a nationally-representative survey of SUD treatment admissions. Finally, Section 6 discusses factors that researchers should take into consideration when defining target populations and implementing the appropriate methods, as well as some limitations and areas for future research.

2. Causal Effects, Notation and Assumptions

Suppose a trial of n participants is conducted, and researchers are interested in generalizing the trial results to a well-defined target population of size N . Define S to be an indicator of trial membership: $S_i = 1$ indicates that individual i is in the trial, while $S_i = 0$ indicates that they are in the population but not a trial participant. Note that since we are discussing generalizability, S simply indicates trial membership, and all individuals in the trial are still considered to come from the target population of interest, even when the trial and population data sets are disjoint. If the study sample is totally separate from the target population, (e.g. a trial is conducted in a sample in Los Angeles and researchers wish to extrapolate its findings to a population in New York) then it becomes a matter of transportability instead of generalizability (Lesko et al., 2017; Pearl & Bareinboim, 2011).

Let Y denote the outcome of interest, $Y_i(1)$ denote the potential outcome for subject i under treatment, and $Y_i(0)$ denote the potential outcome for subject i under control. The causal effect for an individual is defined as the difference in potential outcomes under treatment and control conditions, $Y_i(1) - Y_i(0)$ (Rubin, 1974). The challenge of causal inference is that, in practice, it is not possible to observe both $Y_i(1)$ and $Y_i(0)$ for individual i , as, at any particular point in time, each individual receives either treatment or control, not both. It is therefore common to estimate the *average* treatment effect (ATE), defined as the mean over the individual level causal effects (Kern et al., 2016). The sample average treatment effect (SATE) is defined as

$$SATE = E[Y(1) - Y(0) \mid S = 1],$$

and can be unbiasedly estimated in an RCT by $\frac{1}{n} \sum_{i=1}^n (Y_i(1) - Y_i(0) \mid S_i = 1)$. However, the estimand of interest here is the target population average treatment effect (TATE), which is defined as

$$TATE = E[Y(1) - Y(0) \mid S = 0]$$

If the data were available, this could be estimated by $\frac{1}{N} \sum_{i=1}^N (Y_i(1) - Y_i(0) | S_i = 0)$. Since the intervention is assumed be unavailable to the population at the time of the trial, outcomes under treatment are not observed in the population and therefore this quantity can not be calculated directly. This challenge motivates the generalizability methods presented in Section 3.

In addition to the common structural assumptions required for randomized trials' internal validity, several additional assumptions are needed when estimating the TATE using data from a trial and a target population. We assume the following:

- A1)** All members of the target population have a nonzero probability of being selected for the trial.
- A2)** There are no unmeasured variables associated with sample selection and treatment effect given the observed variables.
- A3)** When considering the set of pre-treatment covariates associated with treatment effect, the ranges of such effect modifiers in the target population are covered by their respective ranges in the trial.
- A4)** In the trial, treatment assignment is independent of sample selection, as well as of potential outcomes, given the pre-treatment covariates.

A1 is similar to the positivity assumption for drawing causal inference in non-experimental studies. A2 is comparable to the assumption of "unconfounded treatment assignment" in non-experimental studies. This is a strong assumption that is unrealistic in some settings; while a trial may measure all variables related to treatment effect, a data set representing the target population may be limited in the scope of variables measured. A3 regarding coverage should be highly considered when defining the target population. For example, if the age range in a trial is 18-30, there is no evidence from the trial to estimate the population average treatment effect for 50 year olds. A4 is satisfied by nature of the randomization in RCTs.

3. Methods

In this section, we first describe the probability of trial participation and its use, then we discuss how to assess the generalizability of a trial, followed by an overview of several methods for estimating the population average treatment effect.

3.1. Probability of Trial Participation

Traditionally used in non-experimental studies for assessing balance between treatment groups and for matching (Rosenbaum & Rubin, 1983; Rubin, 2001), propensity score-type methods are also highly useful for generalizability. Here, they are used to model the probability of trial sample membership based on a set of baseline covariates. The probabilities are then used to assess differences between the trial sample and the population (Section 3.2), and also to construct weights to estimate the TATE (Section 3.3.1).

Trial participation probabilities can be estimated using several methods; here, we focus on three: logistic regression, Random Forests, and Lasso. Estimation using logistic regression involves specifying a sample selection model based on a linear combination of the pretreatment covariates of interest and then obtaining the predicted values. Random Forests are a decision tree-based regression method that have shown good performance for propensity score estimation (Lee et al., 2010). Lasso is a penalty approach that places constraints on the model coefficients and aids in model selection by allowing certain coefficients to shrink to zero (Tibshirani, 1996). Both Random Forests and Lasso are quite flexible models of trial membership and do not require specification of the specific model form. All three of these estimation methods are supported by the statistical package described in the Appendix.

3.2. Assessing the Generalizability of a Trial

Prior to generalizing existing study results to a target population, it is important to assess how similar or different the study sample is to the target population.

One way to do so is to calculate the absolute standardized mean difference (ASMD) of each covariate between the trial sample and target population. A larger ASMD indicates greater differences between the covariate distribution in the trial and the population, whereas a smaller ASMD indicates that the trial is more similar to the population on that factor. As detailed further below, this metric can also be used to help assess the success of the trial weighting methods described below. However, while this method may reveal covariate-by-covariate differences, it does not assess the joint distribution of the covariates.

Another metric of similarity is a generalizability index proposed by Tipton (2014), which utilizes the trial participation probabilities and therefore captures differences between all of the observed covariates at once. Tipton's generalizability index functions like a "histogram distance" to describe how similar a trial sample is to a random sample drawn from the target population. The index, β , is defined as follows:

$$\beta = \int \sqrt{f_s(s)f_s(p)} ds$$

where $f_s(s)$ and $f_s(p)$ are the distributions of trial participation probabilities in the trial sample and target population given a set of common covariates, respectively. Estimation of β involves binning the trial and population data based on the distribution of their trial participation probabilities and comparing the proportions of each data set that fall within each bin. Tipton's generalizability index has several appealing properties: it is bounded between 0 and 1, does not require any distributional assumptions and has an informative magnitude. An index of 1 signifies that the trial sample is like a random sample drawn from the target population, whereas an index of 0 indicates no overlap between the trial population. Typically, samples with indices greater than .8 are considered highly similar to the population, whereas indices less than .5 are considered dissimilar (Tipton, 2014), which may inform whether generalizing the study results to that target population is appropriate at all.

3.3. Estimating Population Treatment Effects

After assessing differences between the trial and population, there are several approaches for estimating the TATE. We now detail three broad classes of methods for estimating the TATE: one set based on using the probability of trial participation to equate the trial sample and population, one set based on flexible outcome models used to predict outcomes in the population, and a third that combines both together.

3.3.1. Weighting by the Inverse Odds of Trial Participation—One proposed method weights the trial sample by the inverse odds of trial participation, which assigns greater weight to individuals in the trial with greater probability of being in the target population. In doing so, this approach weights the sample to be more similar to the target population. This is similar to the construction of ATT weights using propensity scores in non-experimental settings (Stuart, 2010). The weights are defined as follows:

$$w_i = \begin{cases} 0 & \text{if } S_i = 0 \\ \frac{1 - \hat{e}_i}{\hat{e}_i} & \text{if } S_i = 1 \end{cases}$$

where $\hat{e}_i$ is defined as the predicted probability of individual i being a trial participant, and can be calculated using the methods described in Section 3.1. The TATE is then estimated by fitting a weighted least squares regression model using the trial data (Kern et al., 2016).

3.3.2. Outcome Model Based Approach—Another set of approaches estimate the TATE by modeling the outcome in a flexible way. Machine learning algorithms have become increasingly popular in estimating causal effects, as, compared to parametric regression models, they implement more flexible models that do not require linearity or additivity assumptions (Kern et al., 2016). Bayesian Additive Regression Trees (BART) is one such algorithm that has been used to estimate treatment effects (J. L. Hill, 2011). The algorithm operates as a “sum of trees,” fitting many regression models that each have a small contribution to the overall model. In the context of generalizability, BART is used to fit the outcome model on the trial data and then estimate the TATE by predicting outcomes under treatment and control in the target population. Draws from the posterior distribution of the individual causal effects are then averaged across the population data set to obtain the TATE estimate (Kern et al., 2016).

3.3.3. Combining weighting and outcome modeling: TMLE—Lastly, Targeted Maximum Likelihood Estimation (TMLE) is a method that combines both strategies. It models both the outcome and the trial participation using pre-treatment covariates, and is robust to whether or not one of those models is incorrect (Gruber & Van Der Laan, 2009; Rudolph et al., 2014). In the generalizability context, the outcome model is first used to predict outcomes under treatment conditions in both the trial and population data, which are then essentially offset by a function of the participation probabilities, generated by the selection model. The updated predicted outcomes in the full data are then used to estimate the TATE.

4. Preparing Data for Method Implementation

In order to implement the methods described in Section 3, several data pre-processing steps must be taken. First, it is important to identify a data set that describes the target population of interest and measures an overlapping set of covariates with the trial data that may impact treatment effect heterogeneity and/or trial membership.

Next, trial and population data must be harmonized across that common set of covariates. This may involve categorizing or dichotomizing certain variables across data sources to make measures comparable. It may be useful to identify which data source has fewer variables, and then try to find the maximal overlap with the variable list of the more detailed data source. Data on outcomes and treatment will be missing in the population data set and should be coded as such. The final combined "stacked" data should contain variables for outcomes and treatment in the trial that are missing in the population, a binary indicator for trial participation to distinguish those enrolled in the RCT from those who are not, and the set of overlapping covariates (see Figure 1).

Once the data are formatted in this manner, the methods described in Section 3 can be implemented using "generalize," a package developed for statistical software R (R Core Team, 2017). Currently available on Github (Ackerman, 2018), "generalize" allows researchers to assess and generalize trial findings to a well-defined target population (see Appendix for code).

5. Data Example

We now apply the methods discussed to a trial related to methamphetamine dependence. Trial data were obtained from the CSP-1025 trial of the NIDA-CTN data repository (Johnson, 2015). The phase 2, multi-site, placebo-controlled RCT aimed to determine if topiramate, a therapeutic shown to reduce alcohol and cocaine use (Johnson et al., 2007; Kampman et al., 2004), could reduce methamphetamine use relative to placebo in individuals with methamphetamine dependence. 140 participants were randomized to either topiramate or placebo. For this illustrative example, the outcome of interest is methamphetamine use reported during follow-up. No significant differences between treatment groups were found for this outcome in the initial report of the trial (Elkashef et al., 2012).

Data from the Treatment Episode Data Set: Admissions (TEDS-A) of 2014 were used to represent the population of substance abuse treatment seekers. Managed by the Substance Abuse and Mental Health Services Administration (SAMHSA), TEDS-A consists of annual data regarding all publicly-funded admissions to substance abuse treatment programs in the United States, as required by state law. For better relevance to the CSP-1025 trial, we subset TEDS-A to only include records where methamphetamine was listed as the primary substance abuse problem at time of admission, resulting in 135,264 records in the population dataset.

Eight common covariates were identified across the trial and target population data sets: age, sex, race, ethnicity, marital status, education, employment status and prior

methamphetamine use in the past week. To ensure that measures across each data set were comparable, variables were categorized and dichotomized when needed. For example, the binary variable indicating any prior methamphetamine use in the past week was determined by a variable in the trial that measured the actual number of days of methamphetamine use in the month prior to the study, and a categorical variable in TEDS-A that reported either 1) no methamphetamine use in the past month, 2) 1-3 times in the past month, 3) 1-2 times in the past week, 4) 3-6 times in the past week, or 5) daily.

5.1. Results

Table 1 describes the distributions of pre-treatment covariates in the trial sample and in the target population. The trial sample was older, more predominantly male, less racially and ethnically diverse, and more educated than the target population of individuals in treatment for methamphetamine dependence. A larger proportion of trial participants reported using methamphetamine in the prior seven days than did individuals in the target population. Since none of the trial participants were between the ages of 12 and 15, members of the target population in that age range were excluded from the target population to avoid violating the coverage assumption (Assumption A3).

The distributions of the log(trial participation probabilities) in the trial and target population varied somewhat by method of calculation as well (Figure 2). Here, probabilities calculated using logistic regression depicted greater differences between the trial and target population, while probabilities calculated using Lasso and Random Forests suggested that the trial was slightly more similar to the target population.

The TATE estimates are shown in Figure 3. In the trial sample, there was no significant effect of treatment on decreasing reported methamphetamine use in follow-up (see ‘Unweighted’ estimate). The TATE estimates obtained across all methods suggested a similar non-significant conclusion, indicating that the original findings from within the trial sample still hold when generalized to the target population of interest. It is important to also note that the distribution of the pre-treatment covariates in the trial resembled those in the target population much more closely after weighting the sample by using Random Forests to predict sample membership (Table 1).

6. Discussion

When recruiting fully representative samples or altering study design to strengthen external validity is infeasible, statistical methods for estimating target population effects are helpful tools that allow researchers to better estimate population average treatment effects post-hoc. The application of these methods to real-world data highlights several limitations and challenges.

First, identifying the right data to represent the target population is crucial, and depends on both the policy question at hand and the availability of population data related to the subject matter of the trial (e.g., from a nationally representative survey). Limited covariates available in population-level data sets poses problems of satisfying Assumption A2: that there are no unmeasured variables related to treatment effect and trial participation, once we adjust for

the observed factors. Sensitivity analyses have been recently proposed to test how sensitive TATE estimates are to unobserved effect modifiers, and should be utilized in cases of concern over data availability (Nguyen et al., 2017).

Second, it is important to note that while TEDS-A consisted of 135,264 admissions records, the CTN trial consisted of only 140 participants. Trying to generalize from a small sample to a very large population may impact the performance of the generalizability methods discussed, and is subject to further ongoing research.

Lastly, choosing the most appropriate generalizability method is not always trivial, nor it is obvious when or when not to generalize a trial's results at all. For example, the CTN's mission is to determine the effectiveness of interventions in diversified patient populations, and so the CTN trial described in this paper may actually be more generalizable by design than other RCTs. While the Tipton generalizability index provides a useful summary of differences between a trial and target population based on the predicted trial participation probabilities, one should also assess the balance of the covariates post-weighting, and consider the importance of the variables included in the selection model in terms of how related they are to effects (Kern et al., 2016).

In this paper, we highlighted and implemented several methods to estimate population average treatment effects, providing practical considerations for researchers to follow. Assessing and improving the external validity of RCTs is an important step in improving how clinical findings are used in practice (i.e., determining whether to train providers to administer a new intervention based on its potential effect in their population). While data availability and quality may be scarce, the methods discussed and the accompanying R package are useful tools to evaluate the generalizability of a trial's results, and should be carefully implemented prior to drawing population-level inferences from trial data.

Acknowledgments

Role of Funding Sources: Funding for this study was provided by NIDA Grant R01DA036520 [PI: R. Mojtabai]. NIDA hosts the data share from which the randomized trial data used as an illustrative example in this manuscript were obtained. NIDA had no role in the analysis or interpretation of the data, in the manuscript writing, or in the decision to submit the manuscript for publication.

Appendix

In this appendix, we demonstrate the implementation of the methods described in Section 3 on the data example described in Section 5 by using the R package "generalize." The "generalize" package contains two core functions: `assess` and `generalize`.

`Assess` evaluates similarities and differences between the trial sample and the target population based on a specified list of common covariates. This is done in a few ways:

1. **Covariate table:** `assess` provides a summary table of covariate means in the trial and the population, along with absolute standardized mean differences (ASMD) between the two sources of data.

2. Trial participation probabilities: assess estimates the probability of trial participation based on a specified vector of covariate names and statistical method, and summarizes their distribution across the trial and target populations. For this, logistic regression is the default method, but estimation using Random Forests or Lasso is currently supported by the package as well.
3. Generalizability index: assess utilizes the estimated trial participation probabilities to calculate the Tipton generalizability index described in Section 3.2.
4. Target population “trimming”: assess can check for any violations of the coverage assumption (A3). If the parameter `trim_pop` is set to equal `TRUE`, then assess returns a “trimmed” data set excluding all individuals in the target population with covariate values outside the ranges of the respective trial covariates, and reports how many individuals in the population were excluded.

After assessing the generalizability, the `generalize` function can be used to implement the TATE estimation methods described in Section 3.3. Weighting by the inverse odds using logistic regression is the default method, though weights based on other models (Lasso or Random Forests) or using BART or TMLE are available for use as well.

We now demonstrate how to use the “generalize” package to compare the CSP-1025 trial to the TEDS-A-2014 population, and then to estimate the TATE for the outcome “methamphetamine use in followup.” Note that in the code below, the stacked data set will be referred to as “meth_data,” and that these results are purely illustrative.

First, we install and load the package from Github using the “devtools” package (Wickham & Chang, 2017):

```
devtools::install_github("benjamin-ackerman/generalize")
library(generalize)
```

For convenience, we define a vector of covariate names:

```
covariates = c("age", "sex", "race", "ethnicity", "maritalstatus",
               "education", "employment", "methprior")
```

Next, to assess the differences between the trial (CSP-1025) and the population (TEDS-A-2014), we use the `assess` function, estimating the trial participation probabilities using Random Forests. To check the coverage assumption, we set the parameter `trim_pop` to equal `TRUE`:

```
assess_object = assess(trial = "trial", selection_covariates = covariates,
                       data = meth_data, selection_method = "rf", trim_pop = TRUE)
```

```
summary(assess_object)

## Probability of Trial Participation:
##
## Selection Model: trial ~ age + sex + race + ethnicity + maritalstatus + ed
ucation + employment + methprior
##
##               Min.   1st Qu.   Median     Mean 3rd Qu.    Max.
## Trial (n = 137)      0  0.002019  0.01075  0.011820  0.01739  0.04932
## Population (n = 126344)  0  0.000000  0.00000  0.001618  0.00134  0.05630
##
## Estimated by Random Forests
## Generalizability Index: 0.604
## =====
## Covariate Distributions:
##
## Population data were trimmed for covariates to not exceed trial covariate
bounds
## Number excluded from population: 8923
##
##
##               trial population  ASMD
## age18.20          0.0219      0.0454 0.113
## age21.24          0.0365      0.1241 0.266
## age25.29          0.1387      0.2072 0.169
## age30.34          0.1241      0.2132 0.218
## age35.39          0.2409      0.1492 0.257
## age40.44          0.1825      0.1067 0.245
## age45.49          0.1679      0.0775 0.338
## age50.54          0.0730      0.0503 0.104
## age55.            0.0146      0.0264 0.073
## sexMale           0.6350      0.5386 0.193
## raceBlack         0.0219      0.0433 0.105
## raceNative.Hawaiian 0.0292      0.0129 0.144
## raceOther         0.1022      0.1828 0.209
## raceWhite         0.8321      0.7411 0.208
## ethnicityNot.Hispanic.Latino 0.8613      0.7856 0.185
## ethnicityUnknown.Not.Given 0.0365      0.0070 0.355
## maritalstatusMarried.Partnered 0.2263      0.0943 0.452
## education12       0.4015      0.4602 0.118
## education13.15    0.3285      0.1717 0.416
## education16.      0.1460      0.0291 0.695
## education9.11     0.1022      0.2862 0.407
## employmentNot.in.labor.force 0.0657      0.3689 0.629
```

```
## employmentPart.time          0.2482      0.0701 0.697
## employmentUnemployed         0.2409      0.4526 0.425
## methprior                     0.9124      0.4243 0.988
```

The assess function creates an object of the class “generalize_assess.” The summary of a “generalize_assess” object returns the selection model, the distribution of the trial participation probabilities by data source, and the method of trial participation probability estimation. It also returns the calculated Tipton generalizability index, the number of individuals excluded due to coverage violations, and a table of the covariate distributions. Since we set trim_pop = TRUE, all of the results generated by assess used the “trimmed” data set.

Lastly, we estimate the effect of treatment on reported methamphetamine use at followup (“methfollowup”) by using the generalize function. Here, we estimate the TATE using weighting by the inverse odds, where the probabilities are estimated by Random Forests. Since there were a large number of individuals violating the coverage assumption (n=8923), we again “trim” the target population here:

```
generalize_object = generalize(outcome = "methfollowup", treatment = "treat",
  trial = "trial", selection_covariates = covariates, data = meth_data,
  method = "weighting", selection_method = "rf", trim_pop = TRUE)

summary(generalize_object)

## Average Treatment Effect Estimates:
##
## Outcome Model: methfollowup ~ treat
##
##      Estimate Std. Error 95% CI Lower 95% CI Upper
## SATE -0.1260684  0.1149249  -0.3513211  0.09918434
## TATE -0.1218059  0.1162635  -0.3496825  0.10607059
##
## =====
## TATE estimated by Weighting
## Weights estimated by Random Forests
##
## Trial sample size: 137
## Population size: 126344
## Population data were trimmed for covariates to not exceed trial covariate
## bounds
## Number excluded from population: 8920
##
## Generalizability Index: 0.606
##
```

```
## Covariate Distributions after Weighting:
##
##               trial (weighted) population  ASMD
## age18.20              0.0170      0.0454 0.136
## age21.24              0.0856      0.1241 0.117
## age25.29              0.2464      0.2072 0.097
## age30.34              0.1754      0.2132 0.092
## age35.39              0.1892      0.1492 0.112
## age40.44              0.1635      0.1067 0.184
## age45.49              0.0854      0.0775 0.029
## age50.54              0.0332      0.0503 0.078
## age55.                0.0043      0.0264 0.138
## sexMale                0.5565      0.5386 0.036
## raceBlack              0.0208      0.0433 0.111
## raceNative.Hawaiian   0.0127      0.0129 0.002
## raceOther              0.1219      0.1828 0.158
## raceWhite              0.8430      0.7411 0.233
## ethnicityNot.Hispanic.Latino 0.9356      0.7856 0.366
## ethnicityUnknown.Not.Given 0.0050      0.0070 0.024
## maritalstatusMarried.Partnered 0.1340      0.0943 0.136
## education12            0.4789      0.4602 0.037
## education13.15         0.2620      0.1717 0.239
## education16.           0.0312      0.0291 0.012
## education9.11          0.1972      0.2862 0.197
## employmentNot.in.labor.force 0.1584      0.3689 0.436
## employmentPart.time    0.1104      0.0701 0.158
## employmentUnemployed   0.6345      0.4526 0.365
## methprior              0.8574      0.4243 0.877
```

The generalize function creates an object of the class “generalize.” The summary of a “generalize” object returns a table with the SATE and TATE estimates, along with their standard errors and 95% confidence intervals (or credible intervals, when BART is used). When weighting is the specified method of TATE estimation, a covariate distribution table is printed as well, where the covariate means in the trial are weighted by the trial participation weights.

References

- Ackerman B (2018). Generalize: an R package for generalizing average treatment effects from rcts to target populations. Retrieved from <http://www.github.com/benjamin-ackerman/generalize>
- Cole SR, & Stuart EA (2010). Generalizing evidence from randomized clinical trials to target populations: The actg 320 trial. *American Journal of Epidemiology*, 172(1), 107–115. [PubMed: 20547574]
- Dababnah S, & Parish SL (2016). A comprehensive literature review of randomized controlled trials for parents of young children with autism spectrum disorder. *Journal of Evidence-Informed Social Work*, 13(3), 277–292. [PubMed: 26177069]

- Elkashef A, Kahn R, Yu E, Iturriaga E, Li S-H, Anderson A, et al. (2012). Topiramate for the treatment of methamphetamine addiction: A multi-center placebo-controlled trial. *Addiction*, 107(7), 1297–1306. [PubMed: 22221594]
- Flay BR (1986). Efficacy and effectiveness trials (and other phases of research) in the development of health promotion programs. *Preventive Medicine*, 15(5), 451–474. [PubMed: 3534875]
- Gruber S, & Van Der Laan MJ (2009). Targeted maximum likelihood estimation: A gentle introduction.
- Hill JL (2011). Bayesian nonparametric modeling for causal inference. *Journal of Computational and Graphical Statistics*, 20(1), 217–240.
- Imai K, King G, & Stuart EA (2008). Misunderstandings between experimentalists and observationalists about causal inference. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 171 (2), 481–502.
- Insel TR (2006). Beyond efficacy: The star* d trial. *American Journal of Psychiatry*, 162(1), 5–7. [dataset].
- Johnson BA (2015) “NIDA-CSP-1025.” National Institute on Drug Abuse Data Share Website. Retrieved from datashare.nida.nih.gov/study/nida-csp-1025
- Johnson BA, Rosenthal N, Capece JA, Wiegand F, Mao L, Beyers K, et al. (2007). Topiramate for treating alcohol dependence: A randomized controlled trial. *Jama*, 298(14), 1641–1651. [PubMed: 17925516]
- Kampman KM, Pettinati H, Lynch KG, Dackis C, Sparkman T, Weigley C, et al. (2004). A pilot trial of topiramate for the treatment of cocaine dependence. *Drug and Alcohol Dependence*, 75(3), 233–240. [PubMed: 15283944]
- Kern HL, Stuart EA, Hill J, & Green DP (2016). Assessing methods for generalizing experimental impact estimates to target populations. *Journal of Research on Educational Effectiveness*, 9(1), 103–127. doi:10.1080/19345747.2015.1060282 [PubMed: 27668031]
- Lee BK, Lessler J, & Stuart EA (2010). Improving propensity score weighting using machine learning. *Statistics in Medicine*, 29(3), 337–346. [PubMed: 19960510]
- Lesko CR, Buchanan AL, Westreich D, Edwards JK, Hudgens MG, & Cole SR (2017). Generalizing study results: A potential outcomes perspective. *Epidemiology*, 28(4), 553–561. [PubMed: 28346267]
- Moradveisi L, Huibers M, Renner F, & Arntz A (2014). The influence of patients' preference/attitude towards psychotherapy and antidepressant medication on the treatment of major depressive disorder. *Journal of Behavior Therapy and Experimental Psychiatry*, 45(1), 170–177. [PubMed: 24172145]
- Nguyen TQ, Ebnesajjad C, Cole SR, Stuart EA, & others. (2017). Sensitivity analysis for an unobserved moderator in rct-to-target-population generalization of treatment effects. *The Annals of Applied Statistics*, 11(1), 225–247.
- Pearl J, & Bareinboim E (2011). Transportability of causal and statistical relations: A formal approach In *Data mining workshops (icdmw)*, 2011 *IEEE 11th international conference on* (pp. 540–547). IEEE.
- Peto R, Collins R, & Gray R (1995). Large-scale randomized evidence: Large, simple trials and overviews of trials. *Journal of Clinical Epidemiology*, 48(1), 23–40. [PubMed: 7853045]
- R Core Team. (2017). R: A language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing Retrieved from <https://www.R-project.org/>
- Rosenbaum PR, & Rubin DB (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70(1), 41–55.
- Rubin DB (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology*, 66(5), 688.
- Rubin DB (2001). Using propensity scores to help design observational studies: Application to the tobacco litigation. *Health Services and Outcomes Research Methodology*, 2(3), 169–188.
- Rubin DB (2008). For objective causal inference, design trumps analysis. *The Annals of Applied Statistics*, 808–840.
- Rudolph KE, Díaz I, Rosenblum M, & Stuart EA (2014). Estimating population treatment effects from a survey subsample. *American Journal of Epidemiology*, 180(7), 737–748. [PubMed: 25190679]

- Shadish WR, Cook TD, & Campbell DT (2002). Experimental and quasi-experimental designs for generalized causal inference. Wadsworth Cengage learning.
- Stuart EA (2010). Matching methods for causal inference: A review and a look forward. *Statistical Science: A Review Journal of the Institute of Mathematical Statistics*, 25(1), 1. [PubMed: 20871802]
- Stuart EA, & Rhodes A (2017). Generalizing treatment effect estimates from sample to population: A case study in the difficulties of finding sufficient data. *Evaluation Review*, 41(4), 357–388. [PubMed: 27491758]
- Stuart EA, Ackerman B, & Westreich D (2017). Generalizability of randomized trial results to target populations: Design and analysis possibilities. *Research on Social Work Practice*, 1049731517720730.
- Stuart EA, Cole SR, Bradshaw CP, & Leaf PJ (2011). The use of propensity scores to assess the generalizability of results from randomized trials. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 174(2), 369–386.
- Susukida R, Crum RM, Ebnesajjad C, Stuart EA, & Mojtabai R (2017). Generalizability of findings from randomized controlled trials: Application to the national institute of drug abuse clinical trials network. *Addiction*.
- Susukida R, Crum RM, Stuart EA, Ebnesajjad C, & Mojtabai R (2016). Assessing sample representativeness in randomized controlled trials: Application to the national institute of drug abuse clinical trials network. *Addiction*, 111(7), 1226–1234. [PubMed: 26825855]
- Tibshirani R (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, 267–288.
- Tipton E (2014). How generalizable is your experiment? An index for comparing experimental samples and populations. *Journal of Educational and Behavioral Statistics*, 39(6), 478–501.
- Weisberg HI, Hayden VC, & Pontes VP (2009). Selection criteria and generalizability within the counterfactual framework: Explaining the paradox of antidepressant-induced suicidality? *Clinical Trials*, 6(2), 109–118. [PubMed: 19342462]
- Wickham H, & Chang W (2017). Devtools: Tools to make developing r packages easier. Retrieved from <https://CRAN.R-project.org/package=devtools>
- Zhai F, Raver CC, Jones SM, Li-Grining CP, Pressler E, & Gao Q (2010). Dosage effects on school readiness: Evidence from a randomized classroom-based intervention. *Social Service Review*, 84(4), 615–655. [PubMed: 21488322]

Highlights:

- RCT results from non-representative samples may not generalize to the population.
- Methods exist to assess a trial's generalizability and estimate population effects.
- Carefully defining a target population is crucial to generalizing RCT findings.

		S (trial membership)	Y (outcome)	A (treatment)	X ₁ (covariate 1)	X ₂ (covariate 2)	...	X _p (covariate p)
Trial data set	1		✓	✓	✓	✓		✓
	0		×	×	✓	✓		✓
Population data set								

Figure 1:
Format of "stacked" data set for implementing generalizability methods

Figure 2:
log(Trial Participation Probabilities) by Method and Sample Membership

Figure 3:
Average Treatment Effect of Topiramate on Methamphetamine Use Reported in Followup by Generalizability Method

Table 1:

Distribution of Covariates in the Trial vs. Population and their Absolute Standardized Mean Difference (ASMD) pre and post-weighting

	CSP-1025 (trial)	TEDS-A-2014 (population)	ASMD (pre-weighting)	ASMD (post-weighting)
Age				
12-14	0.00	0.01	0.04	0.04
15-17	0.00	0.02	0.13	0.13
18-20	0.02	0.04	0.11	0.06
21-24	0.04	0.12	0.26	0.10
25-29	0.14	0.20	0.16	0.08
30-34	0.12	0.21	0.21	0.14
35-39	0.24	0.15	0.27	0.24
40-44	0.18	0.10	0.25	0.21
45-49	0.17	0.08	0.35	0.06
50-54	0.07	0.05	0.11	0.11
55 +	0.01	0.03	0.07	0.15
Sex				
Male	0.64	0.54	0.20	0.03
Race				
Black	0.02	0.04	0.10	0.08
White	0.83	0.74	0.21	0.10
Native Hawaiian	0.03	0.01	0.15	0.04
Other	0.10	0.18	0.21	0.09
Ethnicity				
Hispanic/Latino	0.10	0.21	0.27	0.45
Not Hispanic/Latino	0.86	0.78	0.19	0.45
Marital Status				
Married/Partnered	0.23	0.09	0.46	0.07
Education				
9-11 years	0.10	0.29	0.42	0.24
12 years	0.40	0.45	0.10	0.07
13-15 years	0.33	0.17	0.43	0.31
16 + years	0.15	0.03	0.70	0.04
Employment				
Not in labor force	0.07	0.38	0.64	0.46
Part-time	0.25	0.07	0.70	0.09
Unemployed	0.24	0.45	0.41	0.33
Methamphetamine Used in Past Week				
Yes	0.91	0.42	0.99	0.81