STATISTICAL METHODS FOR TRANSPORTABILITY: ADDRESSING EXTERNAL VALIDITY AND MEASUREMENT ERROR CONCERNS IN RANDOMIZED TRIALS

by

Benjamin Ackerman

A dissertation submitted to Johns Hopkins University
in conformity with the requirements for the degree of
Doctor of Philosophy

Baltimore, Maryland
April, 2020

Abstract

Randomized trials are considered the gold standard for estimating causal effects, and evidence from trials is highly regarded in decision making processes that impact entire populations. While rigorous in design, RCTs can still be flawed; leveraging data and information from additional non-experimental or “real world" studies can be advantageous for addressing statistical issues and improving inferences. This dissertation addresses two complications that arise in trials and can be addressed in this way: poor external validity and measurement error. To deal with both of these issues, it is important to consider (and account for) differences in baseline covariates between the RCT sample and the external data source. In other words, it is crucial to address how “transportable" inferences are between the two studies. This work focuses on transportability between an RCT and an external non-experimental study in two contexts: 1) when generalizing RCT findings to a well-defined target population and 2) when correcting for outcome measurement error in an RCT.

Thesis Committee

Primary Readers

Elizabeth A. Stuart (Primary Advisor)
Associate Dean for Education
Professor
Department of Mental Health
Department of Biostatistics
Department of Health Policy and Management
Johns Hopkins Bloomberg School of Public Health

Margaret Daniele Fallin
Professor, Chair
Department of Mental Health
Johns Hopkins Bloomberg School of Public Health

Catherine R. Lesko
Assistant Professor
Department of Epidemiology
Johns Hopkins Bloomberg School of Public Health

Elizabeth Ogburn
Associate Professor
Department of Biostatistics
Johns Hopkins Bloomberg School of Public Health

Alternate Readers

Scott L. Zeger
Professor
Department of Biostatistics
Johns Hopkins Bloomberg School of Public Health

Joanne Katz
Professor
Department of International Health
Johns Hopkins Bloomberg School of Public Health

Acknowledgments

First and foremost, to my PhD advisor, Liz Stuart. I feel incredibly lucky and fortunate for all of your guidance throughout this journey. Right from the beginning, you've provided me with countless opportunities to grow and learn as a researcher, writer, and communicator. Along the way, you've also taught me the importance of a healthy work-life balance, showing me how to stay productive without sacrificing or undervaluing my personal life and well-being. You've helped me find my voice and my confidence as an applied statistician, and you've inspired me to identify statistical questions embedded in impactful health-related problems. I can't thank you enough for believing in me, supporting me, and for always being such a grounding presence throughout graduate school. I look forward to staying in touch, and to applying the lessons you've taught me in my next position and beyond.

To my final dissertation committee members, Katie Lesko, Dani Fallin and Betsy Ogburn, thank you for taking the time to read my dissertation and provide such thoughtful feedback. Additionally, thank you to Ramin Mojtabai and John Jackson, members of my Preliminary Oral Exam Committee and Thesis Advisory Committee, for also monitoring my dissertation progress and helping to provide a well-rounded public health perspective to my thesis research. All of your input, guidance and mentorship has made me a better student, researcher and person!

To all of my Stuart Lab colleagues (past and present), thank you for pushing me to be more rigorous and thorough in my research. I've grown from all of your ideas and suggestions, and I am grateful for the chance to address so many interesting research questions with you. Trang Nguyen, you are an absolute rock star, and I thank you for your wisdom, patience and support.

To Karen Bandeen-Roche, for her fierce dedication as Chair of the Biostatistics Department to the success and well-being of all students. Karen, I will always cherish our conversations, and I am grateful for the generosity and compassion you have shown me as a graduate student. Thank you for all of your support and encouragement, and for fostering such a special departmental culture. To Brian Caffo, former director of the Biostatistics PhD program, for helping me persevere with the first two years of coursework. Your perspective and reassurance helped me build confidence when I struggled to feel like I belonged here - thank you for believing in me. To Mary Joy Argo, for knowing *every* answer to my *many* questions. Thank you for your wealth of knowledge and kindness. To Kara Schoenberg, Ashley Johnson, Marti Gilbert and Patty Hubbard, for always brightening up my day with a visit to the front office. To Scott Zeger, whose introductory course inspired me as a sophomore Public Health Studies major at JHU to pursue a career in Biostatistics. Scott, thank you for the many years of mentorship. To Margaret Taub and Leah Jager, for allowing me to give back to the course that started me on this journey, and for encouraging me and cheering me on at every step of the way.

To Tonia Poteat, Stefan Baral, Heather Volk and Xiaobin Wang, for allowing me to dig my hands into their data and contribute meaningfully to public health research. To Helen Pentikis, for the opportunities to impact cancer care through drug development consulting projects over the years. To Rayid Ghani and everyone at the Data Science for Social Good Fellowship, for the deep dive on making positive, long-lasting societal change through data science and machine learning. Working with all of you has been such a joy, and has served as an important reminder that every data point represents a person and a life. Thank you for enriching my graduate school career with applied research, collaboration and data!

To all of my friends at JHSPH and in the Biostat department: thank you for creating such a vibrant and supportive student culture that made coming into school so much fun. To Elizabeth Sweeney, Mandy Mejia, John Muschelli and David Lenis, thank you for your friendship and for leading by example as role models. I looked up to each of you as a new PhD student, and am so grateful for your guidance. To my Mental Health Grad Network peers, thank you for fighting to end the stigma around graduate student mental health. Keep advocating for better access to the high quality mental health resources that graduate students at JHSPH deserve! To Sophie Bérubé and Bonnie Smith: without you both, I'd definitely still be re-taking Probability Theory right now. You've been great study pals, and even better friends, confidants and boozy milkshake sharers. To my boys: Jacob Fiksel, Lamar Hunt and Matt Cole. Oh boy... it's been a wild ride, and I seriously couldn't have gotten through any of this without you guys. We did it, and we made our moms proud!

To all of my "nons-Hopkins" Baltimore friends, and to the Harbor Minyan and Hinenu communities, thank you for making this city truly feel like home. You've helped distract me from the stresses of grad school, cheered me on in many Baltimore Running Festivals, and consumed lots of ice cream and Johnny Rad's pizza with me. The friendships and communities we've built and maintained over the last nine years are incredibly special to me, and I feel so lucky to have you all in my life. To all of the good dogs that I've befriended in grad school, thank you for bringing a huge smile to my face and wiping away all of my biggest worries with your presence. A special thank you to Sir Charles, the geriatric Welsh Corgi of Cleveland (formerly of Baltimore), for being the best boy in the whole world. You are the dog I always wanted and never had. Thank you Danielle LeVeck for sharing Chuck with me and for being such a dear friend.

To my boyfriend, Michael, thank you for being a voice of reason, a calming presence, and somebody I can always depend on. You've helped me stay focused on what matters, and being with you brings me so much happiness and makes me a better person. I love you.

Last and most certainly not least, to my loving parents and sister. You know better than anyone that this degree has not been easy for me at all. Thank you for never giving up on me, for comforting me at my lowest moments, and for cheering me on at my proudest moments. I love you very much, and will forever cherish everything that you've given me.

Dedication

For my grandmother, Roz Klarman, z"l.

Table of Contents

List of Tables

List of Figures

Chapter 1

Introduction

Randomized controlled trials (RCTs) are considered the gold standard for estimating causal effects, and evidence from trials is highly regarded in decision making processes that impact entire populations. Randomization of treatment assignment helps yield strong internal validity and allows for unbiased estimation of the sample average treatment effect (ATE). While rigorous in design, RCTs can still be flawed. Leveraging data and information from additional non-experimental or “real world” studies can be advantageous for addressing various statistical issues and improving inferences drawn from RCTs. When doing so, it is important to consider the relationship between each study sample’s demographics, and how they each relate to a common target population of interest. In other words, it is crucial to address how “transportable" inferences are between the two studies, such as by comparing the baseline covariate distributions between them. Failure to address transportability when supplementing RCTs with external data can add further biases to ATE estimates, sometimes more-so than if the supplemental data were not used at all.

This dissertation research focuses on the development and dissemination of statistical methods for transportability when addressing two complications that arise in trials: poor generalizability and outcome measurement error. The first issue is a matter of external validity, where supplemental data are being used to extrapolate ATE estimates *from* an RCT to a broader population. The second issue is a matter of internal validity, where supplemental data are being used to model a measurement error structure that is then applied *to* the RCT sample. Nevertheless, the statistical methodology discussed and proposed in both of these cases are complementary.

Chapter 2 builds upon existing statistical methods for generalizing RCT inferences to well-defined target populations. Findings from RCTs are often used to inform health policy and public health program implementation, yet their results may not generalize well to a policy-relevant target population due to potential differences in effect moderators between the trial and population (Imai, King, & Stuart, 2008). This issue has been frequently raised about trials across various fields in health (Dababnah & Parish, 2016; Susukida, Crum, Ebnesajjad, Stuart, & Mojtabai, 2017), social work (Stuart, Ackerman, & Westreich, 2017) and education (Tipton & Olsen, 2018), and can often be attributed to the convenience sampling recruitment strategies implemented to acquire the trial sample. While there are trial design approaches to improving external validity (Flay, 1986; Insel, 2006), there are many barriers to changing recruitment for medical trials, such as time, money and strict exclusion criteria established for safety purposes. Recently, post-hoc statistical methods have been developed to generalize trial findings to a target population, and to assess when such generalizations are even possible (Kern, Stuart, Hill, & Green, 2016; Stuart, Cole, Bradshaw, & Leaf, 2011). One such generalization method draws from the propensity score literature by modeling the probability of trial selection conditional on pre-treatment characteristics, and weighting the trial so that it better resembles the target population (Cole & Stuart, 2010). This approach (as well as other model-based approaches) requires finding external data that are a simple random sample of the target population, which can be challenging to do in practice (Stuart & Rhodes, 2017).

One promising source of population data are large health-related government surveys; they often have an extensive set of measured covariates that describe a wide range of populations. However, given their complex survey design, these datasets are *not* representative of their respective target populations without the incorporation of survey weights. Existing generalization methods do not account for this type of population data study design, and applying current methods using a population survey could therefore produce incorrect (or biased) estimates of the population average treatment effect (PATE). In Chapter 2, we formally show that the PATE depends on both the RCT-to-survey transportability weights and the survey's inverse probability of selection weights. We then propose and evaluate an extension to existing generalization weighting methods: a two-stage weighting approach that incorporates survey weights from supplementary population survey data when generalizing trial findings.

Chapter 3 similarly draws upon the propensity score weighting literature to address transportability when correcting for measurement error in lifestyle intervention trials. Lifestyle intervention trials aim to establish how changes to human behavior, such as physical activity or food intake, can impact and improve health outcomes. In such trials, it is important to obtain accurate measures on these behaviors; however, reliable measures are often expensive and burdensome for participants to collect. Self-reported outcomes, such as dietary intake, are often therefore used in order to assess the intervention's effectiveness. While less costly and challenging to obtain, these measures are subject to measurement error, which can lead to biased estimates of the average treatment effect (Rothman, Greenland, & Lash, 2008; Willett, 2012). Methods have been developed to correct for measurement error by using external validation studies, which measure both the self-reported outcome and an accompanying biomarker, to model the measurement error structure (Wong, Day, Bashir, & Duffy, 1999). Much of the attention in the measurement error literature has been paid to when measurement error is present in either the exposure of interest or in covariates (Buonaccorsi, 2010; Carroll, Ruppert, Stefanski, & Crainiceanu, 2006; Keogh & White, 2014). Less work, however, has focused on correcting for misclassification or measurement error in study outcomes (Keogh, Carroll, Tooze, Kirkpatrick, & Freedman, 2016), which is particularly worrisome for trials that focus on self-reported behavioral outcomes (Spring et al., 2012; Spring et al., 2018). Additionally, external validation samples typically only collect measures under a "usual care" setting, which we must assume is equivalent to the control conditions of a randomized trial. This makes it infeasible to directly correct for the error under both the treatment *and* control conditions based on the information available to researchers.

Siddique et al. (2019) developed methodology for modeling outcome measurement error under the control condition using an external validation sample, followed by sensitivity analyses to obtain a range of plausible values for the treatment effect. The existing literature, for both covariate and outcome measurement error, often assumes that measurement error models from external validation studies apply directly to the variable in the trial of interest; however, there is growing concern that such error corrections may not transport well due to pre-treatment characteristic differences between the two samples. We show that poor transportability can lead to further biases in estimating the average treatment effect. We then evaluate the relationship between such covariate imbalance and measurement error correction through simulation, and propose the use of propensity score-type weighting methods to improve upon error correction transportability. Chapter 3 concludes with guidance for researchers on how to check if their validation sample inferences would transport well to the trial of interest, with the hopes of making this work easy to implement in practice.

Finally, Chapter 4 returns to generalizability with a methods tutorial paper published in *Addictive Behaviors* (Ackerman et al., 2019). This publication provides an overview of existing statistical methods for assessing and generalizing findings from randomized trials to a well-defined target population, with an applied research audience in mind. The paper highlights how to approach several pragmatic issues when making generalizations, such as where to look for population data, how to harmonize the external data with the trial data, and how to check for violations of key assumptions. Accompanying this tutorial is the development of an R package, “generalize." This R package wraps existing methods for assessing and improving upon RCT external validity, and is available to download and install from Github at http://benjamin-ackerman.github.io/generalize. The software provides user-friendly functions for implementing the propensity score-type weighting generalization method described in Chapter 2. It also has functions to implement outcome-modeling generalization approaches, where flexible models of the outcome conditional on observed covariates are fit in the trial, and then outcomes under treatment conditions are predicted in the target population. This can be done using Bayesian Additive Regression Trees (BART) (Hill, 2011; Kern et al., 2016) or Targeted Maximum Likelihood Estimation (TMLE) (Rudolph, Díaz, Rosenblum, & Stuart, 2014). Lastly, “generalize" allows researchers to compare the baseline demographic distributions between a trial and target population, both individually (as standardized mean differences) and jointly through a generalizability index (Tipton, 2014).

Each chapter of this dissertation provides methodological advances for improving inferences drawn from randomized trials. Chapter 3 demonstrates the consequences of applying outcome measurement error correction from a validation study that is not transportable to the RCT of interest, and illustrates the benefit in applying propensity score-type weighting methods to improve transportability. Chapter 2 identifies a pragmatic issue that arises when implementing generalization methods using target population data from a complex survey, and offers a methodological fix to ensure that findings are generalized to the correct target population of interest. Chapter 4 aims to bridge the gap between method development and method implementation for generalizability, providing applied researchers with guidance and software for generalizing trial findings. This work highlights the importance of taking transportability into consideration when supplementing randomized trials with external data.

References

Ackerman, B., Schmid, I., Rudolph, K. E., Seamans, M. J., Susukida, R., Mojtabai, R., & Stuart, E. A. (2019). Implementing statistical methods for generalizing randomized trial findings to a target population. *Addictive behaviors*, *94*, 124–132.

Buonaccorsi, J. P. (2010). *Measurement error: Models, methods, and applications*. Chapman and Hall/CRC.

Carroll, R. J., Ruppert, D., Stefanski, L. A., & Crainiceanu, C. M. (2006). *Measurement error in nonlinear models: A modern perspective*. Chapman and Hall/CRC.

Cole, S. R., & Stuart, E. A. (2010). Generalizing evidence from randomized clinical trials to target populations: The actg 320 trial. *American journal of epidemiology*, *172*(1), 107–115.

Dababnah, S., & Parish, S. L. (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.

Flay, B. R. (1986). Efficacy and effectiveness trials (and other phases of research) in the development of health promotion programs. *Preventive medicine*, *15*(5), 451–474.

Hill, J. L. (2011). Bayesian nonparametric modeling for causal inference. *Journal of Computational and Graphical Statistics*, *20*(1), 217–240.

Imai, K., King, G., & Stuart, E. A. (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, T. R. (2006). Beyond efficacy: The star* d trial. *American Journal of Psychiatry*, *163*(1), 5–7.

Keogh, R. H., & White, I. R. (2014). A toolkit for measurement error correction, with a focus on nutritional epidemiology. *Statistics in medicine*, *33*(12), 2137–2155.

Keogh, R. H., Carroll, R. J., Tooze, J. A., Kirkpatrick, S. I., & Freedman, L. S. (2016). Statistical issues related to dietary intake as the response variable in intervention trials. *Statistics in medicine*, *35*(25), 4493–4508.

Kern, H. L., Stuart, E. A., Hill, J., & Green, D. P. (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. eprint: http://dx.doi.org/10.1080/19345747.2015.1060282

Rothman, K. J., Greenland, S., Lash, T. L., et al. (2008). *Modern epidemiology*. Wolters Kluwer Health/Lippincott Williams & Wilkins Philadelphia.

Rudolph, K. E., Díaz, I., Rosenblum, M., & Stuart, E. A. (2014). Estimating population treatment effects from a survey subsample. *American journal of epidemiology*, *180*(7), 737–748.

Siddique, J., Daniels, M. J., Carroll, R. J., Raghunathan, T. E., Stuart, E. A., & Freedman, L. S. (2019). Measurement error correction and sensitivity analysis in longitudinal dietary intervention studies using an external validation study. *Biometrics*.

Spring, B., Schneider, K., McFadden, H. G., Vaughn, J., Kozak, A. T., Smith, M., . . . Hedeker, D., et al. (2012). Multiple behavior changes in diet and activity: A randomized controlled trial using mobile technology. *Archives of internal medicine*, *172*(10), 789–796.

Spring, B., Pellegrini, C., McFadden, H., Pfammatter, A. F., Stump, T. K., Siddique, J., . . . Hedeker, D. (2018). Multicomponent mhealth intervention for large, sustained change in multiple diet and activity risk behaviors: The make better choices 2 randomized controlled trial. *Journal of medical Internet research*, *20*(6), e10528.

Stuart, E. A., & 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.

Stuart, E. A., Cole, S. R., Bradshaw, C. P., & Leaf, P. J. (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.

Stuart, E. A., Ackerman, B., & Westreich, D. (2017). Generalizability of randomized trial results to target populations: Design and analysis possibilities. *Research on Social Work Practice*, 1049731517720730.

Susukida, R., Crum, R. M., Ebnesajjad, C., Stuart, E. A., & Mojtabai, R. (2017). Generalizability of findings from randomized controlled trials: Application to the national institute of drug abuse clinical trials network. *Addiction*.

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.

Tipton, E., & Olsen, R. B. (2018). A review of statistical methods for generalizing from evaluations of educational interventions. *Educational Researcher*, *47*(8), 516–524.

Willett, W. (2012). *Nutritional epidemiology*. Oxford university press.

Wong, M., Day, N., Bashir, S., & Duffy, S. (1999). Measurement error in epidemiology: The design of validation studies i: Univariate situation. *Statistics in medicine*, *18*(21), 2815–2829.

Chapter 2

Generalizing Randomized Trial Findings to a Target Population using Complex Survey Population Data

2.1 Introduction

Randomized controlled trials (RCTs) are considered the gold standard for estimating the causal effect of a new treatment or intervention; however, they often suffer from poor external validity, or generalizability (Imai, King, & Stuart, 2008; Shadish, Cook, & Campbell, 2002). Evidence from RCTs is frequently used when formulating health policy and implementing new large-scale health programs, but poor generalizability may hinder policymakers' abilities to make correct policy decisions for their populations. When feasible, trial designs that strategically sample from the target population of interest to improve representativeness have been shown to also improve upon the generalizability of RCTs (Insel, 2006; Peto, Collins, & Gray, 1995; Tipton &

Matlen, 2019); however, particularly in medical trials, there are many barriers to doing so, such as time, money and location. Recruitment strategies for RCTs that do not consider the ultimate target population of interest may lead to non-representative trial samples. More formally, if the trial sample differs from the target population on characteristics that moderate treatment effect, then the average treatment effect in the trial sample (SATE) will not equal the average treatment effect in the target population (PATE) (Cole & Stuart, 2010).

Several classes of post-hoc statistical methods have been developed to address concerns of generalizability once a trial has already been completed. One broad strategy uses propensity score-type methods to weight the trial so that it better resembles the target population on baseline covariates (Westreich, Edwards, Lesko, Stuart, & Cole, 2017). Note that this is similar to using propensity score weighting to estimate the average treatment effect on the treated (ATT) in non-experimental studies, where instead of fitting a model of treatment selection, a model of sample membership (i.e. trial participation vs. not) is specified. A second approach involves modeling the outcome as a flexible function of the observed covariates in the trial, and then predicting outcomes under treatment conditions in the target population. This can be done using Bayesian Additive Regression Trees (BART) (Hill, 2011; Kern, Stuart, Hill, & Green, 2016) or Targeted Maximum Likelihood Estimation (TMLE) (Rudolph, Díaz, Rosenblum, & Stuart, 2014). Lastly, doubly robust methods have been proposed, in which models are fit for both the outcome and the probability of sample membership (Dahabreh, Hernán, Robertson, Buchanan, & Steingrimsson, 2019).

The implementation of these methods requires the identification of a dataset for the target population of interest, one that contains individual-level data on all relevant treatment effect modifiers in the trial. While data availability and quality make this challenging to do (Stuart & Rhodes, 2017), in practice, large nationally representative surveys collected by government agencies are often good sources of information on policy-relevant populations. For example, the National Health and Nutrition Examination Survey (NHANES) consists of a series of annual surveys that collect information on participants' demographics, socioeconomic status, dietary behaviors and health outcomes, with supplemental laboratory tests and medical examinations (Johnson, Dohrmann, Burt, & Mohadjer, 2014). NHANES is designed to be representative of the non-institutionalized civilian US population across all 50 states and Washington D.C., and may therefore be a promising source of population data for implementing generalizability methods.

While surveys like NHANES may provide a wealth of information on the target population of interest, the analytic datasets on their own are themselves *not* representative of the target population. These raw datasets are the result of complex survey sampling designs that systematically over-sample and under-sample certain demographic groups. Such designs may involve stratifying the target population (e.g., first by state, then by county or Census tract) and then selecting primary sampling units (e.g., households, schools, individuals) by pre-specified rates, perhaps defined by demographic categories. Some surveys implement additional levels of stratification, for example, sampling counties first and then selecting individuals within the sampled counties. Selected participants in the final sample are then assigned sampling weights inversely proportional to their probability of being selected. Additional corrections for non-response and post-stratification are also often applied (Valliant, Dever, & Kreuter, 2013). These sampling weights are typically included as a variable in the final analytic datasets, though note that not all variables used to construct the weights are always available for researchers to use. For example, sampling may occur at the zipcode level, but for confidentiality reasons, zipcode may be omitted from the final public-use dataset, while a correlated variable, such as state or region, may be included.

Given these complex survey design elements, any inferences made by weighting a trial to look like one of these survey raw datasets will generally not be accurate for the true target population, rather they will just reflect the survey sample's demographics. In other words, when using NHANES as target population data without utilizing NHANES' survey weights, one would be generalizing to the NHANES sample, *not* to the non-institutionalized civilian US population. While several studies have applied these generalizability methods using population data from complex surveys, no previous work has formalized an approach for properly incorporating survey weights when doing so.

Although the proper incorporation of complex survey design elements has not been not been addressed in the generalization context, there are some methodological similarities to be found in a limited, yet growing set of papers on using propensity score methods to estimate causal effects in non-experimental complex survey data. However, even in that context, there is no consensus on how to best use the survey weights when specifying a treatment assignment model, whether as weights or as covariates. Zanutto (2006) argue that survey weights do not need to be used in propensity score estimation when using matching methods, so long as the survey weights are used in modeling the outcome. Through simulation studies, DuGoff, Schuler, and Stuart (2014) show benefit in using the survey weights as predictors in the propensity score model, but not in using them to weight the propensity score model. Ridgeway, Kovalchik, Griffin, and Kabeto (2015) provide theoretical justification for weighting the propensity score model using the survey weights, and then weighting the outcome model by the resulting propensity score weights *multiplied* by the survey weights. Lenis, Nguyen, Dong, and Stuart (2017) observe no difference through simulation in how the survey weights are incorporated in the propensity score model, and Austin, Jembere, and Chiu (2018) similarly report inconclusive findings on the optimal specification of the propensity score model. Overall, though, researchers tend to agree that ignoring survey weights altogether yields causal estimates that do not generalize to the target population in which a survey was conducted, and may produce invalid inferences when using propensity score methods. An important distinction to make is that here, we are not using the survey weights to estimate an effect *within* the survey itself, rather we are using the survey as a target population to generalize *to*. Other recent relevant work by Yang, Ganesh, Mulrow, and Pineau (2018) demonstrates the benefit of a propensity score-type weighting approach when combining a non-probability sample with a companion probability sample to enhance population-level estimation. While their approach can be extended to our context by viewing

RCTs as non-probability samples and surveys as population-level data, this work does not provide detailed methodological justification on the proper use of the probability-sample's survey weights.

Given this existing relevant literature, we hypothesize that it is crucial to incorporate the survey weights, which relate the survey sample back to the target population of interest, in order to correctly generalize RCT findings to the target population of interest. The rest of this paper is structured as follows: In Section 2.2, we formally evaluate the consequences of ignoring survey weights when generalizing RCT findings to a target population on which data are available from a complex survey. We then propose an approach to estimating the population average treatment effect while incorporating survey weights in Section 2.3. In Section 2.4, we examine our hypothesis by conducting a simulation study to investigate when the proposed approach improves our population-level inferences. We then apply the methods to two generalization examples where population data come from complex surveys in Section 2.5, and we conclude by summarizing the findings and discussing future work in Section 2.6.

2.2 Transporting to a Complex Survey Population Dataset

2.2.1 Definitions and Assumptions

Suppose the goal of a randomized trial is to estimate the population average treatment effect (PATE), defined as $E[Y(1) - Y(0)]$ where $Y(a)$ is the potential outcome Y under treatment a ($a = 1$ denotes treatment and $a = 0$ denotes

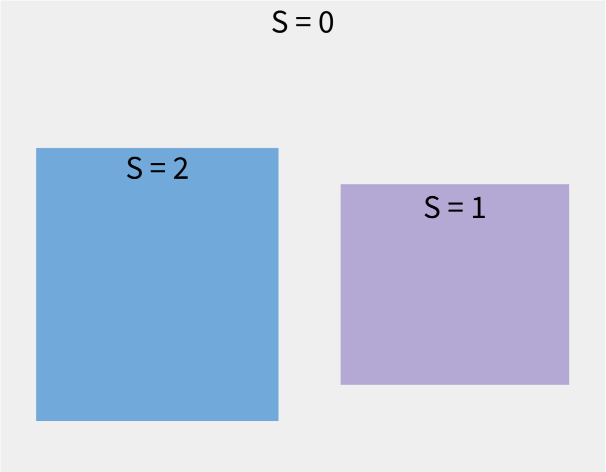


Figure 2.1: Scenario of how data sources relate to each other and to the target population. The entire grey region denotes the target population, $S = 1$ denotes the RCT, $S = 2$ denotes the complex survey sample, and $S = 0$ denotes members of the target population not sampled into either study. Only individuals with $S = 1$ or $S = 2$ are observed, while data on individuals with $S = 0$ are assumed unavailable. This three-level “S" variable also assumes no overlap between trial and survey participants. This is a plausible assumption to make for policy-relevant scenarios, where the target population may be the entire US, and the study sample sizes are on the magnitudes of a few thousand.

control). This expectation is defined across a well-defined target population of interest. Let S denote sample membership, where $S = 1$ denotes trial membership, $S = 2$ denotes survey membership, and $S = 0$ denotes the individual is in the target population, but not the trial nor the survey sample (See Figure 2.1)[1]. Here, we assume no overlap between the trial and survey samples, which is plausible for policy-relevant scenarios where the target population is the entire US and the study sample sizes are comparatively small. Additionally, let A denote treatment assignment and let X denote a set of pre-treatment covariates.

[1] Extensions of this work could consider settings in which the trial and survey samples overlap (i.e. having two indicator variables, one for trial selection and one for survey selection).

Note that the population of interest is the union of all S levels; however, in practice, we often do not have any data on the full population, nor do we observe outcomes for each level of S. Suppose all we have are data from the trial itself ($S = 1$). If the RCT is a simple random sample of the target population, then we can unbiasedly estimate the PATE using the trial data alone. However, if the treatment effect in the trial is moderated by covariate X, *and* if the distribution of X differs between the trial and the target population, then the naive estimate in the trial will be a biased estimate of the PATE (Olsen, Orr, Bell, & Stuart, 2013).

In such cases, we can supplement the trial with survey data ($S = 2$) and transport the estimate of the trial to the survey to obtain an unbiased estimate of the PATE (Westreich et al., 2017). Note that this requires the survey data to have all Xs related to sample selection and treatment effect heterogeneity fully observed, while treatment assignment and outcomes may be missing. Estimating the PATE by transporting the trial findings to a complex survey sample require making the following assumptions:

$1A$ All members of the target population have nonzero probability of being selected into the trial.

$1B$ All members of the target population also have nonzero probability of being selected into the survey.

$2A$ There are no unmeasured variables associated with treatment effect and trial sample selection.

$2B$ There are also no unmeasured variables associated with treatment effect, trial sample selection *and* survey sample selection.

3 Treatment assignment in the trial is independent of trial sample selection and the potential outcomes given the pre-treatment covariates.

4 The survey sample is a simple random sample of the target population (in other words, the survey is “self-weighting").

The plausibility of assumptions 1A, 2A and 3 have been discussed and established in previous work on generalizability. For instance, Nguyen, Ackerman, Schmid, Cole, and Stuart (2018) address assumption 2A by developing sensitivity analysis methods for unobserved moderators. When using population data that come from complex surveys, however, assumptions 2B and 4 must also be made. These two assumptions are under-discussed in the existing generalizability literature, and are also highly unrealistic assumptions to make given the complex survey designs of most publicly available government surveys. We now describe how biased the transported estimate will be as an estimate of the PATE when assumptions 2B and 4 are violated, and particularly, when the complex survey weights are ignored.

2.2.2 Consequence of ignoring survey weights in the PATE

Recall that the estimand of interest here is the PATE, defined as $\Delta = E[Y(1) - Y(0)]$. This estimand can be expanded upon and expressed as:

$$\Delta = E\left[\frac{\mathbb{1}_{S=1}AY}{e(\emptyset)\delta^{-1}(X)} - \frac{\mathbb{1}_{S=1}(1-A)Y}{e(\emptyset)\delta^{-1}(X)}\right]$$

where $e(\emptyset) = P(A = a)$ and $\delta(X) = \frac{P(S=2|X)}{P(S=1|X)} \times \frac{1}{P(S=2|X)}$. In other words, the PATE can be re-written in terms of the trial data ($S = 1$) and the relationship between the trial sample and the target population ($\delta(X)$). This extends upon a result from Cole and Stuart (2010) by recognizing that

$$\underbrace{\left(\frac{P(S=2|X)}{P(S=1|X)}\right)}_{\text{Transportability weights}} \times \underbrace{\left(\frac{1}{P(S=2|X)}\right)}_{\text{Survey weights}} = \left(\frac{1}{P(S=1|X)}\right) \tag{2.1}$$

Furthermore, when $P(S = 1|X)$ cannot be estimated directly, as is often the case since RCTs are not equipped with "trial selection weights," it can be conveniently decomposed into two estimable quantities: the inverse odds of sample vs. survey membership (transportability weights) and the inverse probability of survey sampling (survey weights).

Note that survey weights are commonly included as variables in publicly available government complex surveys. While some researchers have, in practice, incorporated survey weights when transporting from a trial to a complex survey sample, none have provided methodological details on how exactly they were used, nor have they provided any justification for their use. Without such reasoning, it is plausible that some researchers may apply current generalization methods with complex survey population data while neglecting to incorporate the survey weights. Suppose we were to *ignore* the survey weights altogether. We can refer to this quantity as follows:

$$\Delta_{\text{transport}} = E[Y(1) - Y(0)|S = 2] = E\left[\frac{\mathbb{1}_{S=1}AY}{e(\emptyset)\gamma^{-1}(X)} - \frac{\mathbb{1}_{S=1}(1-A)Y}{e(\emptyset)\gamma^{-1}(X)}\right]$$

Note that $\Delta_{\text{transport}}$ differs from Δ in that we substitute $\delta(X)$ for $\gamma(X)$ such that $\gamma(X) = \frac{P(S=2|X)}{P(S=1|X)}$. Observe that

$$\Delta_{\text{transport}} = \Delta \times P(S = 2|X)$$

In other words, if survey weights are ignored, then the estimate of $\Delta_{\text{transport}}$ will be biased as an estimate for Δ, the PATE, by a factor of $P(S = 2|X)$, or the probability of being sampled for the survey given covariates X. Note that $\Delta_{\text{transport}}$ will only be equal to Δ when $P(S = 2|X) = 1$, or when the survey is either a simple random sample of the population, *or* it is the entire finite target population.

2.3 Estimating the PATE, Δ

We now discuss three different potential estimators to estimate Δ, the last of which will incorporate the complex survey weights. First, if we were to use the trial data alone ($S = 1$) to estimate Δ, we could use the following naive estimator:

$$\hat{\Delta}_{\text{naive}} = \frac{\sum_i \mathbb{1}_{S_i=1} A_i Y_i}{\sum_i \mathbb{1}_{S_i=1} A_i} - \frac{\sum_i \mathbb{1}_{S_i=1}(1 - A_i) Y_i}{\sum_i \mathbb{1}_{S_i=1}(1 - A_i)}$$

However, recall from Section 2.2 that $\hat{\Delta}_{\text{naive}}$ will be a biased estimate of the PATE if the treatment effect is moderated by a pre-treatment covariate and sample selection also depends on that covariate. To improve upon this, we can transport the estimate to the survey ($S = 2$) with the following inverse-odds of sample membership weighted estimator:

$$\hat{\Delta}_{\text{transport}} = \frac{\sum_i \mathbb{1}_{S_i=1} A_i Y_i \hat{\gamma}_i}{\sum_i \mathbb{1}_{S_i=1} A_i \hat{\gamma}_i} - \frac{\sum_i \mathbb{1}_{S_i=1}(1 - A_i) Y_i \hat{\gamma}_i}{\sum_i \mathbb{1}_{S_i=1}(1 - A_i) \hat{\gamma}_i}$$

where $\hat{\gamma}_i = \gamma(X_i, \hat{\beta})$ and $\gamma(X, \beta) = \frac{P(S=2|X)}{P(S=1|X)}$. Note that $\hat{\gamma}_i(X_i, \hat{\beta})$ can be estimated parametrically by fitting a logistic regression model of sample membership (trial vs. survey) conditional on pre-treatment observables in a dataset in which the trial and survey data have been concatenated. While $\hat{\Delta}_{\text{transport}}$ may be unbiased for $\Delta_{\text{transport}}$ (Westreich et al., 2017), it will still be a biased estimate of the PATE, Δ, if the complex survey is not "self-weighting." We therefore propose a modified version of this estimator, one that incorporates the complex survey weights relating the survey sample ($S = 2$) to the target population:

$$\hat{\Delta}_{\text{svy.wtd}} = \frac{\sum_i \mathbb{1}_{S_i=1} A_i Y_i \hat{\delta}_i}{\sum_i \mathbb{1}_{S_i=1} A_i \hat{\delta}_i} - \frac{\sum_i \mathbb{1}_{S_i=1} (1 - A_i) Y_i \hat{\delta}_i}{\sum_i \mathbb{1}_{S_i=1} (1 - A_i) \hat{\delta}_i}$$

where $\hat{\delta}_i = \delta(X_i, \hat{\beta})$ and $\delta(X, \beta) = \frac{P(S=2|X)}{P(S=1|X)} \times \frac{1}{P(S=2|X)}$. Here, $\hat{\delta}_i$ can be estimated parametrically by fitting a model for $\frac{P(S=2|X)}{P(S=1|X)}$, and multiplying the resulting estimated transportability weights by the survey weights. If all related covariates are observed and accounted for, then this estimator is unbiased for the PATE, directly following a result from Buchanan et al. (2018) by applying the equality in Equation 2.1. We will now present a simple example to compare each of these estimators when weighting a trial to a target population based on a single covariate.

2.3.1 Toy Example

In order to highlight the consequences of ignoring survey weights when estimating the PATE, consider the scenario in Table 2.1. Suppose that in the true target population of interest, 50% of people are above the age of 40, while the other 50% are 40 or younger. Suppose data on the full target population are not available, but a survey is conducted among the target population members, where 200 individuals over the age of 40 and 300 individuals who are 40 or younger are sampled. In order for the survey to be representative of the target population according to dichotomous age, survey weights are constructed as the inverse probability of being sampled into the survey given age category. The older category individuals are given a weight of $\frac{5}{2}$ while the younger category individuals are assigned a weight of $\frac{5}{3}$. In doing so, older survey participants receive greater weight than younger ones to reflect that older individuals are undersampled in the survey.

Table 2.1: Toy example of a population (not observed), a survey sampled from the population with weights to reflect the population demographics distribution, and a trial sampled from the population (by convenience sampling)

	$E[Y(1) - Y(0)\|X]$	Target pop	Survey	RCT
age > 40	2	500	200	100
age $\leq$ 40	4	500	300	50

Next, suppose a randomized trial is conducted among a convenience sample from the population, and among the recruited participants, $\frac{2}{3}$ of them are 40 years or older. Additionally, suppose that the treatment effect is truly moderated by age, where younger participants experience twice the average effect as older participants. Observe that while older members are *undersampled* in the survey, they are *oversampled* in the trial, and since age moderates treatment effect and differs between the trial and population, the RCT findings will not generalize well to this target population.

First, the true PATE can be calculated by averaging over the stratum-specific treatment effects in the target population:

$$\Delta = \sum_x E[Y(1) - Y(0)|X = x]P(X = x) = 2 \times 0.5 + 4 \times 0.5 = 3$$

Next, the naive trial estimator for the PATE can be estimated as follows:

$$\hat{\Delta}_{\text{naive}} = \sum_x E[Y(1) - Y(0)|X = x]P(X = x|S = 1) = 2 \times \frac{2}{3} + 4 \times \frac{1}{3} = 2.67$$

As expected, the naive estimate is an underestimate of the PATE because the trial oversampled older participants, while the treatment has a stronger effect for younger participants. If we apply the standard transportability weighting methods using the survey as the target population dataset, and if we *ignore the survey weights*, we would weight trial members by the inverse odds of trial participation conditional on their age category. Older trial participants would be given a weight of $\frac{200}{100} = 2$, and younger trial participants would be given a weight of $\frac{300}{50} = 6$. We would therefore estimate the transported estimate as follows:

$$\begin{aligned}
\hat{\Delta}_{\text{transport}} &= \frac{\sum_x E[Y(1) - Y(0)|X = x]P(X = x|S = 1)\frac{P(S=2|X=x)}{P(S=1|X=x)}}{\sum_x P(X = x|S = 1)\frac{P(S=2|X=x)}{P(S=1|X=x)}} \\
&= \frac{2 \times \frac{2}{3} \times 2 + 4 \times \frac{1}{3} \times 6}{\frac{2}{3} \times 2 + \frac{1}{3} \times 6} \\
&= 3.2
\end{aligned}$$

As a result, the estimate is unbiased for the ATE in the *survey*; however, it is still biased as an estimate of the PATE. Our inferences here reflect that *older* participants are oversampled in the survey, and so in this case we are overestimating the true PATE. Finally, if we utilize the survey weights by *multiplying* the inverse odds transportability weights by the inverse probability of survey selection, we would obtain weights of $\frac{200}{100} \times \frac{500}{200} = 5$ and $\frac{300}{50} \times \frac{500}{300} = 10$ for the older and younger trial participants, respectively, thereby accurately weighting them to the *target population* age distribution. We would estimate the PATE using this approach as follows:

$$\begin{aligned}
\hat{\Delta}_{\text{svy.wtd}} &= \frac{\sum_x E[Y(1) - Y(0)|X = x] P(X = x|S = 1) \frac{P(S=2|X=x)}{P(S=1|X=x)} \frac{1}{P(S=2|X=x)}}{\sum_x P(X = x|S = 1) \frac{P(S=2|X=x)}{P(S=1|X=x)} \frac{1}{P(S=2|X=x)}} \\
&= \frac{2 \times \frac{2}{3} \times 5 + 4 \times \frac{1}{3} \times 10}{\frac{2}{3} \times 5 + \frac{1}{3} \times 10} \\
&= 3
\end{aligned}$$

Observe that our estimate of the PATE is now unbiased, as we are accounting for the fact that our survey is not “self-weighting" and the survey weights must therefore be used to make inferences relevant to the true target population of interest.

2.3.2 Estimating $\hat{\Delta}_{\text{svy.wtd}}$ with a weighted sample membership model

When accounting for a small set of covariates, such as in the example above, one can directly construct and multiply the transportability weights by the survey weights. When using a survey equipped with pre-estimated survey weights, though, this is not plausible. We therefore propose a two-stage weighting approach, where we first weight the sample membership model using the survey weights before constructing the inverse odds transportability weights. This is equivalent to the multiplication of weights in the simple approach above, because by weighting survey participants in the sample membership model, we are recognizing that each participant represents a particular number of individuals in the true target population. For example, if a survey participant has a probability of survey selection of 0.02, the corresponding weight of $\frac{1}{0.02} = 50$ suggests that the individual should count for 50 people in the population when estimating population effects with the survey. Weighting the survey participants in the sample membership model allows us to therefore compare the trial demographics to the target population, and *not* to the survey sample.

The first step entails fitting a weighted logistic regression model of sample membership using a pseudo-likelihood approach (Pfeffermann, 1993), where trial participants are assigned a weight of 1 while survey participants are assigned weights equal to their inverse probability of survey selection. Again, these weights are typically included in complex survey datasets and are meant to be used in analyses to relate the survey back to the target population of interest. The second step entails using the predicted probabilities from the sample membership model, $\hat{e}_i$, to construct the inverse odds weights ($\hat{\delta}_i$) that are used to estimate $\hat{\Delta}_{\text{svy.wtd}}$, where trial participants are assigned a weight of $\frac{1-\hat{e}_i}{\hat{e}_i}$ and survey participants are assigned a weight of 0. It is important to note that, in theory, this approach will yield an unbiased estimate of Δ only when we account for *all* covariates that impact treatment effect heterogeneity and sample selection. However, it may be the case that certain variables used to construct the survey weights may not be available in the trial data, or even in the survey dataset itself. In other words, if a moderator is accounted for in the survey weights, but cannot be directly accounted for in the transportability weights as well, the PATE estimate may still be biased. With this in mind, we now describe a simulation study to compare our two-stage weighting approach to the standard transported estimator and the naive trial sample estimator.

2.4 Simulation

We conducted a simulation study to assess the performance of the two-stage weighting approach described in the Section 2.3. We first simulated a finite population of size $N = 1000000$ with six covariates using the multivariate Normal distribution with mean vector 0, and a variance-covariance matrix where each variable had variance 1, and pair-wise correlation (i.e. X_1 and X_2, X_3 and X_4, X_5 and X_6) of ρ. We paired the covariates in this way and varied ρ to look at scenarios where a covariate related to the sample selection mechanisms was not available in the analytic datasets, but a variable correlated with the missing covariate was available for use in its place. For example, survey participants may be sampled proportional to their zipcode, but the survey dataset might only include state as a geographic indicator for privacy purposes.

We then assigned probabilities of survey selection and trial selection to everyone in the population according to the following two models:

$$P(S_i = 1) = \text{expit}[\gamma_1(X_{1i} + X_{2i} + 2X_{3i} + 0X_{4i} + X_{5i} + 0X_{6i})]$$

$$P(S_i = 2) = \text{expit}[\gamma_2(2X_{1i} + 0X_{2i} + X_{3i} + X_{4i} + X_{5i} + 0X_{6i})]$$

We used scaling parameters γ_1 and γ_2 to control the magnitude of difference between the two samples and the population, while fixing the relative impacts of each covariate for each model. As the scaling parameters increase, the samples differ more greatly from the target population. The coefficients for the covariates were set to different values in each model to ensure that the sampling mechanisms for the trial and the survey differed from one another. Next, we generated potential outcomes for the entire population as $Y(0) \sim N(0, 1)$ and $Y(1) \sim N(2 + \gamma_3[\sum_{i=1}^{6} X_i], 1)$, such that the $PATE = 2$, and the γ_3 scaling parameter controlled the amount of treatment effect heterogeneity due to the covariates. Note that when $\gamma_3 = 0$ (no treatment effect heterogeneity), all of the PATE estimates should be unbiased.

In each simulation run, we then randomly sampled approximately 600 trial participants and approximately 4000 survey participants according to each individual's respective selection probabilities. In order to do this, we scaled each individual's originally generated $P(S_i = 1)$ by 0.0006 and their $P(S_i = 2)$ by 0.004, and estimated their probability of *not* being selected into either study as $P(S_i = 0) = 1 - P(S_i = 1) - P(S_i = 2)$. This type of scaling combined the specified selection models with the desired sampling proportions from the population, and allowed us to then randomly generate an S of $0, 1$ or 2 for each individual using a multinomial distribution. For the survey participants

($S = 2$), we retained their $P(S_i = 2)$ as their known survey sampling probabilities to construct survey weights. For the trial participants, we generated a randomized binary treatment variable A, as well as the observed outcome $Y = A \times Y(1) + (1 - A) \times Y(0)$.

Once the trial and survey data were simulated, we estimated the PATE in the following three ways: 1) Naive trial estimator ($\hat{\Delta}_{\text{naive}}$), 2) transported estimator (trial-to-survey) while ignoring the survey weights ($\hat{\Delta}_{\text{transport}}$) and 3) transported estimator (trial-to-survey) using the survey weights to fit a weighted sample membership model ($\hat{\Delta}_{\text{svy.wtd}}$). For the two transportability estimators, we predicted the probabilities of sample membership by fitting models with logistic regression, generalized boosted models (GBM) and the Super Learner. GBM is a flexible, iterative algorithm that has been demonstrated to perform well when used to estimate propensity scores in non-experimental studies, capturing nonlinear relationships between covariates and treatment assignment (McCaffrey, Ridgeway, & Morral, 2004). The Super Learner fits a series of models based on a user-specified library of methods, combining the resulting estimates such that the overall performance is no worse than the performance of the best individual method (Van der Laan, Polley, & Hubbard, 2007). We considered two Super Learner libraries (Luedtke & van der Laan, 2016; Moodie & Stephens, 2017), and fit each of the estimators described above using the 'WeightIt' package in R (Greifer, 2019; R Core Team, 2019).

Lastly, in order to investigate scenarios where variables used to construct survey weights are omitted from the survey dataset, we fit the sample membership model by using all of the covariates, by omitting X_1, and by omitting

X_1, X_3 *and* X_5. To evaluate the performance of each method, we calculated the bias and the empirical 95% coverage of each estimator, using $PATE = 2$ as the truth. Standard error estimates were obtained by using a standard sandwich variance estimator. We also calculated coverage for a subset of simulation scenarios using a stratified double bootstrapping approach, in which both the trial and the survey were sampled with replacement upon each bootstrap run. Strata for survey re-sampling were defined by survey probability deciles, and survey weights in each bootstrap sample were adjusted according to Valliant et al. (2013) (see Appendix A for details). The results presented in the next section are averaged over 1000 simulation runs, and are stable to the 2nd decimal place across different seeds.

2.4.1 Simulation Results

We now present the findings of the simulation study. Given the number of parameters to vary, we present figures where $\rho = 0.3$ (pairwise X correlation) and $\gamma_3 = 0.3$ (treatment effect heterogeneity). Note though that as expected, when $\gamma_3 = 0$, all estimators were unbiased for the PATE across all scenarios.

When $\rho = 0.3$ and $\gamma_3 = 0.3$, Figure 2.2 shows the bias of the three PATE estimators across simulation scenarios. Each column signifies a different setting regarding which variables are omitted from the sample membership model: on the left, all variables are included, and on the far right, X_1, X_3 and X_5 are all missing from the analytic datasets, but they were used to calculate the survey weights in the survey. Within each plot, the x-axis depicts the absolute standardized mean difference (ASMD) of the predicted probability of survey sampling between the survey sample and the target population (see Figure A.2 for the relationship between γ_2 and ASMD). In other words, moving from left to right along the x-axis, the survey sample becomes increasingly different from the target population on baseline covariates. The top row depicts when $\gamma_1 = 0$, or when the trial is a simple random sample from the target population. Notice that the naive estimate is unbiased, as is the transported estimate that uses the survey weights. However, when the ATE is transported from a representative trial to a *non*-representative survey and the survey weights are not used, the transported ATE estimate becomes increasingly biased as the survey becomes less representative of the population. This suggests that if findings from a trial are already generalizable, yet researchers implement transportability weighting methods without survey weights to a complex survey that is not representative of the target population, then they may actually obtain a more biased PATE estimate than had they not transported at all.

As the trial differs more greatly from the target population (moving down the rows, $\gamma_1 = 0.3$ to $\gamma_1 = 0.9$), the naive trial estimate becomes increasingly biased as expected. When the survey is slightly different from the target population of interest, the transported estimate that ignores survey weights is *less* biased then the naive estimate. However, once the survey differs enough from the target population, ignoring the survey weights when transporting yields similar bias to the naive estimator, and in some cases, even greater bias. On the other hand, the transported estimate that uses the survey weights to fit a weighted sample membership model is uniformly less biased than the other estimators across all scenarios. In other words, it seems as though using the survey weights in the sample membership model can help prevent any additional bias introduced from the survey not being a simple random sample from the population.

Between the different methods used to fit the trial membership model, there is little to no difference in terms of PATE bias for $\hat{\Delta}_{\text{svy.wtd}}$, except for when the trial differs greatly from the target population. In such cases, predicting the probability of trial membership using GBM appears to yield the least biased ATE estimates, with notable differences in performance between the two SuperLearner libraries considered.

Next, observe that the transported estimators perform best when the selection model is fit using all covariates used to calculate the survey weights. However, when one of the variables influencing survey selection (i.e. X_1) is not available in the survey dataset, the bias of the transported estimators increases, and continues to increase as fewer variables impacting survey selection are included in the analytic dataset. However, as the pairwise correlation of the missing and non-missing covariates increases, the bias decreases. In other words, and not surprisingly, if X_1 is unavailable to use in the sample membership model, but X_2 is available, the more X_2 and X_1 are correlated, the less it matters that X_1 is missing in terms of bias.

Figure 2.3 shows the empirical 95% coverage of the three estimators across simulation scenarios. Note that a standard sandwich variance estimator was used for all weighting approaches here, and results were fairly similar when using the double bootstrap approach as well (see Figure A.1). Across the top row, where the trial is representative of the target population, the coverage

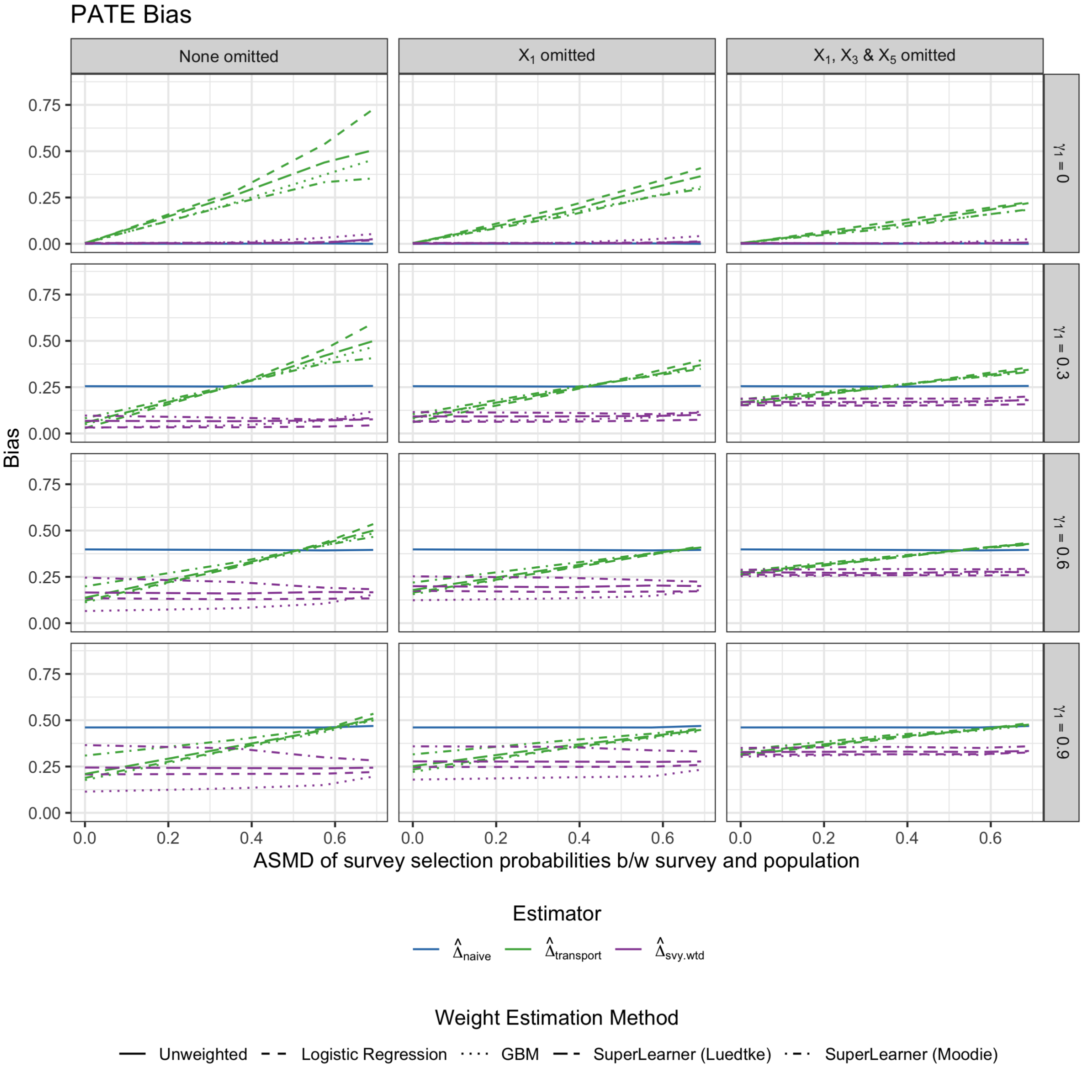


Figure 2.2: Bias of estimating the PATE by weighting method. Each column represents a different scenario of missing a variable used to calculate survey weights in the analytic survey dataset. From top to bottom row, the γ_1 "scale" parameter for how much the trial differs from the population by the Xs increases. The different colors represent the different weighting approaches: Naive trial estimate (blue), transported estimate ignoring the survey weights (green), and transported estimate using the survey weights (purple). This figure appears in color in the electronic version of this article.

of the naive estimator is around 95%, as expected (as is the coverage of the transported estimator using the survey weights). However, the coverage of the transported estimator that ignores the survey weights rapidly decreases as the survey becomes less representative of the population. Note that this corresponds to when the bias of the transported estimator without survey weights increases as well. As the trial becomes more different from the population, the empirical coverage of the naive estimator drops to zero. The transported estimator that incorporates the survey weights maintains much better coverage than the estimator that ignores the survey weights as the survey becomes less representative of the population. Also, when the trial differs substantially from the target population, the $\hat{\Delta}_{\text{svy.wtd}}$ estimate using GBM to fit the trial membership model results in the best coverage of the $\hat{\Delta}_{\text{svy.wtd}}$ estimates. The variability in the performance of $\hat{\Delta}_{\text{svy.wtd}}$ using the two Super Learner libraries is also notable, highlighting the method's sensitivity to library choice. Lastly, note that the transported estimator performs best when all variables included in the survey selection model are available in the survey dataset, and the empirical coverage declines as fewer of those variables become available for use in the sample membership transportability model (as ρ increases, the empirical coverage improves slightly across scenarios as well).

2.5 Data Examples

We now present two applications of these methods to generalizing trial findings to well-defined target populations. First, we generalize findings from

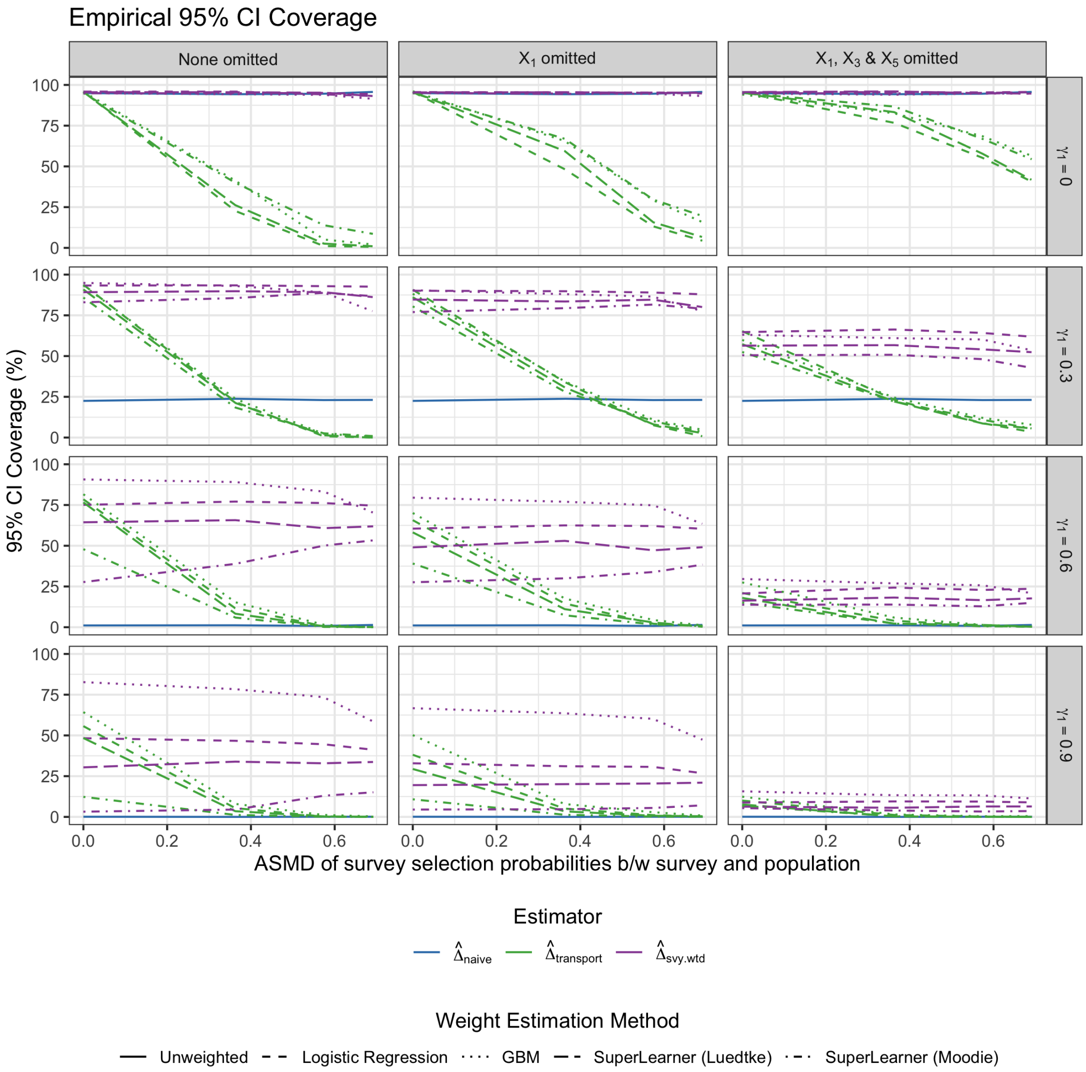


Figure 2.3: Empirical 95% coverage of the PATE estimates by weighting method. Each column represents a different scenario of missing a variable used to calculate survey weights in the analytic survey dataset. From top to bottom row, the γ_1 "scale" parameter for how much the trial differs from the population by the Xs increases. The different colors represent the different weighting approaches: Naive trial estimate (blue), transported estimate ignoring the survey weights (green), and transported estimate using the survey weights (purple). This figure appears in color in the electronic version of this article.

PREMIER, a lifestyle intervention trial for reducing blood pressure, to the National Health and Nutrition Examination Survey (NHANES). Next, we generalize results from CTN-0044, a trial examining the use of a web-based intervention for substance use disorder (SUD) treatment, to the National Survey on Drug Use and Health (NSDUH). In both examples, data on the respective target populations come from publicly available government surveys with complex survey sampling designs, where each participant is assigned a survey weight indicative of the number of individuals they represent in the target population. For each example, we illustrate the importance of utilizing the survey weights when comparing covariate distributions between the trial and survey, and demonstrate how the use of the survey weights affects PATE estimation. Given the simulation findings, we fit the sample membership model using GBM in both examples.

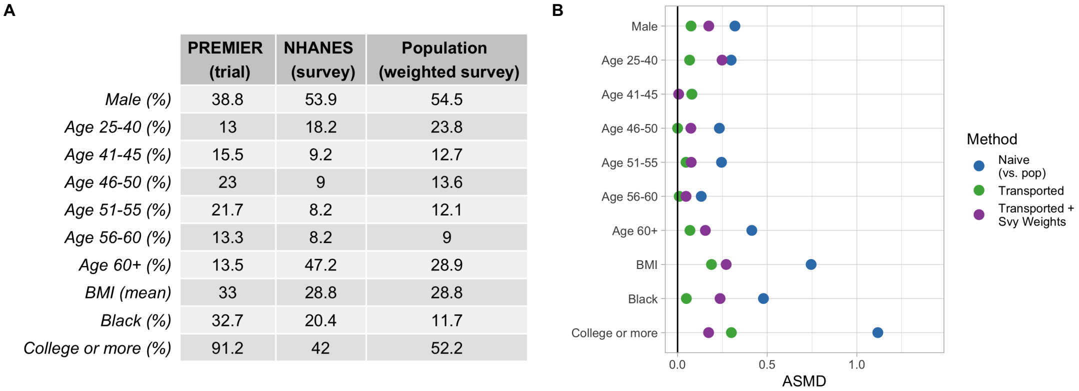
2.5.1 Lifestyle Intervention Trial for Blood Pressure Reduction

PREMIER was a multi-center randomized trial in which 810 participants were randomized to either one of two behavioral interventions, comprised of a mix of diet and exercise recommendations, or to standard care. The primary goal of the trial was to study the effect of these lifestyle interventions on blood pressure reduction. The original report on the trial found evidence supporting the interventions' effectiveness on blood pressure reduction, and concluded that "results from PREMIER should influence policy pertaining to implementation of lifestyle modification in the contemporary management of patients with above-optimal blood pressure through stage 1 hypertension" (Svetkey et al., 2003). For illustrative purposes, we combine the two intervention arms into a single “lifestyle intervention treatment" group, and select our outcome of interest as change in systolic blood pressure (SBP) between baseline and 6-month study followup.

We will now further investigate how these findings generalize to a potentially policy-relevant target population. To do so, we use population data from NHANES, a national survey funded by the Centers for Disease Control and Prevention (CDC) with extensive measures on participants’ dietary behaviors and health outcomes. Using a complex and multistage probability-based sampling design, NHANES participants are carefully sampled according to sex, age, race, ethnicity and income, resulting in a sample that is representative of the entire non-institutionalized civilian US population (CDC, 2003). To define the target population of interest, we subset the NHANES sample to individuals who are 25 years of age or older with BMI between 18.5 and 40 (due to PREMIER inclusion criteria). To better determine how PREMIER findings may impact a population of adults with “above-optimal blood pressure," we further limit the NHANES sample to individuals with either SBP greater than or equal to 120 *or* diastolic blood pressure (DBP) greater or equal to 80. This results in a sample size of 2180 representing a population of over 85 million US adults.

Figure 2.4A shows the covariate distributions in the trial and survey samples, as well as in the weighted survey sample (indicating the target population of interest). Observe that while some variables, such as sex and BMI, are distributed quite similarly between the unweighted and weighted NHANES samples, other variables, such as race, age and education, differ a fair amount between the two. These differences show the NHANES survey sampling methodology, and how the true population characteristics may differ from the raw analytic sample. If we generalize to the NHANES survey sample (i.e. fit the transportability estimator ignoring the survey sampling weights), we would be generalizing to a population that is younger, less educated, and more racially diverse than our true target population of interest. Figure 2.4B shows the covariate balance between the trial and target population before and after transport weighting. Note that weighting the trial to resemble either the NHANES sample or the target population results in better covariate balance; however, only the latter is truly relevant to our interests.

The effect of the lifestyle intervention on change in SBP is shown in Figure 2.5, with the naive trial estimate on the left, the transported estimate in the middle, and the transported estimate using survey weights on the right. The naive trial estimate of -4.66 and 95% confidence interval of (-6.10, -3.23) indicate a positive effect of the lifestyle intervention recommendations in lowering systolic blood pressure among study participants, as originally reported in the trial findings. In this example, there are no substantial differences between the naive estimates and the transported estimates, nor between the two transported estimates (ignoring vs. incorporating the survey weights). Note, though, that both weighted estimators have larger standard errors. Given the consistent estimates, these generalized findings provide further evidence to support the original trial's claims, that PREMIER's results should be used to influence blood pressure management policies related to persons with



	PREMIER (trial)	NHANES (survey)	Population (weighted survey)
Male (%)	38.8	53.9	54.5
Age 25-40 (%)	13	18.2	23.8
Age 41-45 (%)	15.5	9.2	12.7
Age 46-50 (%)	23	9	13.6
Age 51-55 (%)	21.7	8.2	12.1
Age 56-60 (%)	13.3	8.2	9
Age 60+ (%)	13.5	47.2	28.9
BMI (mean)	33	28.8	28.8
Black (%)	32.7	20.4	11.7
College or more (%)	91.2	42	52.2

Figure 2.4: A) Covariate Distributions in PREMIER (trial) and NHANES (survey sample), along with the weighted NHANES sample (target population). B) Absolute standardized mean difference (ASMD) of covariates between the trial and target population. Points in blue reflect covariate differences between the raw trial sample and the weighted survey sample (i.e. the target population demographics). Points in green show the differences between the transport-weighted trial and survey sample. Points in purple show the differences between the transport-weighted trial and population (where the trial is weighted to be more similar to the target population).

above-optimal blood pressure in the United States.

2.5.2 Web-Based Intervention for Treating Substance Use Disorders

We now turn to our second illustrative example using a trial from the Clinical Trials Network (CTN), a publicly available data repository for substance use-related RCTs funded by the National Institute of Drug Abuse (NIDA). The trial of interest, CTN-0044, evaluated the effectiveness of Therapeutic Education System (TES), a web-based behavioral intervention including motivational incentives, as a supplement to SUD treatment. A total of 507 individuals in treatment for SUDs were randomized to either treatment as usual or treatment plus TES, and the reported trial results suggested that TES successfully reduced treatment dropout and improved upon abstinence (Campbell et al.,

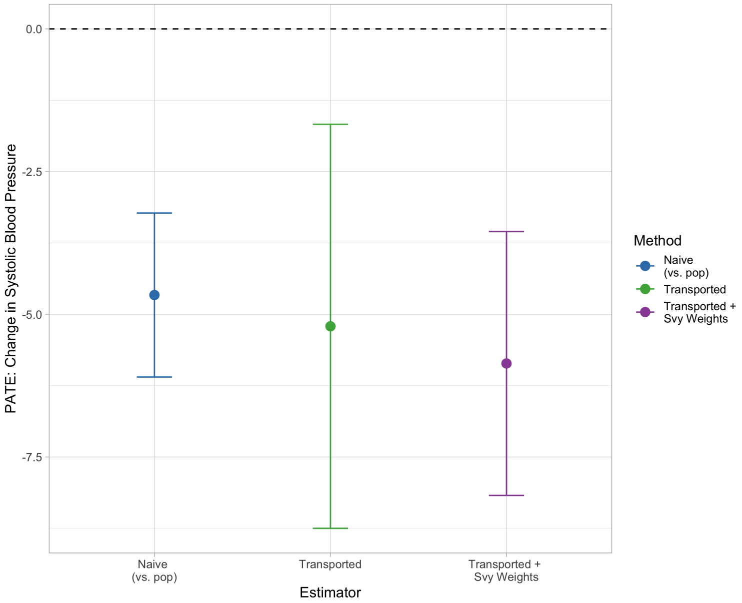


Figure 2.5: Blood pressure reduction PATE estimates by transportability method. Points in blue reflect the naive PATE estimate, points in green show the transported PATE estimate ignoring survey weights. Points in purple show the survey-weighted transportability estimate.

2014). Our outcome of interest is a binary indicator of drug and alcohol abstinence in the last week of the study.

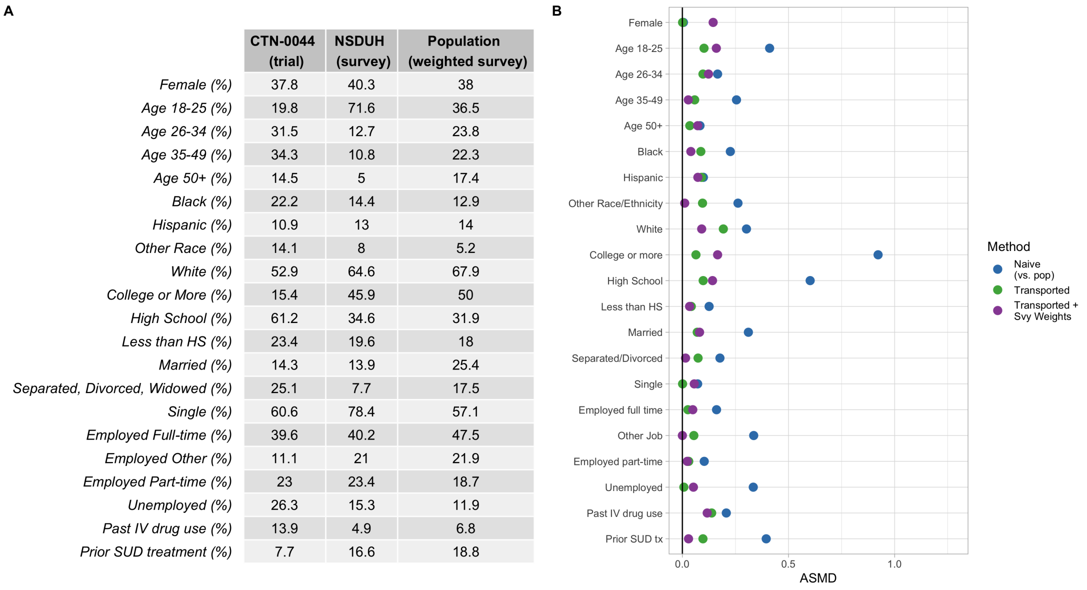
We generalize these findings from CTN-0044 to a population of US adults seeking treatment for substance use disorders using NSDUH, a survey on drug use in the United States. In its sampling design, NSDUH systematically over-samples adults over the age of 26 in order to better estimate drug use and mental health issues in the US. This suggests that the raw NSDUH survey sample is likely not reflective of the target population on key demographics. We subset the NSDUH sample to individuals over the age of 18 who have reported any illicit drug use in the past 30 days in order to best reflect our target population of interest. The resulting NSDUH sample has 5645 participants representing a target population of around 20 million people.

The distribution of covariates across the trial and survey samples are shown in Figure 2.6A. Note that, pre-transport-weighting, there are substantial differences between the trial and raw survey samples with respect to age, though when the survey sample is weighted to the target population using the survey weights, these age differences decrease. Other variables like race, education and prior substance use treatment are actually *more* different between the trial and target population than they are between the trial and unweighted NSDUH sample. This further highlights the importance of incorporating the survey weights in order to make inferences on the true target population of interest when transporting. Figure 2.6B shows the covariate balance between the trial sample and (survey-weighted) target population before and after weighting. Points in green show covariate balance when the trial is weighted to the raw survey sample, while points in purple show covariate balance when the trial is weighted to the target population (the survey-weighted survey sample). Overall, both weighting methods yield better balance (and therefore better resemblance) between the trial and the population, though it should be noted again that only the points in purple reflect when the trial is weighted to resemble the true target population (i.e. the survey weights are used in the sample membership model).

Figure 2.7 depicts the three PATE estimates, or the odds ratio of substance abstinence. As reported in the original trial, the naive estimate is statistically significant, with an odds ratio of 1.5 and 95% confidence interval of (1.02, 2.24), suggesting that TES was effective in increasing substance abstinence in the trial. However, when this estimate is transported to the NSDUH sample (middle, green), this point estimate drops to around 0.8 and the confidence interval width increases (0.28, 2.48). While the lower odds ratio may suggest qualitative differences in TES' effectiveness, the transported estimate indicates no significant difference in abstinence rates between the two treatment arms in the NSDUH sample. When the survey weights are included in the sample membership model, and the estimate therefore generalized to the target population of interest, the wide confidence interval of (0.90, 3.55) indicates a similar not-significant conclusion, though the point estimate of 1.79 more closely mirrors what was estimated in the original trial. This example highlights that if the survey weights are left out when making generalizations, different qualitative conclusions may be reached.

2.6 Conclusion

Existing methods for improving RCT generalizability with propensity score-type weights make an implicit assumption about the population data: that they are either 1) a simple random sample drawn from the true target population, or 2) the *complete* finite target population. When transporting trial findings to a population dataset derived from a complex survey, this assumption no longer holds. Our work demonstrates that it is crucial to incorporate the survey weights from the complex survey population data in order to obtain the best estimate of the PATE with these methods. Omitting the survey weights can be thought of as generalizing to an entirely different population, one that has the demographics of the survey sample rather than the target population of



	CTN-0044 (trial)	NSDUH (survey)	Population (weighted survey)
Female (%)	37.8	40.3	38
Age 18-25 (%)	19.8	71.6	36.5
Age 26-34 (%)	31.5	12.7	23.8
Age 35-49 (%)	34.3	10.8	22.3
Age 50+ (%)	14.5	5	17.4
Black (%)	22.2	14.4	12.9
Hispanic (%)	10.9	13	14
Other Race (%)	14.1	8	5.2
White (%)	52.9	64.6	67.9
College or More (%)	15.4	45.9	50
High School (%)	61.2	34.6	31.9
Less than HS (%)	23.4	19.6	18
Married (%)	14.3	13.9	25.4
Separated, Divorced, Widowed (%)	25.1	7.7	17.5
Single (%)	60.6	78.4	57.1
Employed Full-time (%)	39.6	40.2	47.5
Employed Other (%)	11.1	21	21.9
Employed Part-time (%)	23	23.4	18.7
Unemployed (%)	26.3	15.3	11.9
Past IV drug use (%)	13.9	4.9	6.8
Prior SUD treatment (%)	7.7	16.6	18.8

Figure 2.6: A) Covariate Distributions in CTN-0044 (trial) and NSDUH (survey sample), along with the weighted NSDUH sample (target population). B) Absolute standardized mean difference (ASMD) of covariates between the trial and target population. Points in blue reflect covariate differences between the raw trial sample and the weighted survey sample (i.e. the target population demographics). Points in green show the differences between the transport-weighted trial and survey sample. Points in purple show the differences between the transport-weighted trial and population (where the trial is weighted to be more similar to the target population).

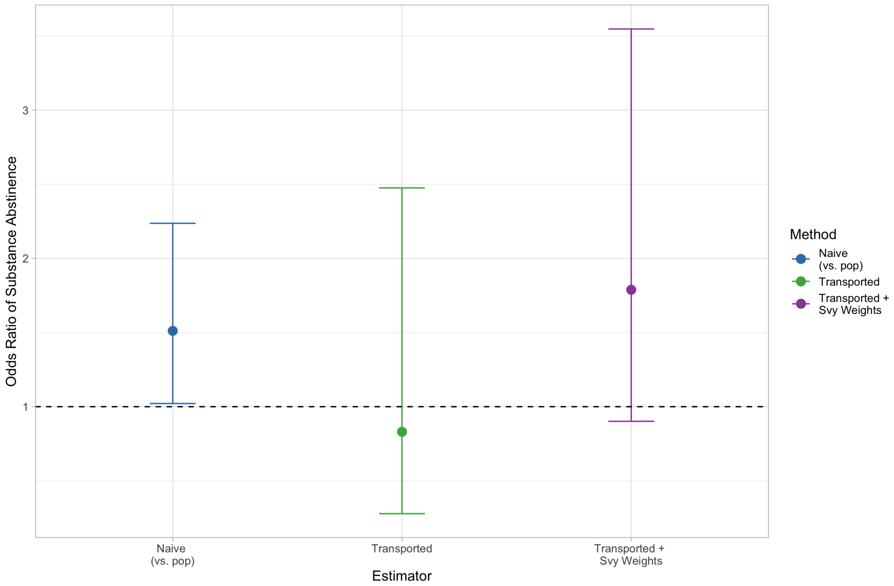


Figure 2.7: Substance abstinence PATE estimates by transportability method. Points in blue reflect the naive PATE estimate, points in green show the transported PATE estimate ignoring survey weights. Points in purple show the survey-weighted transportability estimate.

interest. While the demographic differences between a survey sample and its intended target population may not be that large for some analytic survey datasets, it can be particularly noticeable for others where great amounts of over- or under-sampling of certain groups are implemented.

Our work has shown that fitting a sample membership model weighted by survey weights can only improve upon our ability to draw population-level inferences from RCTs, and that failing to do so (i.e. using standard transportability weights alone) may actually result in *more* biased estimates. Given that complex survey data often come ready for use with a variable containing the necessary survey weights, implementing this approach does not require specifying any additional models other than those needed for the standard transportability weighting methods. Still, there are still a few limitations to this work. First, as noted earlier in this paper, we can obtain an unbiased estimate the PATE when we assume that *all* covariates impacting survey selection, trial selection, *and* treatment effect heterogeneity are fully observed and accounted for in both datasets. In practice, certain variables used to construct the survey weights may not be publicly available at the individual-level in the survey sample. While we demonstrated the performance of these methods when we use a correlated proxy for one such variable, it is also conceivable that certain key covariates may be unobserved in one or both datasets completely. Further research is needed to extend upon sensitivity analyses for partially and fully unobserved treatment effect modifiers, particularly when the population data come from a complex survey. Second, while we explored the benefit of double-bootstrapping methods for variance estimation, there may be additional concerns over uncertainty introduced by using a small survey sample that represents a huge target population. Additional research is warranted to assess the impact of the proportion of the population sampled on estimate variability. Finally, the propensity score-type weighting method explored in this paper is only one post-hoc statistical approach for estimating population effects from RCTs. Outcome-model-based approaches have also been shown beneficial, where a model is fit using trial data, and predictions are generated under treatment and control conditions in the target population data. Future work should build upon such methods when using complex survey population data as well. Nevertheless, our two-stage weighting method will ultimately allow researchers to draw more accurate inferences from trials to be used in policy formation and population-level decision making.

References

Ackerman, B., Schmid, I., Rudolph, K. E., Seamans, M. J., Susukida, R., Mojtabai, R., & Stuart, E. A. (2019). Implementing statistical methods for generalizing randomized trial findings to a target population. *Addictive behaviors*, *94*, 124–132.

Austin, P. C., Jembere, N., & Chiu, M. (2018). Propensity score matching and complex surveys. *Statistical methods in medical research*, *27*(4), 1240–1257.

Bell, B. A., Onwuegbuzie, A. J., Ferron, J. M., Jiao, Q. G., Hibbard, S. T., & Kromrey, J. D. (2012). Use of design effects and sample weights in complex health survey data: A review of published articles using data from 3 commonly used adolescent health surveys. *American Journal of Public Health*, *102*(7), 1399–1405.

Brunell, T. L., & DiNardo, J. (2004). A propensity score reweighting approach to estimating the partisan effects of full turnout in american presidential elections. *Political Analysis*, *12*(1), 28–45.

Buchanan, A. L., Hudgens, M. G., Cole, S. R., Mollan, K. R., Sax, P. E., Daar, E. S., ... Mugavero, M. J. (2018). Generalizing evidence from randomized trials using inverse probability of sampling weights. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, *181*(4), 1193–1209.

Cain, L. E., & Cole, S. R. (2009). Inverse probability-of-censoring weights for the correction of time-varying noncompliance in the effect of randomized highly active antiretroviral therapy on incident aids or death. *Statistics in medicine*, *28*(12), 1725–1738.

Campbell, A. N., Nunes, E. V., Matthews, A. G., Stitzer, M., Miele, G. M., Polsky, D., ... Kyle, T. L., et al. (2014). Internet-delivered treatment for substance abuse: A multisite randomized controlled trial. *American Journal of Psychiatry*, *171*(6), 683–690.

CDC. (2003). National health and nutrition examination survey data.

Cole, S. R., & Stuart, E. A. (2010). Generalizing evidence from randomized clinical trials to target populations: The actg 320 trial. *American journal of epidemiology*, *172*(1), 107–115.

Dahabreh, I. J., Robertson, S. E., Tchetgen, E. J., Stuart, E. A., & Hernán, M. A. (2018). Generalizing causal inferences from individuals in randomized trials to all trial-eligible individuals. *Biometrics*.

Dahabreh, I. J., Hernán, M. A., Robertson, S. E., Buchanan, A., & Steingrimsson, J. A. (2019). Generalizing trial findings in nested trial designs with subsampling of non-randomized individuals. *arXiv preprint arXiv:1902.06080*.

DuGoff, E. H., Schuler, M., & Stuart, E. A. (2014). Generalizing observational study results: Applying propensity score methods to complex surveys. *Health services research*, *49*(1), 284–303.

Gelman, A. et al. (2007). Struggles with survey weighting and regression modeling. *Statistical Science*, *22*(2), 153–164.

Greifer, N. (2019). *Weightit: Weighting for covariate balance in observational studies*. Retrieved from https://CRAN.R-project.org/package=WeightIt

Heckman, J. J., & Todd, P. E. (2009). A note on adapting propensity score matching and selection models to choice based samples. *The econometrics journal*, *12*(suppl_1), S230–S234.

Hill, J. L. (2011). Bayesian nonparametric modeling for causal inference. *Journal of Computational and Graphical Statistics*, *20*(1), 217–240.

Hoertel, N., Le Strat, Y., Blanco, C., Lavaud, P., & Dubertret, C. (2012). Generalizability of clinical trial results for generalized anxiety disorder to community samples. *Depression and Anxiety*, *29*(7), 614–620.

Imai, K., King, G., & Stuart, E. A. (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, T. R. (2006). Beyond efficacy: The star* d trial. *American Journal of Psychiatry*, *163*(1), 5–7.

Johnson, C. L., Dohrmann, S. M., Burt, V., & Mohadjer, L. K. (2014). National health and nutrition examination survey: Sample design, 2011–2014.

Kern, H. L., Stuart, E. A., Hill, J., & Green, D. P. (2016). Assessing methods for generalizing experimental impact estimates to target populations. *Journal of Research on Educational Effectiveness*, *9*(1), 103–127.

Lenis, D., Nguyen, T. Q., Dong, N., & Stuart, E. A. (2017). It's all about balance: Propensity score matching in the context of complex survey data. *Biostatistics*, *20*(1), 147–163.

Luedtke, A. R., & van der Laan, M. J. (2016). Super-learning of an optimal dynamic treatment rule. *The international journal of biostatistics*, *12*(1), 305–332.

Lumley, T. (2010). Complex surveys: A guide to analysis using r. hoboken: Johnwiley & sons. Inc.

McCaffrey, D. F., Ridgeway, G., & Morral, A. R. (2004). Propensity score estimation with boosted regression for evaluating causal effects in observational studies. *Psychological methods*, *9*(4), 403.

Miratrix, L. W., Sekhon, J. S., Theodoridis, A. G., & Campos, L. F. (2018). Worth weighting? how to think about and use weights in survey experiments. *Political Analysis*, *26*(3), 275–291.

Moodie, E. E., & Stephens, D. A. (2017). Treatment prediction, balance, and propensity score adjustment. *Epidemiology*, *28*(5), e51–e53.

Morgan, P. L., Frisco, M. L., Farkas, G., & Hibel, J. (2010). A propensity score matching analysis of the effects of special education services. *The Journal of special education*, *43*(4), 236–254.

Mullinix, K. J., Leeper, T. J., Druckman, J. N., & Freese, J. (2015). The generalizability of survey experiments. *Journal of Experimental Political Science*, *2*(2), 109–138.

Nguyen, T. Q., Ackerman, B., Schmid, I., Cole, S. R., & Stuart, E. A. (2018). Sensitivity analyses for effect modifiers not observed in the target population when generalizing treatment effects from a randomized controlled trial: Assumptions, models, effect scales, data scenarios, and implementation details. *PloS one*, *13*(12), e0208795.

Olsen, R. B., Orr, L. L., Bell, S. H., & Stuart, E. A. (2013). External validity in policy evaluations that choose sites purposively. *Journal of Policy Analysis and Management*, *32*(1), 107–121.

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.

Pfeffermann, D. (1993). The role of sampling weights when modeling survey data. *International Statistical Review/Revue Internationale de Statistique*, 317–337.

R Core Team. (2019). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. Vienna, Austria. Retrieved from https://www.R-project.org/

Ridgeway, G., Kovalchik, S. A., Griffin, B. A., & Kabeto, M. U. (2015). Propensity score analysis with survey weighted data. *Journal of Causal Inference*, *3*(2), 237–249.

Rudolph, K. E., Díaz, I., Rosenblum, M., & Stuart, E. A. (2014). Estimating population treatment effects from a survey subsample. *American journal of epidemiology*, *180*(7), 737–748.

Shadish, W. R., Cook, T. D., & Campbell, D. T. (2002). *Experimental and quasi-experimental designs for generalized causal inference*. Wadsworth Cengage learning.

Stuart, E. A., & 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.

Stuart, E. A., Ackerman, B., & Westreich, D. (2017). Generalizability of randomized trial results to target populations: Design and analysis possibilities. *Research on Social Work Practice*, 1049731517720730.

Svetkey, L. P., Harsha, D. W., Vollmer, W. M., Stevens, V. J., Obarzanek, E., Elmer, P. J., … Aickin, M., et al. (2003). Premier: A clinical trial of comprehensive lifestyle modification for blood pressure control: Rationale, design and baseline characteristics. *Annals of epidemiology*, *13*(6), 462–471.

Tipton, E., & Matlen, B. J. (2019). Improved generalizability through improved recruitment: Lessons learned from a large-scale randomized trial. *American Journal of Evaluation*, 1098214018810519.

Valliant, R., Dever, J. A., & Kreuter, F. (2013). *Practical tools for designing and weighting survey samples*. Springer.

Van der Laan, M. J., Polley, E. C., & Hubbard, A. E. (2007). Super learner. *Statistical applications in genetics and molecular biology*, *6*(1).

Westreich, D., Edwards, J. K., Lesko, C. R., Stuart, E., & Cole, S. R. (2017). Transportability of trial results using inverse odds of sampling weights. *American journal of epidemiology*, *186*(8), 1010–1014.

Yang, M., Ganesh, N., Mulrow, E., & Pineau, V. (2018). Estimation methods for nonprobability samples with a companion probability sample. *Proceedings of the Joint Statistical Meetings, 2018*.

Zanutto, E. L. (2006). A comparison of propensity score and linear regression analysis of complex survey data. *Journal of data Science*, *4*(1), 67–91.

Chapter 3

Transportability of Outcome Measurement Error Correction: from Validation Studies to Intervention Trials

3.1 Introduction

Lifestyle intervention trials aim to establish how changes to human behavior, such as physical activity or food intake, can impact and improve health outcomes. In such trials, it is important to obtain accurate measures on these behaviors; however, reliable measures are often expensive and burdensome for participants to collect. Self-reported measures are therefore commonly collected as proxies of the truth. For example, food frequency questionnaires or interviewer-assisted 24-hour dietary recall may be administered in nutrition studies to quantify sodium intake instead of having participants routinely collect urine samples. While self-reported measures are more feasible to obtain, they are prone to measurement error, as participants may not be able to accurately quantify their behaviors, or may misreport their true actions (Willett, 2012).

Measurement error can lead to biased, less precise estimates of a treatment's effect on the outcome (Rothman, Greenland, & Lash, 2008). In nutrition intervention studies, self-reported dietary intake measures have been shown to differ from the truth both randomly and systematically, which impacts the inferences drawn on the effectiveness of such lifestyle interventions (Espeland et al., 2001; Forster, Jeffery, VanNatta, & Pirie, 1990; Natarajan et al., 2010). Much of the attention in the measurement error literature has been paid to when measurement error is present in either the exposure of interest or in covariates (Buonaccorsi, 2010; Carroll, Ruppert, Stefanski, & Crainiceanu, 2006; Keogh & White, 2014). Less work, however, has focused on correcting for misclassification or measurement error of study outcomes (Keogh, Carroll, Tooze, Kirkpatrick, & Freedman, 2016), which is particularly worrisome for trials that focus on self-reported behavioral outcomes (Spring et al., 2012; Spring et al., 2018). This is in part because measurement error in the outcome will only lead to a biased estimate of the average treatment effect if the error is differential with respect to the intervention (Natarajan et al., 2010).

Existing measurement error methods, whether for covariates or outcomes, often rely on the use of a validation sample, or a group of individuals where both the truth and the observed mis-measured value are recorded (Wong, Day, Bashir, & Duffy, 1999a; Wong, Day, & Wareham, 1999b). Such validation data can either be *internal*, where the reliable biomarkers are collected on a chosen subset of individuals within the primary study, or *external*, where a study is conducted on a separate set of individuals with the sole intention of modeling the error structure. Internal validation studies are typically more feasible to embed within a large observational study cohort (Jenab, Slimani, Bictash, Ferrari, & Bingham, 2009), while it is rather rare to see intervention trials, given added costs of biomarker sample collection and added burden to trial participants. External validation samples also typically only collect measures under a “usual care" setting, which usually corresponds to a control condition in a trial, making it infeasible to directly correct for the error under both the treatment *and* control conditions based on the information available to researchers. Siddique et al. (2019) developed methodology for modeling the measurement error under the control condition using an external validation sample, followed by sensitivity analyses to obtain a range of plausible values for the treatment effect.

While external validation samples play an important role in correcting for measurement error, concerns have been raised over external validation studies not always being “transportable," such that the measurement error correction from an external study may not accurately apply to the main study of interest (Bound, Brown, & Mathiowetz, 2001; Carroll et al., 2006; Courtemanche, Pinkston, & Stewart, 2015). Previous efforts to address transportability have involved combining external validation data with internal validation data (Lyles, Zhang, & Drews-Botsch, 2007), though such an approach cannot be implemented in intervention studies without any internal validation.

In this paper, we address the issue of transportability when using external validation data to correct for measurement error of a continuous outcome in lifestyle intervention trials. Using a potential outcomes framework, we formalize cases when assumptions of measurement error transportability are violated and quantify the resulting additional bias that is introduced when estimating the average treatment effect. We then propose a weighting method for transportability, calibrating the validation data to the intervention trial to better estimate the measurement error.

3.2 Definitions

In order to set up the transportability issue, we will first provide some definitions and describe the measurement error problem more formally. Let A denote treatment assignment (0 = control, 1 = treatment), let S denote sample membership (v = validation, rct = intervention study), and let n_s = the sample size of study s. Let Z denote the outcome measured without error, Y denote Z measured with error, and let X denote a pre-treatment covariate. $Z(a)$ and $Y(a)$ will denote potential outcomes under treatment a, such that $Y(a) = Z(a) + \epsilon(a)$, where $\epsilon(a) \sim N(\mu_a, \sigma_a^2)$. Here, we are assuming a simple classical measurement error structure, where the error terms are Normally distributed such that their distributions can differ across treatment groups. To expand upon this notation, let $Y^s(a)$ denote the outcome measured with error under treatment a in dataset s. We can define the potential outcomes measured with error under different treatment conditions in different study samples in the following way:

$$Y^{rct}(0) = Z(0) + \epsilon^{rct}(0) \qquad \epsilon^{rct}(0) \sim N(\mu_0^{rct}, \sigma_0^{rct^2})$$

$$Y^{rct}(1) = Z(1) + \epsilon^{rct}(1) \qquad \epsilon^{rct}(1) \sim N(\mu_1^{rct}, \sigma_1^{rct^2})$$

$$Y^{v}(0) = Z(0) + \epsilon^{v}(0) \qquad \epsilon^{v}(0) \sim N(\mu_0^{v}, \sigma_0^{v2})$$

$$Y^{v}(1) = Z(1) + \epsilon^{v}(1) \qquad \epsilon^{v}(1) \sim N(\mu_1^{v}, \sigma_1^{v2})$$

Note that $Z(a)$ does not differ by study sample, because conceptually, we do not consider someone's underlying true potential outcomes to differ by which study they are in. However, we use the sample superscripts to suggest that a person's potential outcomes *measured with error* can, in fact, differ by study sample, because the measurement error parameters could differ by study sample. Using these definitions, we will now formalize when the average treatment effect in the intervention trial will be biased when estimated using an outcome variable measured with error instead of a variable measured without error.

3.2.1 ATE Bias under Outcome Measurement Error

Suppose the estimand of interest in an intervention trial is the average treatment effect (ATE) of the intervention on the outcome measured without error, defined as $\Delta = E[Z(1) - Z(0)]$. However, since we do not observe $Z(0)$ or $Z(1)$ in the intervention study, we can only attempt to estimate Δ using Y, the observed outcome measured with error, in the intervention study with the following naive estimator:

$$\hat{\Delta} = \frac{\sum_{i=1}^{n_{rct}} Y_i A_i}{\sum_{i=1}^{n_{rct}} A_i} - \frac{\sum_{i=1}^{n_{rct}} Y_i (1 - A_i)}{\sum_{i=1}^{n_{rct}} (1 - A_i)} \tag{3.1}$$

Using Equation 3.1, we can derive the bias of $\hat{\Delta}$ as an estimate for Δ as:

$$\text{bias}_{\hat{\Delta}} = \mu_1^{rct} - \mu_0^{rct} \tag{3.2}$$

In other words, estimating Δ using the mis-measured outcome will be a biased estimate if the means of the outcome error under treatment and control conditions are different. Note that if the measurement error is either classical or differential with respect to treatment, but the errors under treatment and control conditions are centered around the same value, then $\hat{\Delta}$ will still be an unbiased estimate of Δ (its variance may be inflated, though this is not the focus of the current paper). Throughout this paper, we will therefore focus on the case where the measurement error is differential with respect to treatment, such that in the trial, $\mu_0^{rct} \neq \mu_1^{rct}$.

One challenge to correcting for $\text{bias}_{\hat{\Delta}}$ is that the true outcomes measured without error, $Z(0)$ and $Z(1)$, are typically unobserved in an intervention trial. Therefore, the error means μ_0^{rct} or μ_1^{rct} cannot be estimated using the trial data alone. One strategy is to utilize information from an external validation study to estimate μ_0^{rct} or μ_1^{rct} by making an assumption of transportability. However, external validation studies typically only measure the outcomes under a single control condition, as described in Table 3.1 below.

Note that if the potential outcomes under treatment, $Z(1)$ and $Y(1)$, were

Table 3.1: Observables by study sample. Cells shaded in grey denote observed measures, while blank cells denote unobserved quantities.

	Z(0)	Z(1)	Y(0)	Y(1)	X
validation S = v	✓		✓		✓
intervention study S = *rct*			✓	✓	✓

also observed in the validation sample, then the validation study would be considered an intervention trial in itself. Such a scenario is highly unlikely, given the intention and design of external validation studies. To account for this, Siddique et al. (2019) propose using external validation data to estimate the error mean under control conditions, μ_0^{rct}. We will consider the following naive estimator for μ_0^{rct}:

$$\hat{\mu}_0^{naive} = \frac{1}{n_v} \sum_{i=1}^{n_v} (Y_i - Z_i) \tag{3.3}$$

Observe that $\hat{\mu}_0^{naive}$ is an unbiased estimate of μ_0^v, the error mean under control in the *validation sample*. By assuming that transportability holds, we are also assuming that it is an unbiased estimate for μ_0^{rct}, the error mean under control in the *intervention study*.

Lastly, Siddique et al. (2019) conduct sensitivity analyses around the error under treatment (μ_1^{rct}) to obtain a plausible range of estimates for Δ. We build on this work by proposing a solution to when the transportability assumption is violated, and $\hat{\mu}_0^{naive}$ is therefore a biased estimate for the error mean under control in the intervention study, μ_0^{rct}.

The remainder of the paper is structured as follows: in Section 3.3, we describe the transportability assumption evoked to estimate μ_0^{rct} using validation data and discuss its plausibility. We then formalize when the transportability assumption will be violated, and how much bias is introduced as a result. Then, in Section 3.4, we propose the utilization of propensity score-type weighting methods to decrease the bias of estimating μ_0^{rct} using validation data, followed by a simulation illustrating the performance of the weighting methods in Section 3.5. We apply the methods to a data example in Section 3.6, and conclude with a discussion of outcome measurement error and the utilization of validation data, along with some limitations, in Section 3.7.

3.3 Transportability

Suppose $\hat{\Delta}$ is a biased estimate of Δ in an intervention trial, and external validation data is therefore utilized to partially account for $\text{bias}_{\hat{\Delta}}$ by estimating the mean error under control, μ_0^{rct}. In order to obtain an estimate for μ_0^{rct}, a transportability assumption must be made. Formally, the assumption is as follows:

$$f(Y^v(1), Y^v(0)|Z(1), Z(0), X) = f(Y^{rct}(1), Y^{rct}(0)|Z(1), Z(0), X)$$

This transportability assumption implies that, under Normality, $\mu_0^v = \mu_0^{rct}$ and $\sigma_0^{v2} = \sigma_0^{rct^2}$ (and also that $\mu_1^v = \mu_1^{rct}$ and $\sigma_1^{v2} = \sigma_1^{rct^2}$). In other words, the assumption states that the measurement error structures for the potential outcomes in the validation sample are the same as they are in the intervention study. This assumption also implies that $\hat{\mu}_0^{naive}$, which estimates the error mean under control using validation data, is an unbiased estimate of μ_0^{rct}. However, the transportability assumption may not hold in some cases, which would introduce additional bias when estimating the Δ. We will now describe when the transportability assumption will be violated.

3.3.1 Bias from transportability violation (bias$_{\hat{\mu}_0}$)

In order to formalize when there will be bias in estimating μ_0^{rct} using validation data, consider the case where we have a single covariate, X, such that $X = \beta_0 + \beta_1 \mathbb{1}_{S=v} + \epsilon_X$, where ϵ_X has mean 0 and variance σ_X^2. Observe that $\beta_0 = E[X|S = rct]$ and $\beta_1 = E[X|S = v] - E[X|S = rct]$. In other words, β_1 represents the difference in mean of covariate X across the two datasets (trial and validation data).

Next, recalling the classical measurement error structure of $Y = Z + \epsilon$, consider when the error term is distributed as $\epsilon \sim N(\alpha_0 + \alpha_1 A + \alpha_2 X, \sigma_Y^2)$ such that the measurement error is differential with respect to both treatment and X. By performing a substitution for X, we obtain the following:

$$\epsilon \sim N(\alpha_0 + \alpha_1 A + \alpha_2\{\beta_0 + \beta_1 \mathbb{1}_{S=v} + \epsilon_X\}, \sigma_Y^2) \tag{3.4}$$

Recall that $\mu_a^s = E[Y^s(a)] - E[Z(a)]$. The measurement error mean parameters under each treatment condition in each dataset can therefore be expressed as follows:

$$\mu_0^{rct} = \alpha_0 + \alpha_2\beta_0$$

$$\mu_1^{rct} = \alpha_0 + \alpha_2\beta_0 + \alpha_1$$

$$\mu_0^{v} = \alpha_0 + \alpha_2(\beta_0 + \beta_1)$$

$$\mu_1^{v} = \alpha_0 + \alpha_2(\beta_0 + \beta_1) + \alpha_1$$

First, notice that $\text{bias}_{\hat{\Delta}}$, which is the difference in error means between treatment and control conditions in the trial, can be expressed as α_1. Next, we can derive the bias of $\hat{\mu}_0^{naive}$, which uses validation data, as an estimate of μ_0^{rct} as follows:

$$\text{bias}_{\hat{\mu}_0} = \mu_0^{v} - \mu_0^{rct} = \alpha_2\beta_1 \tag{3.5}$$

The transportability assumption will therefore be violated if $\alpha_2 \neq 0$ *and* $\beta_1 \neq 0$. In other words, if a covariate X impacts the measurement error structure ($\alpha_2 \neq 0$), *and* the distribution of X differs across the trial and the validation sample ($\beta_1 \neq 0$), then $\hat{\mu}_0^{naive}$ will be a biased estimate of μ_0^{rct}. This also extends to when there are multiple covariates that meet these two conditions.

The transportability assumption violation, and the introduction of $\text{bias}_{\hat{\mu}_0}$, may also increase $\text{bias}_{\hat{\Delta}}$. Observe that if we substitute the estimate $\hat{\mu}_0^{naive}$ for

μ_0^{rct} in Equation 3.2, we obtain the following:

$$\widetilde{\text{bias}}_{\hat{\Delta}} = \mu_1^{rct} - \hat{\mu}_0^{naive}$$

$$\mathbb{E}[\widetilde{\text{bias}}_{\hat{\Delta}}] = \mu_1^{rct} - \mu_0^{rct} - \alpha_2\beta_1$$

$$= \alpha_1 - \alpha_2\beta_1$$

This motivates the proposal of a weighted estimator for μ_0^{rct}, that reduces $\text{bias}_{\hat{\mu}_0}$, which we present in Section 3.4.

Table 3.2 summarizes the discussion on the two potential biases, $\text{bias}_{\hat{\Delta}}$ and $\text{bias}_{\hat{\mu}_0}$, providing different cases for researchers to consider when these biases should be of concern. While this may not be an exhaustive list of *all* possible scenarios, we think of these as the most plausible scenarios in practice that researchers may encounter. Only Scenario VI, in which the measurement error is differential in the trial with respect to A, differential in the trial and validation sample with respect to X, and the distribution of X differs between the trial and the validation sample, violates the transportability assumption. Scenario V is technically possible, where $\text{bias}_{\hat{\mu}_0} \neq 0$ while $\text{bias}_{\hat{\Delta}} = 0$. However, note that the motivation for outcome measurement error correction is only really when the measurement error is differential by treatment (i.e. $\text{bias}_{\hat{\Delta}} \neq 0$), so Scenario V is therefore highly unlikely.

3.3.2 Additional Assumptions

In addition to the transportability assumption, we make a parametric assumption that the measurement error model form is the same across the two

Table 3.2: Conditions under which to be concerned about bias in the ATE ($\text{bias}_{\hat{\Delta}}$) and/or bias in the measurement error (ME) correction ($\text{bias}_{\hat{\mu}_0}$)

Scenario	ME differs by A	ME differs by X	X differs by sample	$\text{bias}_{\hat{\Delta}}$	$\text{bias}_{\hat{\mu}_0}$
I				0 $\mu_0^{rct} = \mu_1^{rct}$	0
II	✓			α_1 $\mu_0^{rct} \neq \mu_1^{rct}$	0
III	✓	✓		α_1 $\mu_0^{rct} \neq \mu_1^{rct}$	0 $\alpha_2 \neq 0$, but $\beta_1 = 0$
IV	✓		✓	α_1 $\mu_0^{rct} \neq \mu_1^{rct}$	0 $\beta_1 \neq 0$, but $\alpha_2 = 0$
V		✓	✓	0 $\mu_0^{rct} = \mu_1^{rct}$	$\alpha_2\beta_1$ $\beta_1 \neq 0$ and $\alpha_2 \neq 0$
VI	✓	✓	✓	α_1 $\mu_0^{rct} \neq \mu_1^{rct}$	$\alpha_2\beta_1$ $\beta_1 \neq 0$ and $\alpha_2 \neq 0$

samples. In other words, we assume that if the measurement error is differential with respect to a given covariate in the validation sample, then it is also differential with respect to that covariate in the intervention trial (i.e. if age impacts the measurement error structure in the validation sample, it also does so in the trial). This assumption extends to the presence of higher order terms, such as interactions or quadratic terms, that they be present in both samples. We also must assume that there are no unobserved covariates that impact the measurement error and differ between the two samples. Lastly, we must make an assumption of common support, that the range of all covariates in the validation sample are covered by their respective ranges in the intervention trial. For example, we cannot transport an estimate from a validation study where the oldest participant is fifty years old to an intervention trial with participants over the age of fifty. Another way to frame this is that each trial participant has a nonzero probability of participating in the external validation study. The plausibility of these assumptions are discussed further in Section 3.7.

3.4 Weighting-Based Approach to Reduce Transportability Bias

We will now describe the use of propensity score-type weights to reduce the transportability bias. Propensity scores have been traditionally used in non-experimental studies, where treatment is not randomized, to make treatment groups more similar on pre-treatment covariates using matching or weighting methods (Rosenbaum & Rubin, 1983). This approach has since been applied to the fields of transportability and generalizability, where propensity scores are used to model the conditional probability of trial participation (instead of treatment assignment). The probabilities are subsequently used to weight a randomized trial so it better resembles a well-defined target population on observed pre-treatment covariates (Dahabreh, Robertson, Tchetgen, Stuart, & Hernán, 2018; Kern, Stuart, Hill, & Green, 2016; Stuart, Ackerman, & Westreich, 2018).

Previous work has demonstrated similar benefits of implementing propensity score-type weighting methods when using external validation data to adjust for missing confounders (McCandless, Richardson, & Best, 2012) and when evaluating disease prediction models in samples that differ from the target population (Powers, McGuire, Bernstein, Canchola, & Whittemore, 2019). Here, we are interested in addressing the transportability violation by weighting the external validation sample so that it better resembles the intervention study of interest on a set of observed pre-treatment covariates.

In brief, we will do so by modeling the probability of study membership (trial vs. validation study), and then weighting the validation sample before estimating μ_0. Consider the following model of study participation:

$$\text{logit}(P(S = rct|X)) = \theta^t X \tag{3.6}$$

Where X is a set of observed covariates measured in both the trial and the validation data. We can then predict the probability of trial participation as

$$\hat{e}_i = \hat{e}(X) = \text{expit}(\hat{\theta}^t X) \tag{3.7}$$

Similar to ATT weighting for non-experimental studies, we then construct the following weights:

$$\hat{w}_i = \begin{cases} \frac{\hat{e}_i}{1-\hat{e}_i} & \text{if } S = v \\ 0 & \text{if } S = rct \end{cases} \tag{3.8}$$

Using these weights, we can then estimate μ_0^{rct} using validation data with the following estimator:

$$\hat{\mu}_0^{weighted} = \frac{\sum_{i=1}^{n_v} \hat{w}_i (Y_i - Z_i)}{\sum_{i=1}^{n_v} \hat{w}_i} \tag{3.9}$$

Individuals in the validation sample that are more similar to the trial participants will have greater predicted probabilities of being trial participants, and will therefore have larger weights. Members of the external validation sample that are most dissimilar to the trial sample will be down-weighted towards zero. In this way, we can obtain a weighted estimate of the error mean under control in the validation sample, μ_0^v, such that we reduce the bias of this estimate as an estimate for μ_0^{rct} due to covariate differences across samples. Details on estimating the standard error for inference can be found in the Supplemental Materials.

3.4.1 Weighting under Misspecification of the Sample Membership Model

Recall that the error estimate in the validation sample will be a biased estimate of the error in the trial if there exists a set of covariates that impact the measurement error structure and that also differ by sample. We therefore want to include all observed covariates that fall into this category when fitting the model of sample membership. If we fit the *correct* model of sample membership accounting for all such Xs in the true form, then we should be able to eliminate the bias of our μ_0 estimate through this weighting procedure (as is the case when transporting trial results to a target population using inverse odds weighting, see proof in Westreich, Edwards, Lesko, Stuart, and Cole (2017)). In practice, however, it can be quite challenging to fit the correct sample membership model (i.e. there may be complex interactions or higher order terms in the true model that are omitted). Fitting a simpler, misspecified sample membership model may lead to a smaller reduction of the bias when weighting. In the next section, we describe a simulation study, where we demonstrate the performance of the proposed weighting methods under increasingly complex true sample membership models, and varying amounts of model misspecification.

3.5 Simulation

We now conduct a simulation study to assess the weighting methods described in Section 3.4 on decreasing $\text{bias}_{\hat{\mu}_0}$. We consider four covariates, and vary the following: (1) the degree to which each X impacts the measurement error model, (2) the degree to which each X impacts the trial membership model, and (3) the degree of misspecification of the trial membership model that is fit using the validation sample.

3.5.1 Simulation Setup

Consider the following measurement error model of Y:

$$Y = \alpha_0 + \alpha_1 Z + \alpha_2 A + \alpha_3 X_1 + \alpha_4 X_2 + \alpha_5 X_3 + \alpha_6 X_4 + \epsilon_Y$$

where $\epsilon_Y \sim N(0, \sigma_Y^2)$. Since we only require that the Xs impact the error structure in *some* capacity, we do not consider other, more complex true structures of the measurement error model in this simulation study.

We vary the true underlying models of sample membership by considering the following seven model forms:

1. $S \sim X_1 + X_2 + X_3 + X_4$
2. $S \sim X_1 + X_2 + X_3 + X_4 + X_3^2$
3. $S \sim X_1 + X_2 + X_3 + X_4 + X_4^2$
4. $S \sim X_1 + X_2 + X_3 + X_4 + X_3 X_4$
5. $S \sim X_1 + X_2 + X_3 + X_4 + X_3 X_4 + X_3^2 + X_4^2$
6. $S \sim X_1 + X_2 + X_3 + X_4 + X_1 X_4$
7. $S \sim X_1 + X_2 + X_3 + X_4 + X_1 X_3$

We parameterize the coefficients for the covariates X_1, X_2, X_3, and X_4 as $\{\gamma_1, 0, \frac{1}{2}\gamma_1, 2\gamma_1\}$ in the measurement model, and as $\{0, \gamma_2, 2\gamma_2, \frac{1}{2}\gamma_2\}$ in the trial membership model (See Supplemental Table B.1). In doing so, we establish that covariate X_1 impacts sample membership but not measurement error, X_2 impacts measurement error but not sample membership, X_3 weakly impacts sample membership and strongly impacts measurement error, and X_4 strongly impacts sample membership and weakly impacts measurement error. We set any quadratic term coefficients to 50% of the original X's coefficient (i.e. if X_3 has a coefficient of $\frac{1}{2}\gamma_1$, then X_3^2 would have a coefficient of $\frac{1}{4}\gamma_1$). For interaction terms, the coefficient is set to the average of the two Xs' original coefficients (i.e. the coefficient for a X_1X_3 interaction term would be $\frac{3}{4}\gamma_1$).

The two parameters γ_1 and γ_2 function as scaling parameters, varying from 0 to 1 by increments of 0.2. Observe that when $\gamma_1 = 0$, the four covariates do not impact sample membership at all, and the trial membership probabilities are expected to be 0.5 in both study samples. As γ_1 increases to 1, the impact of the variables on sample membership increases, and the overlap of the probabilities across the two samples decreases. In this way, γ_1 can be considered a function of the absolute standardized mean difference (ASMD) of the participation probabilities between the trial and the validation sample. When $\gamma_2 = 0$, then the measurement error is not differential with respect to any of the covariates, and as γ_2 increases, so does the impact of the covariates on the measurement error structure. Note that when either $\gamma_1 = 0$ or $\gamma_2 = 0$, then we expect that $\text{bias}_{\hat{\mu}_0} = 0$.

For each of the seven true trial membership models, we fit both the true model, which is correctly specified, as well as a main-effects-only model. The main-effects-only model will be misspecified when the true model has interaction and/or quadratic terms. This type of misspecification illustrates a plausible scenario, in which researchers may fit a simple multi-variable logistic regression model to estimate the trial participation probabilities, ignoring potential complexities in the underlying true model form.

In order to quantify the degree of model misspecification (DoM), Lenis, Ackerman, and Stuart (2018) propose a unit-independent, informative metric:

$$\eta_S = \frac{1}{N}\sum_{i=1}^{N}\frac{|\hat{\pi}_i - \hat{\pi_i^C}|}{\sigma_{\hat{\pi}^C}}$$

where $\hat{\pi}_i$ is the predicted probability of being in the trial under the specified model, and $\hat{\pi}_i^C$ is the predicted probability of being in the trial under the true selection model. We use the DoM metric to relate the amount of model misspecification across scenarios.

In total, there are 756 simulation scenarios that vary by: (6 values for γ_1) × (6 values for γ_2) × (7 true trial membership models) × (3 weighting options: unweighted naive estimator, weighted estimator with correctly specified weights, and weighted estimator with misspecified weights). We iterate over each scenario 1000 times, and will now describe the data generation process.

3.5.2 Data Generation

Consider one particular scenario from the 756 scenarios outlined above. We start by simulating a population of four X covariates ($N = 1000000$) according to a multivariate Normal distribution with mean 0, variance 1, where there is no correlation between the Xs. Based on the scenario's γ_1 value and true trial membership model form - suppose the simplest true form for example - we generate the probability of being in the trial (vs. the validation sample) for the whole population as follows:

$$p_i = \text{expit}(\gamma_1 X_{1i} + 0X_{2i} + \frac{1}{2}\gamma_1 X_{3i} + 2\gamma_1 X_{4i})$$

for $i = 1, ..., N$. Next, we generate the binary sample membership variable S for each member of the population as $S_i \sim \text{Bernoulli}(p_i)$ for $i = 1, ..., N$. Note that while each p_i is determined by the scenario's specified parameters, S is assigned with a degree of randomness, such that each person in the population theoretically has a chance of being “in the trial" or “in the validation sample" across each different simulation run.

After assigning S, we randomly sample members for the trial and validation samples, each of size $n = 1000$. In this step, we observe the differences in the covariates across the two samples as specified by the γ_1 scaling parameter. We then generate the true potential outcomes $Z(0)$ and $Z(1)$ as $Z(0) \sim N(0, 1)$ and $Z(1) \sim N(2, 1)$, and the mis-measured potential outcomes $Y(0)$ and $Y(1)$ as:

$$Y(a) \sim N\big(Z(a) + 0X_1 + \gamma_2 X_2 + 2\gamma_2 X_3 + \frac{1}{2}\gamma_2 X_4, 1.5\big)$$

such that the variance of $Y(a)$ is 1.5 times the variance of $Z(a)$.

We assign treatment A as $\text{Bernoulli}(0.5)$ in the trial and 0 in the validation sample. Lastly, we generate the observed outcomes Z and Y as $Z = A \times Z(1) + (1 - A) \times Z(0)$ and $Y = A \times Y(1) + (1 - A) \times Y(0)$.

3.5.3 Simulation Results

Figure 3.1 shows the absolute $\text{bias}_{\hat{\mu}_0}$, or the transportability bias, across simulation scenarios. Each column represents a different underlying true sample membership model structure, increasing in complexity and degree of misspecification from left to right (see Supplemental Figure B.1 for DoM by sample membership model form). The strength of the impact of the Xs on the measurement error (γ_2) increases by row, from top to bottom. The x axis depicts the absolute standardized mean difference (ASMD) of the true selection probabilities across the samples, representing an increasing difference between the X distributions across the samples (γ_1) (see Supplemental Figure B.2). Note that scenarios where the ASMD is greater than 1 represent fairly extreme, less realistic settings. The three lines represent the absolute bias of the naive estimator, the weighted estimator with misspecified weights, and the weighted estimator with correctly specified weights.

First, note that in the top row, the absolute bias is zero under all cases, because the measurement error is not differential with respect to any of the covariates (Table 3.2, Scenarios II and IV). Next, note that in all of the plots, the absolute bias is zero when the ASMD is zero, or when the distribution of covariates in the trial and validation sample do not differ (Table 3.2, Scenarios II and III). Under the same true sample membership model (for a given column), as the impact of the Xs on the measurement error model increases (from top to bottom row), and as the distributional difference of the Xs increases between the samples (from left to right of the x axis), the absolute bias$_{\hat{\mu}_0}$ also increases.

When the selection model is correctly specified, and the resulting predicted probabilities are used to construct the weights, the weighted estimator is nearly unbiased in all scenarios (except for when the impacts of the Xs on the measurement error model and on sample membership are extremely, somewhat unrealistically, large). In practice, though, model misspecification

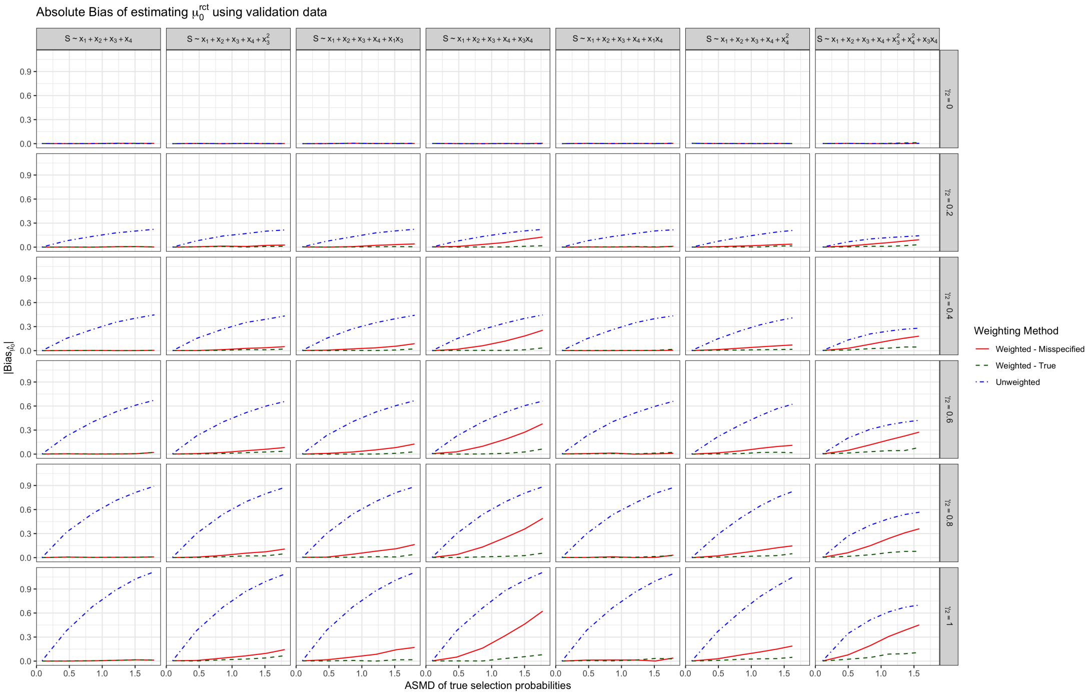


Figure 3.1: Bias of estimating the error mean under control using validation data. Each column represents a different true sample membership model. From top to bottom row, the γ_2 “scale" parameter for the impact of the Xs on the measurement error increases, meaning the strength of the relationship between Y and the Xs is increasing in magnitude. The different line types and colors represent the different weighting approaches: Unweighted (blue dotted dash), weighted by fitting the simplest additive model (“Weighted - Misspecified", red solid), and weighted by fitting the true selection model (“Weighted - True", green dash). This figure appears in color in the electronic version of this article.

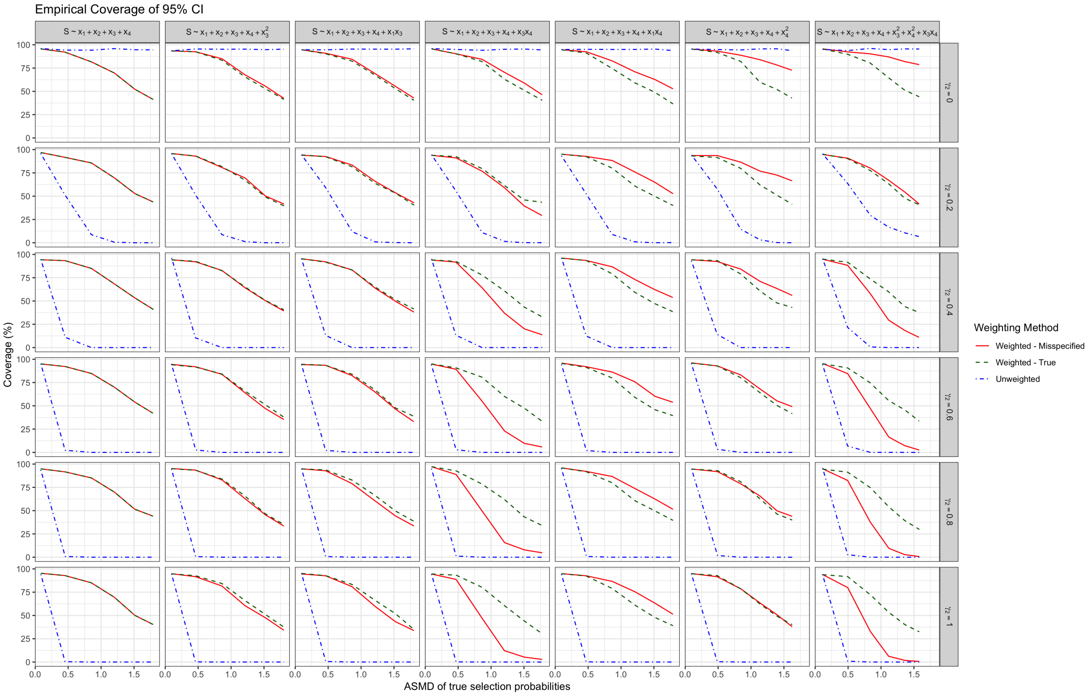


Figure 3.2: Empirical coverage of 95% CI for estimators of μ_0^{rct}. Each column represents a different true sample membership model. From top to bottom row, the γ_2 "scale" parameter for the impact of the Xs on the measurement error increases, meaning the strength of the relationship between Y and the Xs is increasing in magnitude. The different line types and colors represent the different weighting approaches: Unweighted (blue dotted dash), weighted by fitting the simplest additive model ("Weighted - Misspecified", red solid), and weighted by fitting the true selection model ("Weighted - True", green dash). This figure appears in color in the electronic version of this article.

is very plausible, as it may be common for researchers to fit a main-effects-only model of the covariates to predict the sample membership probabilities, unable to identify the true model form. Under varying amounts of model misspecification, we see that the weighted estimator still performs fairly well in reducing bias$_{\hat{\mu}_0}$. As the severity of the transportability assumption violation increases, the weighted estimator with misspecified weights does appear to perform worse than the weighted estimator with correctly specified weights under certain scenarios, particularly when the omitted interaction and/or quadratic terms are for covariates that more strongly impact sample membership and measurement error. For example, in column 4, the main-effects-only model omits an interaction between X_3 and X_4, the two variables that impact *both* models. That weighted estimator performs worse than the main-effects-only model that omits an X_1X_3 interaction (column 3), and the model that omits an X_1X_4 interaction (column 4), as X_1 does not impact the measurement error structure at all.

Figure 3.2 shows the empirical 95% confidence interval coverage of the different estimators for μ_0^{rct}. First, observe that across all scenarios in which the covariates impact the measurement error (rows 2-6), the coverage of the naive estimator sharply decreases towards zero as the intervention trial differs more greatly from the validation sample on pre-treatment covariates. Note also, though, that the weighting approaches (even with a misspecified model) generally yield substantially better confidence interval coverage than does the naive approach. This aligns with the pattern in Figure 3.1, in which the naive estimator becomes increasingly biased as the samples become less similar on the covariates. Next, note that the weighted estimator with correctly specified weights tends to have better coverage than the weighted estimator with misspecified weights, though the coverage also decreases in the more extreme cases of covariate differences between the samples. Still, the rate of coverage decline for the weighted estimators by covariate differences is far smaller than that of the naive estimator. Also, note that for the more plausible scenarios (ASMD < 1), the coverage of the weighted estimators is still fairly good, and far better than that of the naive estimator.

In the cases where the measurement error is *not* differential with respect to the Xs (top row), the weighted estimators seem to have worse coverage than the naive estimator, even though they are all unbiased. This is due to the presence of large/extreme weights, which shift the point estimates of μ_0^{rct} further away from the truth and therefore increase the variability of the weighted estimators, even though they are still unbiased. Trimming extreme weights (especially in the less plausible scenarios where ASMD > 1), resulted in better coverage across these settings. However, in cases where the measurement error is not differential with respect to covariates, and the transportability assumption is not believed to be violated, then the weighting approach may not be preferable.

Overall, though, it appears that *any* weighting, whether by fitting a main-effects-only model, the true selection model, or anything in between, greatly improves the transportability of the control group error mean from the validation sample to the trial. When there are concerns about measurement error corrections not transporting properly from an external validation sample to an intervention trial, these simulation results highlight that the weighting method proposed in Section 3.4 can help reduce the bias and improve coverage due to poor transportability.

3.6 Data Example

We now apply the methods described in Section 3.4 to a lifestyle intervention trial, PREMIER, using OPEN, an external validation sample.

In PREMIER, 810 individuals were randomized to either one of two behavioral/dietary recommendations, or to standard care, to estimate the effect of the intervention on blood pressure reduction (Svetkey et al., 2003). For illustrative purposes, instead of blood pressure, we focus on self-reported sodium intake, measured by 24-hour recall, as the outcome of interest. We also combine the two intervention groups into one "treatment" group. PREMIER is a rather unique intervention trial, in that urinary sodium intake was *also* collected at each time point in addition to the self-reported intake, providing an opportunity for us to evaluate the method performance. Note that this is atypical for an intervention trial to collect. We use sodium intake at 18 months followup as the outcome of interest (though note that in the original trial, the primary time point of interest for analysis was 6 months), and we limit the sample only to those who have both self-reported and urinary sodium intake measures at 18 months (n = 670).

OPEN is an external validation study that measures both self-reported sodium (via 24-hour recall) and urinary sodium (via a 24-hour urine sample) on a sample of 484 study participants, with the goal of understanding the

Table 3.3: Distribution of Covariates by Study

	OPEN (n=484)	PREMIER (n=810)	ASMD
Male	0.54	0.38	0.33
Age			
$\leq$ 40	0.03	0.14	0.42
41-45	0.17	0.17	0.01
46-50	0.21	0.24	0.06
51-55	0.20	0.21	0.03
56-60	0.15	0.13	0.05
$\geq$ 61	0.24	0.12	0.30
BMI	27.87	33.06	0.95
Black	0.06	0.34	0.81
Education			
College	0.55	0.59	0.08
Grad School	0.32	0.32	0.00

structure and amount of measurement error among self-reported dietary outcomes (Subar et al., 2003). Using PREMIER and OPEN, we will demonstrate that the measurement error of dietary sodium intake is differential with respect to pre-treatment covariates, and that the distribution of these factors also differ between the intervention trial and the validation study. Therefore, the transportability assumption is violated, and $\hat{\mu}_0^{naive}$, which is estimated in OPEN, is a biased estimate for μ_0^{rct} in PREMIER.

Table 4.1 describes the distribution of covariates across the two studies, along with the ASMD of each covariate between the two studies. Observe that BMI differs greatly between the two studies. Additionally, the OPEN population appears to be older, more male and less racially diverse than PREMIER.

In order to implement the methods described in Section 3.4, we form a “stacked" dataset, comprised of data from PREMIER and OPEN, which contains variables for sample membership (S), treatment (A), self-reported dietary sodium intake (Y), urinary sodium intake (Z), and the following common covariates (X): age category, sex, race, BMI and education. Certain covariates, like age and education, are categorized to ensure consistency in measures across datasets, and race is utilized as a dichotomous variable, indicating if individuals identify as Black or not. Again, note that typically Z would be coded as missing when $S = rct$; however, the unique nature of PREMIER allows us to observe Z in the trial.

By comparing the difference in outcome means by measurement type in PREMIER, it appears that the self-reported dietary sodium measures under-report the true sodium intake in both treatment arms, and $\text{bias}_{\hat{\Delta}}$ at 18 months is estimated to be 0.028 (see Supplemental Table B.2). While this difference is not significant (see Supplemental Table B.3), we still proceed to assess the feasibility of transporting the error mean under control from OPEN to PREMIER for illustrative purposes.

We fit a linear model to determine which covariates significantly impact the measurement error under control conditions, using data from both PREMIER and OPEN (see Supplemental Table B.4). The error under control appears to be differential with respect to sex and race, and also weakly differential with respect to education. Given the output of this model, and the covariate distributions shown in Table 4.1, we therefore have reason to believe that the transportability assumption is violated.

Next, we fit a sample membership model using all five covariates, which includes covariates that impact both the measurement error and sample membership, as well as covariates that just differ between the two samples. To fit the model, we use generalized boosted models (GBM), an algorithm that allows for flexible, nonlinear relationships when modeling study membership by a large number of covariates (McCaffrey, Ridgeway, & Morral, 2004). We examine the distributions of predicted sample membership probabilities (see Supplemental Figure B.3), which have an $ASMD = 1.47$. This is unsurprising, given the large differences between the two samples' distributions of race and BMI. There are some outliers in the resulting validation sample weights that are more than ten times the average of the weights. We therefore implement weight trimming to account for the extreme weights, setting all validation sample weights in the top decile to the 90th percentile weight value (Lee, Lessler, & Stuart, 2011) (see Supplemental Figure B.4 for the distribution of the trimmed weights in the validation sample).

Lastly, we use the weights to estimate $\hat{\mu}_0^{weighted}$. Table 3.4 shows both the unweighted and weighted estimates, along with the estimate of μ_0^{rct} in PREMIER (which again, is usually not estimable when only self-reported outcomes are collected). Observe that for the outcome at 18 months, the absolute bias of the $\hat{\mu}_0$ estimate, $\text{bias}_{\hat{\mu}_0}$, decreases by about 80% after implementing the weighting method.

In the data example, note that the ASMD of the sample membership probabilities between OPEN and PREMIER is quite large (1.47), and that some of the covariates are extremely different from one another. Even after weighting, OPEN is still a bit dissimilar from PREMIER by BMI and race

Table 3.4: Estimated Error Mean under Control Conditions in the Validation Sample, and Associated Bias, by Weighting Method

	6 months			18 months		
Method	$\hat{\mu}_0$	μ_0^{rct}	$\text{bias}_{\hat{\mu}_0}$	$\hat{\mu}_0$	μ_0^{rct}	$\text{bias}_{\hat{\mu}_0}$
Unweighted	-0.228	-0.227	-0.001	-0.228	-0.305	0.077
Weighted	-0.326		-0.099	-0.321		-0.016

(see Supplemental Figure B.5). Additionally, by fitting the selection model using GBM, we are able fitting a model somewhere between the true form and the main-effects-only form. We therefore see that these results reflect the simulation findings under rather extreme cases, suggesting that the weighting may help to a certain extent, but that the differences between the validation sample and trial may lend to sub-optimal performance.

3.7 Discussion

When using self-reported measures as outcomes in a lifestyle intervention study, it is important to correct for any potential measurement error in order to make accurate inferences on the effect of the treatment in the study population. While measurement error is a well documented issue, particularly in nutritional epidemiology, there is still much need for increased method implementation in applied research studies, as well as improved methodology for different types of error (Brakenhoff et al., 2018; Jurek, Maldonado, Greenland, & Church, 2006). We highlight the importance of considering transportability when utilizing external validation studies to correct for outcome measurement error. Using externally estimated measurement error may introduce additional biases to the ATE estimate in cases where validation samples are dissimilar from the primary intervention study of interest. We show that weighting the validation sample to better resemble the intervention study can reduce such biases, and improve upon the transportability of the measurement error estimated in the external sample. However, in some extreme cases, it may still be inappropriate to transport if the validation sample is vastly different from the trial on a set of observed characteristics. Additionally, it is important to remember that while researchers are often concerned about measurement error, it will only lead to a biased ATE estimate when the outcome error is differential with respect to treatment and the error means are thereby different across treatment groups. Such bias would prompt the usage of external validation data for outcome measurement error correction, and thus the concerns about transportability (see Table 3.2).

There are several limitations to the work presented in this paper. First, we assume that the measurement error model structure (i.e. the model relating the measurement error to the covariates) in both the intervention study and the validation sample are the same. It is possible that such relationships may differ between studies, even though in practice, this would be untestable without observing the outcome without measurement error in the trial itself. Further research is needed to understand transportability and to apply the methods proposed in this paper when relaxing this assumption. Second, we assume in this work that we are able to fully observe all covariates that impact the outcome measurement error structure in both the intervention study and the validation sample. Due to data availability, certain important variables may be unobserved in practice, either in one of the datasets, or in both, which may hinder the performance of these methods. Sensitivity analyses should be adapted and applied to address these concerns (Nguyen, Ebnesajjad, Cole, & Stuart, 2017). Lastly, as seen with PREMIER, transportability may vary by time-point with longitudinal outcomes, warranting further investigation into how transportability and measurement error may vary over time.

This work has focused on the use of external validation samples only, and further research is needed to evaluate the differences in transportability between internal and external validation samples. In some cases, internal validation samples may still be preferable when possible to incorporate into study design, particularly such that true outcome measures can be obtained under different treatment conditions. When it is infeasible to collect internal validation data, researchers designing external validation studies should still consider the possible relevant trial study populations to which the validation sample will be applied to. Taking such steps when designing validation studies will also help ensure better transportability when using information from external data sources to correct for outcome measurement error.

References

Bound, J., Brown, C., & Mathiowetz, N. (2001). Measurement error in survey data. In *Handbook of econometrics* (Vol. 5, pp. 3705–3843). Elsevier.

Brakenhoff, T. B., Mitroiu, M., Keogh, R. H., Moons, K. G., Groenwold, R. H., & van Smeden, M. (2018). Measurement error is often neglected in medical literature: A systematic review. *Journal of clinical epidemiology*, *98*, 89–97.

Buonaccorsi, J. P. (2010). *Measurement error: Models, methods, and applications*. Chapman and Hall/CRC.

Carroll, R. J., Ruppert, D., Stefanski, L. A., & Crainiceanu, C. M. (2006). *Measurement error in nonlinear models: A modern perspective*. Chapman and Hall/CRC.

Courtemanche, C., Pinkston, J. C., & Stewart, J. (2015). Adjusting body mass for measurement error with invalid validation data. *Economics & Human Biology*, *19*, 275–293.

Dahabreh, I. J., Robertson, S. E., Tchetgen, E. J. T., Stuart, E. A., & Hernán, M. A. (2018). Generalizing causal inferences from individuals in randomized trials to all trial-eligible individuals. *Biometrics*.

Espeland, M. A., Kumanyika, S., Wilson, A. C., Wilcox, S., Chao, D., Bahnson, J., … GROUP, T. C. R., et al. (2001). Lifestyle interventions influence relative errors in self-reported diet intake of sodium and potassium. *Annals of epidemiology*, *11*(2), 85–93.

Forster, J. L., Jeffery, R. W., VanNatta, M, & Pirie, P. (1990). Hypertension prevention trial: do 24-h food records capture usual eating behavior in a dietary change study? *The American Journal of Clinical Nutrition*, *51*(2), 253–257. doi:10.1093/ajcn/51.2.253. eprint: http://oup.prod.sis.lan/ajcn/article-pdf/51/2/253/24115494/253.pdf

Jenab, M., Slimani, N., Bictash, M., Ferrari, P., & Bingham, S. A. (2009). Biomarkers in nutritional epidemiology: Applications, needs and new horizons. *Human genetics*, *125*(5-6), 507–525.

Jurek, A. M., Maldonado, G., Greenland, S., & Church, T. R. (2006). Exposure-measurement error is frequently ignored when interpreting epidemiologic study results. *European journal of epidemiology*, *21*(12), 871–876.

Keogh, R. H., & White, I. R. (2014). A toolkit for measurement error correction, with a focus on nutritional epidemiology. *Statistics in medicine*, *33*(12), 2137–2155.

Keogh, R. H., Carroll, R. J., Tooze, J. A., Kirkpatrick, S. I., & Freedman, L. S. (2016). Statistical issues related to dietary intake as the response variable in intervention trials. *Statistics in medicine*, *35*(25), 4493–4508.

Kern, H. L., Stuart, E. A., Hill, J., & Green, D. P. (2016). Assessing methods for generalizing experimental impact estimates to target populations. *Journal of research on educational effectiveness*, *9*(1), 103–127.

Lee, B. K., Lessler, J., & Stuart, E. A. (2011). Weight trimming and propensity score weighting. *PloS one*, *6*(3), e18174.

Lenis, D., Ackerman, B., & Stuart, E. A. (2018). Measuring model misspecification: Application to propensity score methods with complex survey data. *Computational statistics & data analysis*, *128*, 48–57.

Lyles, R. H., Zhang, F., & Drews-Botsch, C. (2007). Combining internal and external validation data to correct for exposure misclassification: A case study. *Epidemiology*, *18*(3), 321–328.

McCaffrey, D. F., Ridgeway, G., & Morral, A. R. (2004). Propensity score estimation with boosted regression for evaluating causal effects in observational studies. *Psychological methods*, *9*(4), 403.

McCandless, L. C., Richardson, S., & Best, N. (2012). Adjustment for missing confounders using external validation data and propensity scores. *Journal of the American Statistical Association*, *107*(497), 40–51.

Natarajan, L., Pu, M., Fan, J., Levine, R. A., Patterson, R. E., Thomson, C. A., ... Pierce, J. P. (2010). Measurement error of dietary self-report in intervention trials. *American journal of epidemiology*, *172*(7), 819–827.

Nguyen, T. Q., Ebnesajjad, C., Cole, S. R., Stuart, E. A., et al. (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.

Powers, S., McGuire, V., Bernstein, L., Canchola, A. J., & Whittemore, A. S. (2019). Evaluating disease prediction models using a cohort whose covariate distribution differs from that of the target population. *Statistical methods in medical research*, *28*(1), 309–320.

Price, G. R. (1972). Extension of covariance selection mathematics. *Annals of human genetics*, *35*(4), 485–490.

Rosenbaum, P. R., & Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika*, *70*(1), 41–55.

Rothman, K. J., Greenland, S., Lash, T. L., et al. (2008). *Modern epidemiology*. Wolters Kluwer Health/Lippincott Williams & Wilkins Philadelphia.

Siddique, J., Daniels, M. J., Carroll, R. J., Raghunathan, T. E., Stuart, E. A., & Freedman, L. S. (2019). Measurement error correction and sensitivity analysis in longitudinal dietary intervention studies using an external validation study. *Biometrics*.

Spring, B., Schneider, K., McFadden, H. G., Vaughn, J., Kozak, A. T., Smith, M., . . . Hedeker, D., et al. (2012). Multiple behavior changes in diet and activity: A randomized controlled trial using mobile technology. *Archives of internal medicine*, *172*(10), 789–796.

Spring, B., Pellegrini, C., McFadden, H., Pfammatter, A. F., Stump, T. K., Siddique, J., . . . Hedeker, D. (2018). Multicomponent mhealth intervention for large, sustained change in multiple diet and activity risk behaviors: The make better choices 2 randomized controlled trial. *Journal of medical Internet research*, *20*(6), e10528.

Stuart, E. A., Ackerman, B., & Westreich, D. (2018). Generalizability of randomized trial results to target populations: Design and analysis possibilities. *Research on social work practice*, *28*(5), 532–537.

Subar, A. F., Kipnis, V., Troiano, R. P., Midthune, D., Schoeller, D. A., Bingham, S., . . . Ballard-Barbash, R., et al. (2003). Using intake biomarkers to evaluate the extent of dietary misreporting in a large sample of adults: The open study. *American journal of epidemiology*, *158*(1), 1–13.

Svetkey, L. P., Harsha, D. W., Vollmer, W. M., Stevens, V. J., Obarzanek, E., Elmer, P. J., … Aickin, M., et al. (2003). Premier: A clinical trial of comprehensive lifestyle modification for blood pressure control: Rationale, design and baseline characteristics. *Annals of epidemiology*, *13*(6), 462–471.

Westreich, D., Edwards, J. K., Lesko, C. R., Stuart, E., & Cole, S. R. (2017). Transportability of trial results using inverse odds of sampling weights. *American journal of epidemiology*, *186*(8), 1010–1014.

Willett, W. (2012). *Nutritional epidemiology*. Oxford university press.

Wong, M., Day, N., Bashir, S., & Duffy, S. (1999a). Measurement error in epidemiology: The design of validation studies i: Univariate situation. *Statistics in medicine*, *18*(21), 2815–2829.

Wong, M., Day, N., & Wareham, N. (1999b). Measurement error in epidemiology: The design of validation studies ii: Bivariate situation. *Statistics in medicine*, *18*(21), 2831–2845.

Chapter 4

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

Ackerman et al. (2019)

4.1 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 (Weisberg, Hayden, & Pontes, 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, King,

& Stuart, 2008; Shadish, Cook, & Campbell, 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, Crum, Stuart, Ebnesajjad, and Mojtabai (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, Huibers, Renner, & Arntz, 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, Crum, Ebnesajjad, Stuart, and Mojtabai (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, Ackerman, & Westreich, 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, Stuart, Hill, & Green, 2016), these methods are meant to be implemented after the study is already conducted. In this paper, we provide an introductory overview of several post-trial 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, Cole, Bradshaw, & Leaf, 2011). The paper proceeds as follows: Section 4.2 describes the notation and assumptions. Section 4.3 describes methods for assessing and improving upon the generalizability of RCT findings. Section 4.4 provides guidance for preparing data and implementing the described methods using our R package, “generalize." We illustrate the use of “generalize" in Section 4.5 using data from an RCT related to methamphetamine dependence and a nationally-representative survey of SUD treatment admissions. Finally, Section 4.6 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.

4.2 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

$$SATE = E[Y(1) - Y(0)|S = 1]$$

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

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

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 4.3. 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:

1. All members of the target population have a nonzero probability of being selected for the trial.

2. There are no unmeasured variables associated with sample selection and treatment effect given the observed variables.

3. 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.

4. In the trial, treatment assignment is independent of sample selection, as well as of potential outcomes, given the pre-treatment covariates.

Assumption 1 is similar to the positivity assumption for drawing causal inference in non-experimental studies. Assumption 2 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. Assumption 3 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. Assumption 4 is satisfied by nature of the randomization in RCTs.

4.3 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.

4.3.1 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 4.3.2), and also to construct weights to estimate the TATE (Section 4.3.3.1). 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 pre-treatment 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, Lessler, & Stuart, 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.

4.3.2 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:

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

where $f_s(s)$ and $f_p(s)$ are the distributions of trial participation probabilities in the trial sample and target population given a set of common covariates, respectively. Estimation of 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.

4.3.3 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.

4.3.3.1 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 being a trial participant, and can be calculated using the methods described in Section 4.3.1. The TATE is then estimated by fitting a weighted least squares regression model using the trial data (Kern et al., 2016).

4.3.3.2 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 (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).

4.3.3.3 Combining weighting and outcome modeling: TMLE

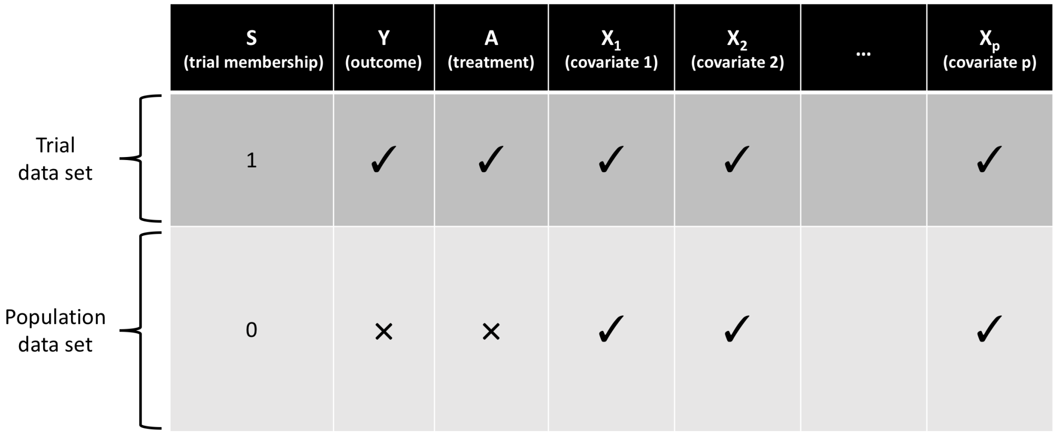
Lastly, Targeted Maximum Likelihood Estimation (TMLE) is a method that combines both strategies. It models both the outcome and the trial participation using pre-treatment covariates, and is robust to whether or not one of those models is incorrect (Gruber & Van Der Laan, 2009; Rudolph, Díaz, Rosenblum, & Stuart, 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.

4.4 Preparing Data for Method Implementation

In order to implement the methods described in Section 4.3, several data preprocessing 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 4.1).

Once the data are formatted in this manner, the methods described in Section 4.3 can be implemented using “generalize," a package developed for



	S (trial membership)	Y (outcome)	A (treatment)	X_1 (covariate 1)	X_2 (covariate 2)	...	X_p (covariate p)
Trial data set	1	✓	✓	✓	✓		✓
Population data set	0	×	×	✓	✓		✓

Figure 4.1: Format of "stacked" data set for implementing generalizability methods

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 C for code).

4.5 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.

4.5.1 Results

Table 4.1 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 3).

Table 4.1: Distribution of Covariates in the Trial vs. Population and their Absolute Standardized Mean Difference (ASMD)

	CSP-1025 (trial)	TEDS-A-2014 (population)	ASMD (unweighted)	ASMD (weighted)
Age				
12-14	0.00	0.01	1.11	1.11
15-17	0.00	0.02	4.12	4.12
18-20	0.02	0.04	0.15	0.07
21-24	0.04	0.12	0.45	0.07
25-29	0.14	0.20	0.19	0.11
30-34	0.12	0.21	0.26	0.11
35-39	0.24	0.15	0.22	0.16
40-44	0.18	0.10	0.20	0.22
45-49	0.17	0.08	0.24	0.21
50-54	0.07	0.05	0.09	0.20
> 55	0.01	0.03	0.09	0.40
Sex				
Male	0.64	0.54	0.20	0.01

Table 4.1: Distribution of Covariates in the Trial vs. Population and their Absolute Standardized Mean Difference (ASMD) *(continued)*

	CSP-1025 (trial)	TEDS-A-2014 (population)	ASMD (unweighted)	ASMD (weighted)
Race				
Black	0.02	0.04	0.14	0.14
White	0.83	0.74	0.25	0.09
Native Hawaiian	0.03	0.01	0.10	0.16
Other	0.10	0.18	0.27	0.08
Ethnicity				
Not Hispanic/Latino	0.86	0.78	0.23	1.07
Unknown/Not Given	0.04	0.01	0.16	0.25
Marital Status				
Married/Partnered	0.23	0.09	0.32	0.08
Education				
9-11 years	0.10	0.29	0.63	0.14
12 years	0.40	0.45	0.10	0.31
13-15 years	0.33	0.17	0.34	0.16
> 15 years	0.15	0.03	0.33	0.05
Employment				
Not in labor force	0.07	0.38	1.25	0.22
Part-time	0.25	0.07	0.41	0.17
Unemployed	0.24	0.45	0.48	0.10
Methamphetamine Use in Past Week				
Yes	0.91	0.42	1.72	0.40

The distributions of the log(trial participation probabilities) in the trial and target population varied somewhat by method of calculation as well (Figure 4.2). 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.

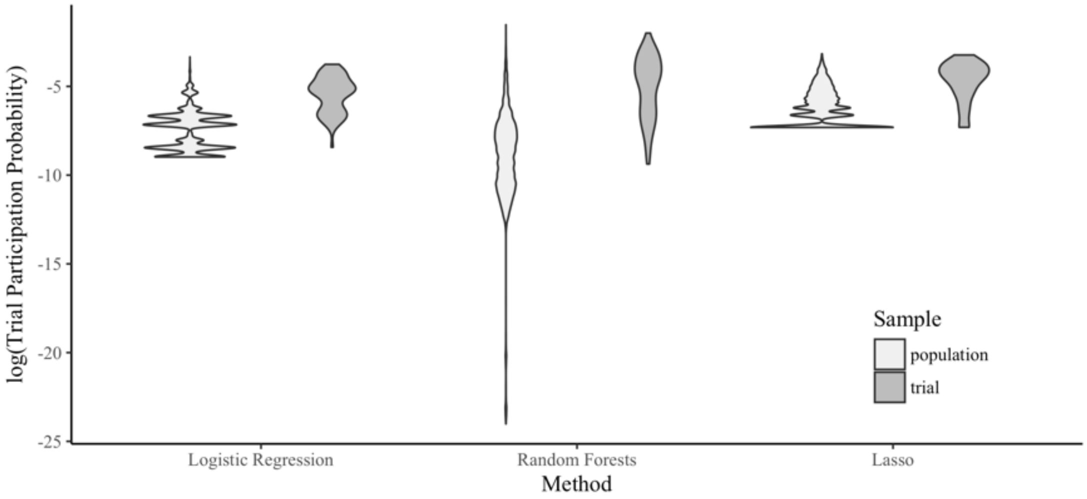


Figure 4.2: log(Trial Participation Probabilities) by Method and Sample Membership

The TATE estimates are shown in Figure 4.3. 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 4.1).

4.6 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

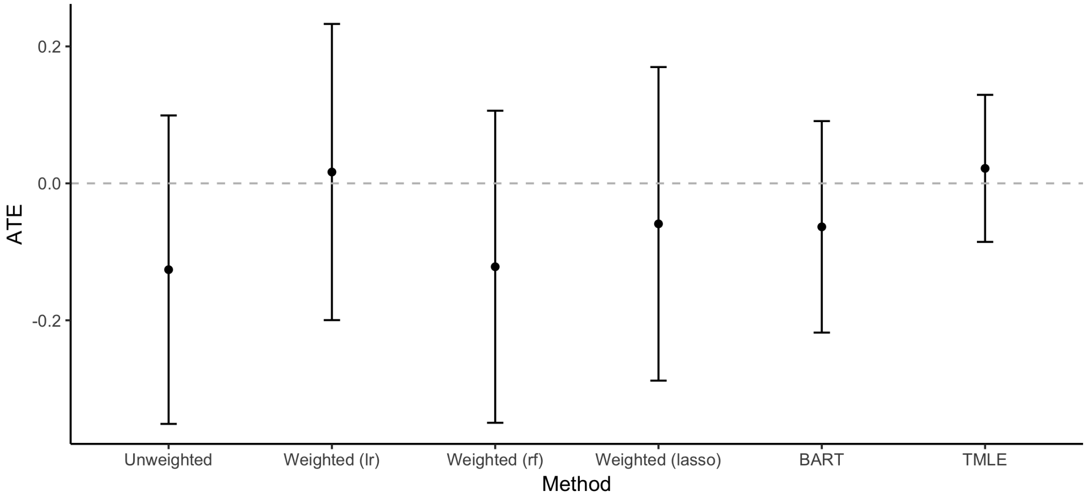


Figure 4.3: Average Treatment Effect of Topiramate on Methamphetamine Use Reported in Followup by Generalizability Method

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 2: 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, Ebnesajjad, Cole, & Stuart, 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.

References

Ackerman, B. (2018). *Generalize: Generalizing average treatment effects from rcts to target populations*. Retrieved from http://www.github.com/benjamin-ackerman/generalizeR

Ackerman, B., Schmid, I., Rudolph, K. E., Seamans, M. J., Susukida, R., Mojtabai, R., & Stuart, E. A. (2019). Implementing statistical methods for generalizing randomized trial findings to a target population. *Addictive behaviors*, *94*, 124–132.

Bhattacharyya, A. (1943). On a measure of divergence between two statistical populations defined by their probability distributions. *Bulletin of the Calcutta Mathematical Society 35, p. 99-109*.

Bhattacharyya, A. (1946). On a measure of divergence between two multinomial populations. *Sankhyā: the indian journal of statistics*, 401–406.

Cole, S. R., & Stuart, E. A. (2010). Generalizing evidence from randomized clinical trials to target populations: The actg 320 trial. *American journal of epidemiology*, *172*(1), 107–115.

Dababnah, S., & Parish, S. L. (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.

Elkashef, A., Kahn, R., Yu, E., Iturriaga, E., Li, S.-H., Anderson, A., … McSherry, F., et al. (2012). Topiramate for the treatment of methamphetamine addiction: A multi-center placebo-controlled trial. *Addiction*, *107*(7), 1297–1306.

Flay, B. R. (1986). Efficacy and effectiveness trials (and other phases of research) in the development of health promotion programs. *Preventive medicine*, *15*(5), 451–474.

Gruber, S., & Van Der Laan, M. J. (2009). Targeted maximum likelihood estimation: A gentle introduction.

Hill, J. L. (2011). Bayesian nonparametric modeling for causal inference. *Journal of Computational and Graphical Statistics*, *20*(1), 217–240.

Imai, K., King, G., & Stuart, E. A. (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, T. R. (2006). Beyond efficacy: The star* d trial. *American Journal of Psychiatry*, *163*(1), 5–7.

Johnson, B. A. (2015). Nida-csp-1025. Retrieved from nida.nih.gov/study/nida-csp-1025

Johnson, B. A., Rosenthal, N., Capece, J. A., Wiegand, F., Mao, L., Beyers, K., … Ciraulo, D. A., et al. (2007). Topiramate for treating alcohol dependence: A randomized controlled trial. *Jama*, *298*(14), 1641–1651.

Kampman, K. M., Pettinati, H., Lynch, K. G., Dackis, C., Sparkman, T., Weigley, C., & OâBrien, C. P. (2004). A pilot trial of topiramate for the treatment of cocaine dependence. *Drug and alcohol dependence*, *75*(3), 233–240.

Kern, H. L., Stuart, E. A., Hill, J., & Green, D. P. (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. eprint: http://dx.doi.org/10.1080/19345747.2015.1060282

Lee, B. K., Lessler, J., & Stuart, E. A. (2010). Improving propensity score weighting using machine learning. *Statistics in medicine*, *29*(3), 337–346.

Lesko, C. R., Buchanan, A. L., Westreich, D., Edwards, J. K., Hudgens, M. G., & Cole, S. R. (2017). Generalizing study results: A potential outcomes perspective. *Epidemiology*, *28*(4), 553–561.

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.

Nguyen, T. Q., Ebnesajjad, C., Cole, S. R., Stuart, E. A., et al. (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.

R Core Team. (2017). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. Vienna, Austria. Retrieved from https://www.R-project.org/

Rosenbaum, P. R., & Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika*, *70*(1), 41–55.

Rubin, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of educational Psychology*, *66*(5), 688.

Rubin, D. B. (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, D. B. (2008). For objective causal inference, design trumps analysis. *The Annals of Applied Statistics*, 808–840.

Rudolph, K. E., Díaz, I., Rosenblum, M., & Stuart, E. A. (2014). Estimating population treatment effects from a survey subsample. *American journal of epidemiology*, *180*(7), 737–748.

Shadish, W. R., Cook, T. D., & Campbell, D. T. (2002). *Experimental and quasi-experimental designs for generalized causal inference*. Wadsworth Cengage learning.

Stuart, E. A. (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.

Stuart, E. A., & 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.

Stuart, E. A., Cole, S. R., Bradshaw, C. P., & Leaf, P. J. (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.

Stuart, E. A., Ackerman, B., & Westreich, D. (2017). Generalizability of randomized trial results to target populations: Design and analysis possibilities. *Research on Social Work Practice*, 1049731517720730.

Susukida, R., Crum, R. M., Stuart, E. A., 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.

Susukida, R., Crum, R. M., Ebnesajjad, C., Stuart, E. A., & Mojtabai, R. (2017). Generalizability of findings from randomized controlled trials: Application to the national institute of drug abuse clinical trials network. *Addiction*.

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, H. I., Hayden, V. C., & Pontes, V. P. (2009). Selection criteria and generalizability within the counterfactual framework: Explaining the paradox of antidepressant-induced suicidality? *Clinical Trials*, *6*(2), 109–118.

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, C. C., Jones, S. M., Li-Grining, C. P., Pressler, E., & Gao, Q. (2010). Dosage effects on school readiness: Evidence from a randomized classroom-based intervention. *Social Service Review*, *84*(4), 615–655.

Chapter 5

Conclusion

This dissertation work has focused on several ways to acquire and utilize different types of external data to improve RCT inferences. After highlighting challenges in finding suitable target population data to make RCT generalizations, Chapter 2 described an opportunity to use complex surveys, and demonstrated that it is crucial to incorporate the complex survey weights when estimating the PATE. Omitting the survey weights can be thought of as generalizing to an entirely different population, one that has the demographics of the survey sample rather than the target population of interest. While the demographic differences between a survey sample and its intended target population may not be that large for some analytic survey datasets, it can be noticeable for others, particularly when surveys are designed to heavily over- or under-sample certain groups. Chapter 3 elaborated on best practices for using external validation studies to measure and properly correct for measurement error in self-reported trial outcomes. Directly applying such measurement error corrections to a continuous trial outcome may introduce additional biases to the ATE estimate when the error is differential with respect to a set of covariates, and those covariates differ in distribution between the trial and validation sample. In such cases, propensity score-type weighting methods can help address this transportability concern; however, under extremely large differences, it may still be inappropriate to transport corrections from an external validation sample to a trial outcome.

In both Chapters 2 and 3, important questions around data compatibility, transportability and data synthesis were raised, accompanied by the development and application of statistical methodology to enhance one's ability to supplement a trial with external sources of data. Chapter 4 then provided practical guidance on seeking and harmonizing such data for generalization purposes, and introduced user-friendly software for method implementation. Each component of this dissertation aimed to address challenges that arise in practice when using RCTs for population-level decision making by providing methods and tools to improve upon one's ability to do so.

Several opportunities for future work present themselves from this dissertation research. One commonly raised challenge, for both making generalizations and for making measurement error corrections, is the public availability of high quality external data. Some data issues are too great to be fixed by statistical methodology alone, and the ability to implement the methods developed and discussed in this dissertation depends on the quality and breadth of external data. For instance, if one wishes to generalize findings from a trial where participants are all over the age of 65, yet the target population dataset consists of individuals between the ages of 18-40 only, then that particular population dataset is simply unsuitable for the research question at hand. Datasets derived from electronic health records (EHR) are promising and expansive sources that may also help address some of these positivity concerns, especially when researchers are interested in generalizing effects from RCTs to particular hospital-based patient populations. Given that EHR and other administrative databases are not necessarily collected with research intentions, further research should be conducted to assess the appropriateness of EHR data for RCT-to-population generalizations.

Similarly, transporting can be challenging when a key variable that moderates treatment effect (for generalizability) or recall bias (for measurement error) is partially or fully unobserved between the trial and external data. Sensitivity analyses for generalizing RCT findings with unobserved effect modifiers have been proposed, and should be a key component of any application with this concern (Chan, 2019; Nguyen, Ackerman, Schmid, Cole, & Stuart, 2018). Additional work is needed to extend such sensitivity analyses when using complex survey population data, pertaining to partially unobserved moderators in both the transportability sample membership model *and* the generation of the survey weights. Future software development for generalizability should also consider incorporating functionality for carrying out such sensitivity analyses.

Finally, this work focused on transportability when addressing issues of internal and external validity in randomized trials separately. An exciting direction of this work would be to tackle both issues *together* methodologically, similar to how Beesley and Mukherjee (2019) handle sample selection bias and outcome misclassification in EHR association studies. Furthermore, there is an ongoing debate on the trade-offs between study design and validity

(Westreich, Edwards, Lesko, Cole, & Stuart, 2019), and future research would benefit from addressing both internal and external validity, in experimental and non-experimental studies alike.

The methodological contributions of this dissertation were demonstrated through applications in randomized trials related to substance use disorders, nutrition and cardiovascular health. Additionally, the issues of generalizability and measurement error addressed have also been documented in trials spanning other disciplines such as medicine (Rubin, 2008), child development (Dababnah & Parish, 2016), social work (Stuart, Ackerman, & Westreich, 2017; Zhai et al., 2010) and education (Tipton & Olsen, 2018). I hope this work on transportability will continue to have a broad impact on public health, public policy-making and social good.

References

Beesley, L. J., & Mukherjee, B. (2019). Statistical inference for association studies using electronic health records: Handling both selection bias and outcome misclassification. *medRxiv*.

Chan, W. (2019). An evaluation of bounding approaches for generalization. *The Journal of Experimental Education*, 1–31.

Dababnah, S., & Parish, S. L. (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.

Nguyen, T. Q., Ackerman, B., Schmid, I., Cole, S. R., & Stuart, E. A. (2018). Sensitivity analyses for effect modifiers not observed in the target population when generalizing treatment effects from a randomized controlled trial: Assumptions, models, effect scales, data scenarios, and implementation details. *PloS one*, *13*(12).

Rubin, D. B. (2008). For objective causal inference, design trumps analysis. *The Annals of Applied Statistics*, 808–840.

Stuart, E. A., Ackerman, B., & Westreich, D. (2017). Generalizability of randomized trial results to target populations: Design and analysis possibilities. *Research on Social Work Practice*, 1049731517720730.

Tipton, E., & Olsen, R. B. (2018). A review of statistical methods for generalizing from evaluations of educational interventions. *Educational Researcher*, *47*(8), 516–524.

Westreich, D., Edwards, J. K., Lesko, C. R., Cole, S. R., & Stuart, E. A. (2019). Target validity and the hierarchy of study designs. *American journal of epidemiology*, *188*(2), 438–443.

Zhai, F., Raver, C. C., Jones, S. M., Li-Grining, C. P., Pressler, E., & Gao, Q. (2010). Dosage effects on school readiness: Evidence from a randomized classroom-based intervention. *Social Service Review*, *84*(4), 615–655.

Appendix A

Supplemental Material for Chapter 2

R Code

All code for the simulation and data example can be found in the following GitHub repository: https://github.com/benjamin-ackerman/generalizability_svys

Derivation of Population Estimand E[Y(a)] for single binary X

$$\begin{aligned}
E[Y(a)] &= \sum_x E[Y(a)|X=x]P(X=x) && \text{Total expectation} \\
&= \sum_x E[Y(a)|X=x,S=1]P(X=x) && S \perp\!\!\!\perp Y(a)|X \\
&= \sum_x E[Y(a)|A=a,X=x,S=1]P(X=x) && A \perp\!\!\!\perp Y(a)|X,S=1 \\
&= \sum_x E[Y|A=a,X=x,S=1]P(X=x) && \text{Consistency} \\
&= \sum_x E[Y|A=a,X=x,S=1]P(X=x|S=1)
\end{aligned}$$

$$\times \frac{P(S=1)}{P(S=1|X=x)} \qquad \text{Bayes thm}$$

$$E[Y(a)] = \sum_x E[Y|A=a, X=x, S=1]P(X=x|S=1)$$

$$\times \left(\frac{P(S=2|X=x)P(S=1)}{P(S=1|X=x)P(S=2)} \right)$$

$$\times \left(\frac{P(S=2)}{P(S=2|X=x)} \right) \qquad \text{multiplying by 1}$$

Double Bootstrap

In order to account for uncertainty in the survey when using it for generalizations, we propose using a double-bootstrapping approach to estimate the variability of the PATE estimates. Similarly to how a bootstrap involves sampling with replacement many times and looking at the distribution of estimates across bootstrap runs, we sample both the trial *and* the survey with replacement in each bootstrap run. Within each bootstrap iteration, we re-sample the trial with replacement (sample size equal to that of the trial). We also re-sample the survey using a stratified approach described by Valliant, Dever, and Kreuter (2013). We define survey strata by deciles of the survey weights. For stratum h with sample size n_h, we sample with replacement $m_h = n_h - 1$ subjects. We adjust the survey weight d_k of subject k to equal

$$d_k^* = d_k \frac{n_h}{n_h - 1} m_{hi}^*$$

where m_{hi}^* is the number of times subject k is sampled for that given bootstrap run. Therefore, if the subject is selected once, their new weight is equal to

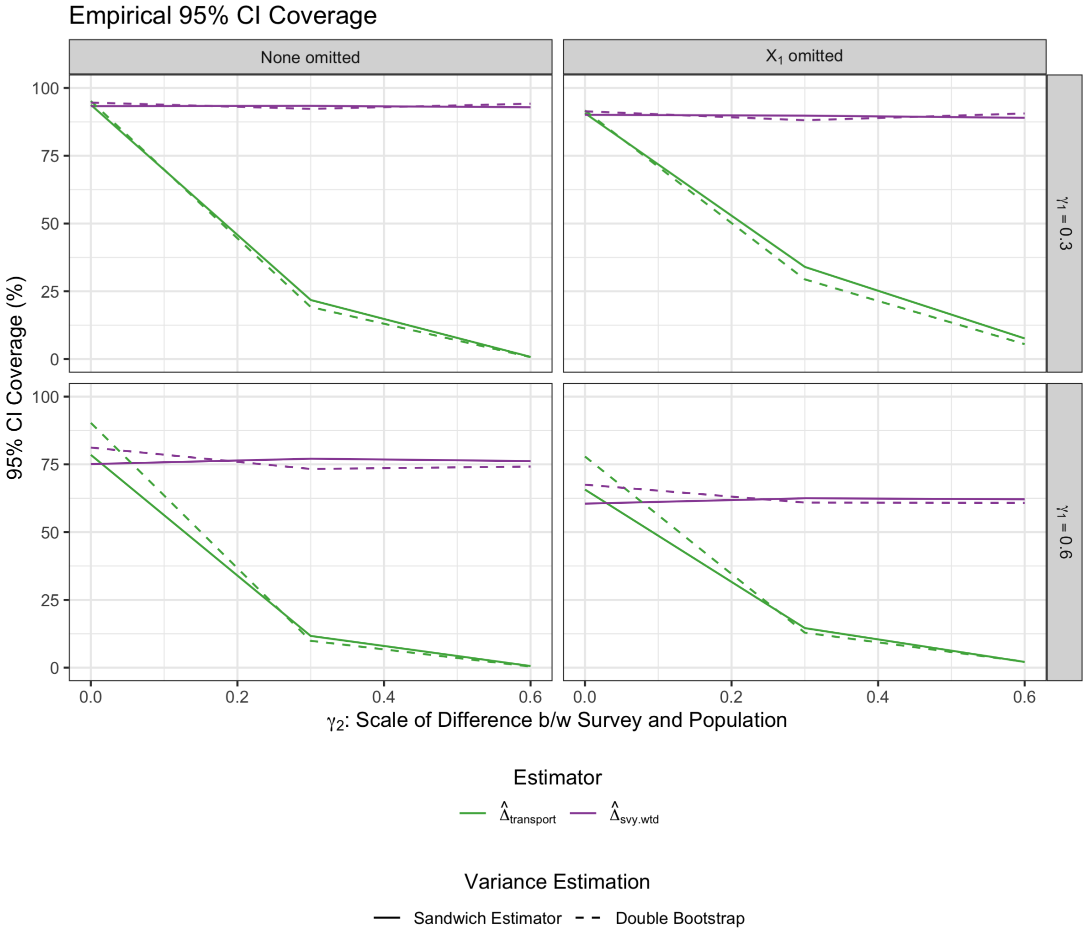


Figure A.1: Empirical coverage of the transportability estimators using the double bootstrap approach to estimate the variance.

$d_k \frac{n_h}{n_h-1}$. If they are selected twice, their new weight is equal to $d_k \frac{2n_h}{n_h-1}$, and so forth. Figure A.1 compares the empirical 95% coverage of the transported estimators using this double bootstrap approach on a subset of the simulation scenarios to the standard sandwich variance estimator used for Figure 2.3. Note that the results across the different approaches are quite similar, though the double bootstrap yields slightly better coverage when the trial differs more from the target population (bottom row).

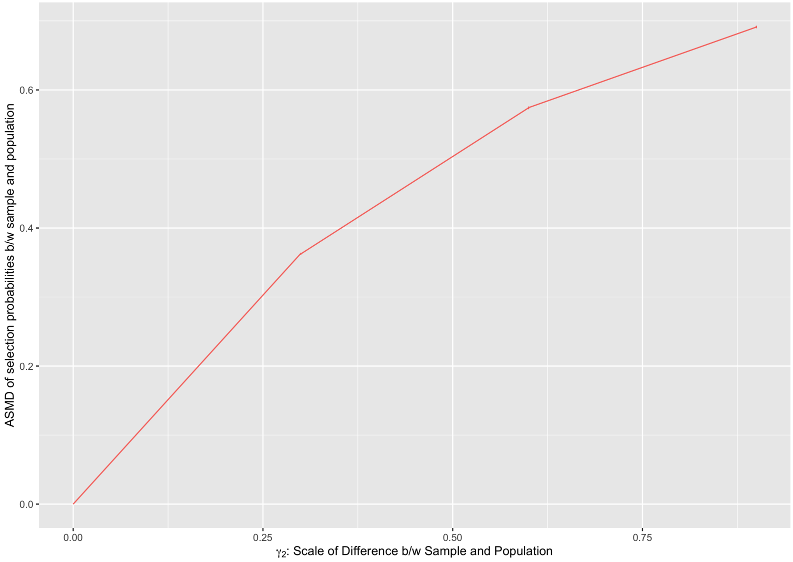


Figure A.2: Relationship between γ_2, the scaling parameter for survey selection, and the ASMD of survey selection probabilities between the survey sample and the target population.

Appendix B

Supplemental Material for Chapter 3

R Code

All code for the simulation and data example can be found in the following GitHub repository:

https://github.com/benjamin-ackerman/ME_transportability

Derivation of Standard Error for $\hat{\mu}_0^{weighted}$

Recall that

$$\hat{\mu}_0^{weighted} = \frac{\sum_{i=1}^{n_v} w_i(Y_i - Z_i)}{\sum_{i=1}^{n_v} w_i}$$

Let Y^* and Z^* be the weighted vectors of Y and Z in the validation sample. Since Y and Z are paired measurements and are not independent, $\text{var}(Y^* - Z^*) = \text{var}(Y^*) + \text{var}(Z^*) - 2\text{cov}(Y^*, Z^*)$.

According to Price (1972),

$$\text{cov}(Y^*, Z^*) = \frac{\left[\sum_i^{n_v} w_i (Y_i - \bar{Y}^*)(Z_i - \bar{Z}^*)\right]}{\sum_i^{n_v} w_i}$$

where

$$\bar{Y}^* = \frac{\sum_{i=1}^{n_v} w_i Y_i}{\sum_{i=1}^{n_v} w_i}$$

$$\bar{Z}^* = \frac{\sum_{i=1}^{n_v} w_i Z_i}{\sum_{i=1}^{n_v} w_i}$$

$$\text{var}(Y^*) = \text{cov}(Y^*, Y^*)$$

$$\text{var}(Z^*) = \text{cov}(Z^*, Z^*)$$

Therefore, $\text{se}_{\hat{\mu}_0^{weighted}} = \sqrt{\text{var}(Y^* - Z^*)/n}$ and can be used to construct a 95% confidence interval.

Supplemental Tables and Figures

Table B.1: Scaled coefficients for covariates by model type

	X_1	X_2	X_3	X_4
S model	γ_1	0	$\frac{1}{2}\gamma_1$	$2\gamma_1$
Y model	0	γ_2	$2\gamma_2$	$\frac{1}{2}\gamma_2$

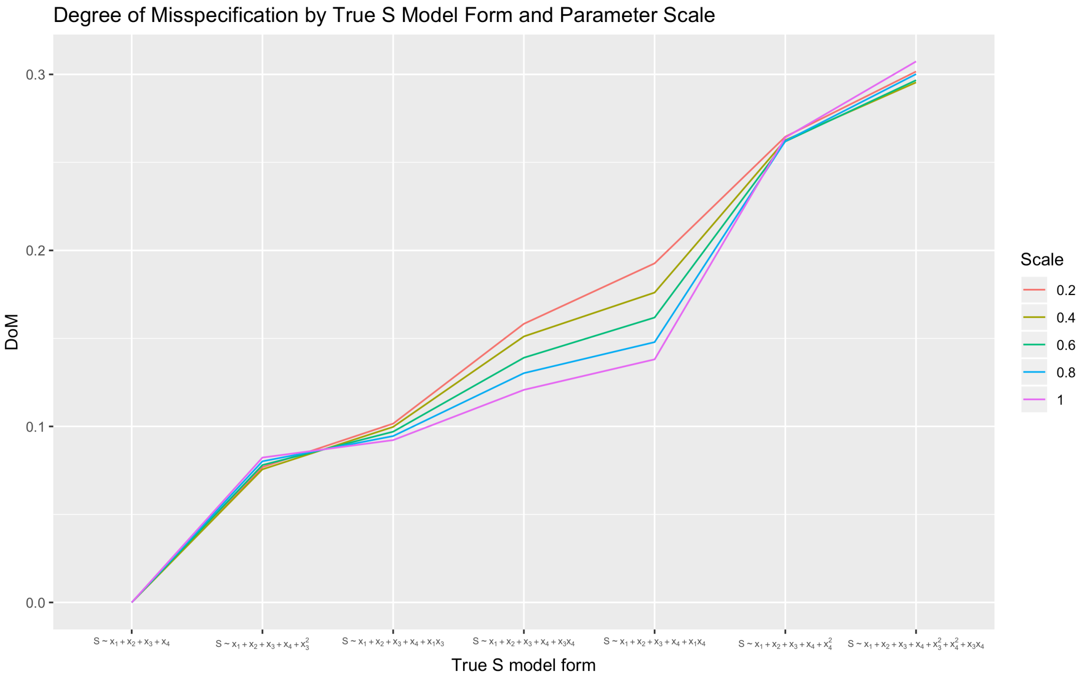


Figure B.1: Degree of Misspecification (DoM) of fitting the main effects model under different true S model forms. As expected, when the true model is the main effects model, there is no misspecification, and the most misspecified model is the model with the most terms. Also note that the degree of misspecification increases when the true S model form depends on the more influential covariates.

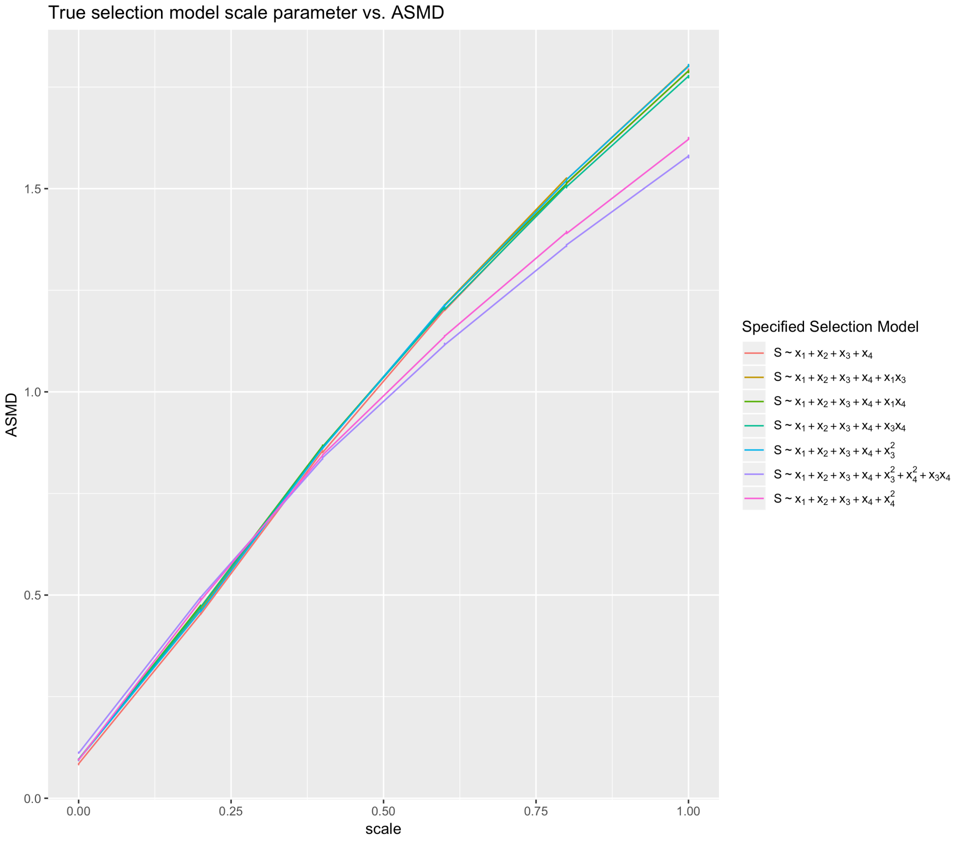


Figure B.2: The relationship between the 'scale' parameter (γ_1) and the ASMD between the true selection probabilities across the trial and the validation data. There's slight variation across the different true selection models, though this could be due to simulation variability.

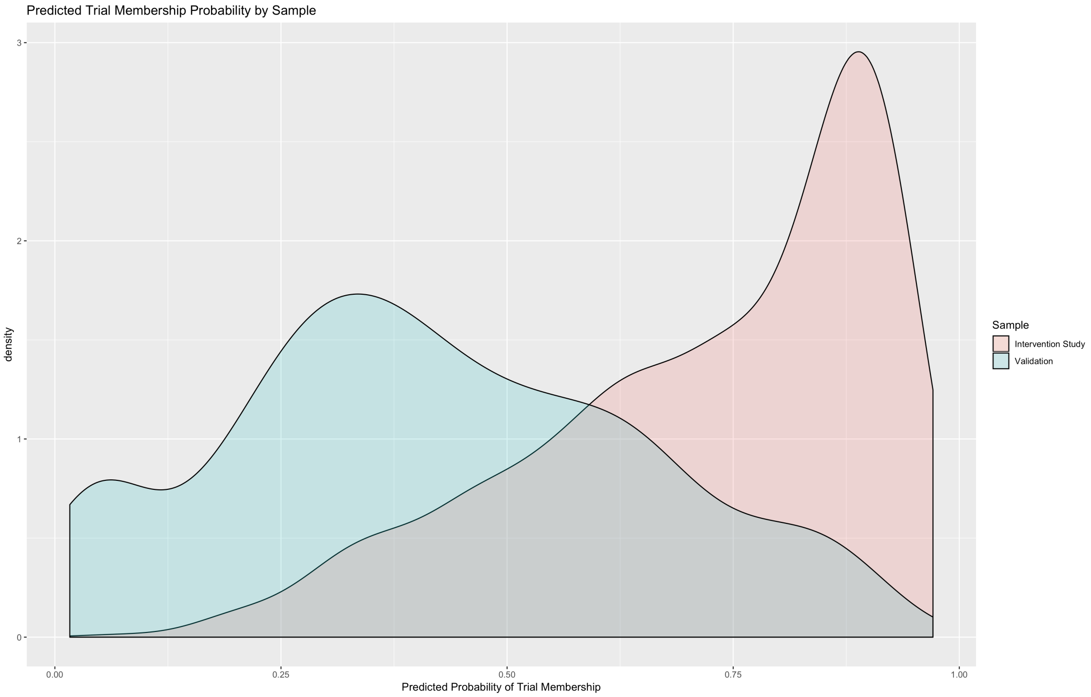


Figure B.3: Predicted Probabilities of Trial Membership by Sample

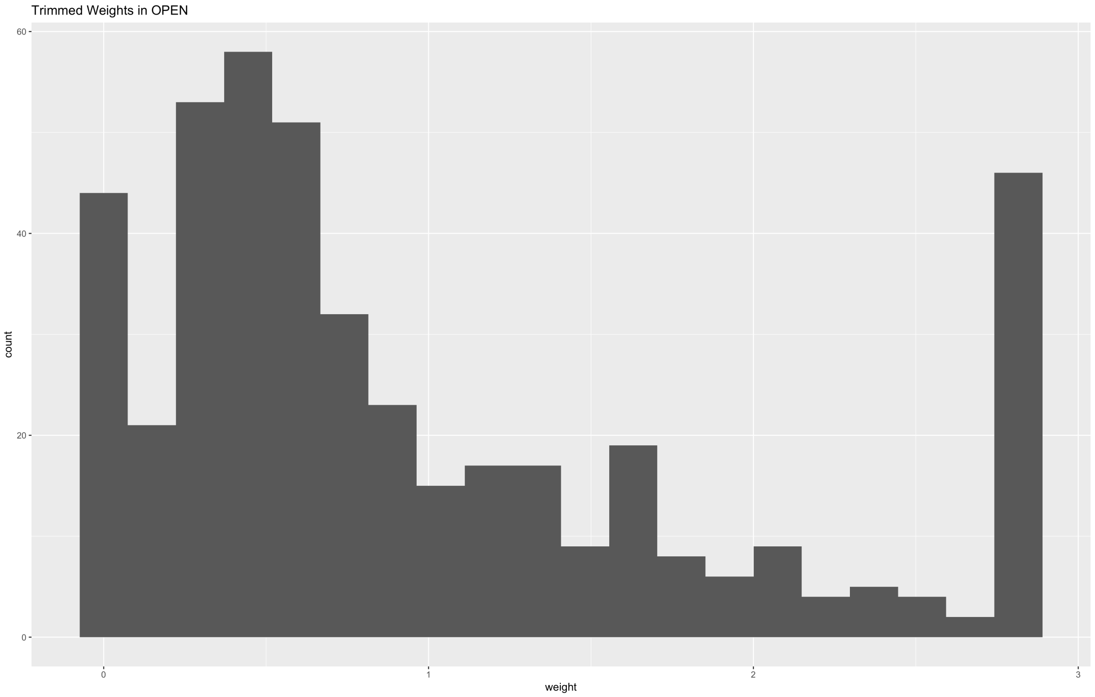


Figure B.4: Distribution of the trimmed weights in OPEN validation sample

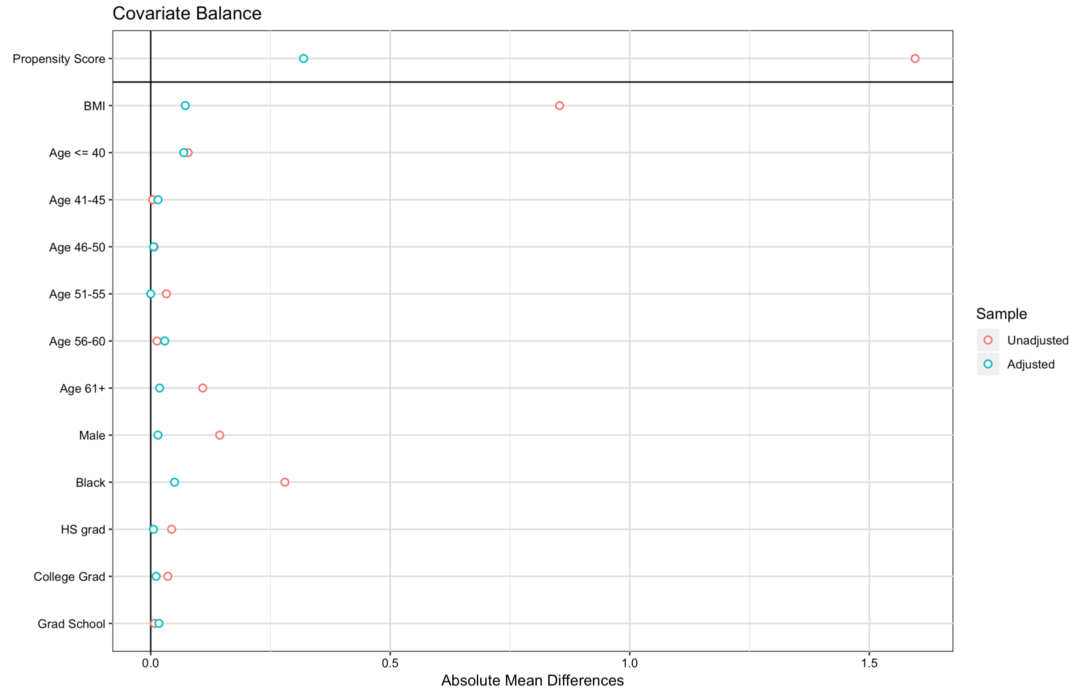


Figure B.5: Love plot comparing the covariate distributions in the intervention trial PREMIER (pink) and validation sample OPEN (blue), pre-and-post weighting the validation sample

Table B.2: Self-reported and Urinary Sodium Outcomes by Study and Treatment Group

	OPEN	PREMIER			
		6 months		18 months	
	Control	Control	Treatment	Control	Treatment
Self-Reported (Y)	8.220	7.850	7.640	7.840	7.690
Urine (Z)	8.450	8.070	7.970	8.140	8.020
μ_a^s	-0.227	-0.228	-0.323	-0.299	-0.327

Table B.3: T-test comparing measurement error by treatment group in PREMIER

Timepoint	bias$_{\hat{\Delta}}$	95% CI	p-value
6 months	0.1108736	(0.0162, 0.206)	0.0217971
18 months	0.0282937	(-0.065, 0.122)	0.5517550

Table B.4: Regression Coefficients for modeling effect of covariates on measurement error term

term	estimate	std.error	statistic	p.value
Intercept	5.90	0.41	14.53	0.0000000
log(sodium urine)	0.24	0.05	4.74	0.0000029
Sex	0.20	0.04	5.33	0.0000002
Age 41-45	0.06	0.10	0.65	0.5178708
Age 46-50	0.06	0.10	0.63	0.5302275
Age 51-55	-0.07	0.10	-0.78	0.4365246
Age 56-60	-0.06	0.10	-0.61	0.5419625
Age > 60	0.00	0.09	0.03	0.9789308
BMI	0.00	0.00	1.21	0.2275847
College Education	0.09	0.05	1.72	0.0864365
Grad School Education	0.10	0.05	1.84	0.0668818
Black	-0.28	0.07	-3.89	0.0001163

Appendix C

Supplemental Material for Chapter 4

In this appendix, we demonstrate the implementation of the methods described in Section 4.3 on the data example described in Section 4.5 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 4.3.2.

4. Target population “trimming": assess can check for any violations of the coverage assumption (Assumption 3). 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 4.3.3. 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 + education + employment + methprior
##
##                  Min.  1st Qu.  Median     Mean 3rd Qu.     Max.
## Trial (n = 137)     0 0.002019 0.01075 0.011820 0.01739 0.04932
## Pop (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.

Benjamin Ackerman

Biostatistician, Data Scientist

☎ (310) 963-0578 | ✉ backer10@jhu.edu | ⌂ www.benjaminackerman.com | benjamin-ackerman | in backer10 | @backerman150

Education

Johns Hopkins Bloomberg School of Public Health
Baltimore, MD

PhD, Biostatistics — *March 2020*

- **Advisor:** Dr. Elizabeth A. Stuart
- **Dissertation Title:** *"Statistical Methods for Transportability: Addressing External Validity and Measurement Error Concerns in Randomized Trials"*

Johns Hopkins University
Baltimore, MD

Bachelor of Arts, Public Health Studies — *May 2015*

- **Minor:** Applied Mathematics and Statistics
- **Honors Thesis:** *"The Association Between Genetic Variants and IQ among Individuals with Autism Spectrum Disorders"* | *Advisor: Dr. Yin Yao*

Professional Experience

SAJE Consulting
Baltimore, MD

Statistician/Programmer — *March 2016 - Present*

- Conducted data analysis and produced publication-ready graphics for various research studies conducted by pharmaceutical and biotechnology companies
- Contributed to analyses for reports to the following regulatory agencies: FDA, EMA and NOMA.

Data Science for Social Good
Chicago, IL

Fellow — *Summer 2018*

- Partnered with AllianceChicago to build a predictive model to identify patients at risk of developing Type 2 Diabetes using de-identified electronic health records (EHR) data.

Research Experience

Graduate Research Assistant, Johns Hopkins Bloomberg School of Public Health
Baltimore, MD

Department of Biostatistics — *June 2016 - Present*

- Worked with thesis advisor Dr. Elizabeth Stuart to evaluate and develop propensity score-type statistical methods for assessing and improving upon the generalizability of randomized controlled trials.
- Developed methods to improve upon the transportability of measurement error correction from external validation samples to lifestyle intervention trials.

Department of Mental Health — *Sept 2016 - Aug 2017*

- Collaborated with Dr. Heather Volk to examine the relationship between prenatal air pollution exposure and risk of Autism Spectrum Disorders, using data from the Boston Birth Cohort.

Center for Public Health and Human Rights — *Oct 2015 - Oct 2017*

- Worked with Dr. Tonia Poteat to estimate HIV risk among MSM, transgender and gender variant populations in Africa
- Merged survey data across eight different countries, and explored issues regarding survey methodology and gathering data on sexual orientation and gender identity.

Undergraduate Research Assistant

Unit on Statistical Genomics, *National Institute of Mental Health* *Summer 2014*

- Evaluated the association between genetic variants and IQ among individuals with Autism Spectrum Disorders using statistical analysis software PLINK and FBAT.

Department of Infectious Diseases, *Assaf Harofeh Medical Center* *Sept - Dec 2013*

- Assessed the epidemiology of carbapenem-resistant enterobacter species in patients at Assaf Harofeh Medical Center in Israel and Detroit Medical Center.

Summer Institute Training in Biostatistics (SIBS), *Columbia University Mailman School of Public Health* *Summer 2013*

- 8-week program supported by the National Heart Lung and Blood Institute
- Conducted research in Center for Behavioral Cardiovascular Health with Dr. Keith Diaz.

Publications

Peer-Reviewed/In Press:

1. Schmid, I., Rudolph, K.E., Nguyen, T.Q., Hong, H., Seamans, M.J., **Ackerman, B.**, Stuart, E.A. (2020). "Comparing the performance of statistical methods that generalize effect estimates from randomized controlled trials to much larger target populations." *Communications in Statistics - Simulation and Computation*.
2. **Ackerman, B.**, Schmid, I., Rudolph, K. E., Seamans, M. J., Susukida, R., Mojtabai, R., Stuart, E. A. (2019). "Implementing statistical methods for generalizing randomized trial findings to a target population." *Addictive Behaviors*, 94, 124-132. `https://doi.org/10.1016/j.addbeh.2018.10.033`
3. Nguyen, T. Q., **Ackerman, B.**, Schmid, I., Cole, S., Stuart, E.A. (2018). "Sensitivity analyses for effect modifiers not observed in the target population when generalizing treatment effects from a randomized controlled trial: Assumptions, models, effect scales, data scenarios, and implementation details." *PLoS One*, `https://doi.org/10.1371/journal.pone.0208795`
4. Lenis, D., **Ackerman, B.**, Stuart, E.A. (2018). "Measuring model misspecification: Application to propensity score methods on complex survey data." *Computational Statistics & Data Analysis*, 128, 48-57. `https://doi.org/10.1016/j.csda.2018.05.003`
5. Poteat, T., **Ackerman, B.**, Diouf, D., Ceesay, N., Mothopeng, T., Odette, K-Z, et al. (2017). "HIV prevalence and behavioral and psychosocial factors among transgender women and cisgender men who have sex with men in 8 African countries: A cross-sectional analysis." *PLoS Med*, 14(11): e1002422. `https://doi.org/10.1371/journal.pmed.1002422`
6. Stuart, E. A., **Ackerman, B.**, Westreich, D. (2017). "Generalizability of randomized trial results to target populations: Design and analysis possibilities." *Research on Social Work Practice*, 28(5), 532-537.
7. Tao, Y., Gao, H., **Ackerman, B.**, Guo, W., Saffen, D., Shugart, Y. Y. (2016). "Evidence for contribution of common genetic variants within chromosome 8p21.2-8p21.1 to restricted and repetitive behaviors in autism spectrum disorders." *BMC Genomics*, 17(1), 163.
8. Lazarovitch, T., Amity, K., Coyle, J. R., **Ackerman, B.**, Tal-Jasper, R., Ofer-Friedman, H., et al. (2015). "The complex epidemiology of carbapenem-resistant enterobacter infections: A multicenter descriptive analysis." *Infection Control and Hospital Epidemiology*,

36(11), 1283-1291.

Non Peer-Reviewed/In Press:

9. Stuart, E.A. and **Ackerman, B.** (2019). “Commentary on Yu et al.: Opportunities and Challenges for Matching Methods in Large Databases.” *Statistical Science*, In Press.

Preprints/Under Review:

10. **Ackerman, B.**, Lesko, C.R., Siddique, J., Susukida, R., Stuart, E.A. (2020). “Generalizing randomized trial findings to a target population using complex survey population data.” Under Review. `http://arxiv.org/abs/2003.07500`
11. Lesko, C.R., **Ackerman, B.**, Webster-Clark, M., Edwards, J.K. (2020). “Target validity: bringing treatment of external validity in line with internal validity.” Under Review.
12. Seamans, M.J., Hong, H., **Ackerman, B.**, Schmid, I., Stuart, E.A. (2020). “Generalizability of subgroup effects.” Under Review.
13. Ackerman, S.E., Gonzalez, J.C., Pearson, C.I., Gregorio, J.D., Hartmann, F.J., Kenkel, J.A., Luo, A., Ho, Po, LeBlanc, H., Kimmey, S.C., Nguyen, M.L., Paik, J.C., Sheu, L.Y., **Ackerman, B.**, ..., Alonso, M.N. (2020). “Immune-stimulating antibody conjugates elicit robust myeloid activation and durable anti-tumor immunity.” Under Review.
14. **Ackerman, B.**, Siddique, J., Stuart, E.A. (2019). “Transportability of outcome measurement error correction: from validation studies to intervention trials.” `http://arxiv.org/abs/1907.10722`

Honors and Awards

2020	**Student Travel Award**, 13th International Conference on Health Policy Statistics (ICHPS)
2019	**Special Award for Outstanding Student Service**, Johns Hopkins Department of Biostatistics
2019	**Best Student Paper Award - 3rd Place**, Joint Statistical Meetings (JSM) - Biopharmaceutical Section
2019	**3 Minute Thesis (3MT) Competition - 3rd Place + Alumni Choice Winner**, Johns Hopkins University
2017	**Delta Omega Poster Competition - 2nd Place (Applied Research)**, Johns Hopkins Bloomberg School of Public Health
2015	**Best Senior Thesis in Public Health**, Johns Hopkins University
2011-2015	**Dean’s List**, Johns Hopkins University

Computing Projects and Resources

generalize (R Package)

- Software for implementing statistical methods to assess and improve upon generalizability of RCTs to well-defined target population
- `https://benjamin-ackerman.github.io/generalize`

How will the House Tax Bill Impact Graduate Students? (R Shiny App)

- Web app to calculate estimated 2018 federal income tax under proposed H.R. 1 tax bill
- *Featured in Science Magazine* (see section on Tuition Waivers)
- `https://benjaminackerman.shinyapps.io/GOPtax2017/`

Professional Activities

Outreach 2017-2018 “This is Public Health” Ambassador for the Association of Schools and Programs of Public Health (ASPPH)

Reviewer *Biometrics, PLOS ONE, Pharmaceutical Statistics, The Journal of Experimental Education,Sexuality Research and Social Policy*

Membership American Statistical Association (ASA)
Eastern North America Region of the International Biometrics Society (ENAR)
Society for Research on Educational Effectiveness (SREE)

Academic Service

Current **Co-President**, JHSPH Mental Health Grad Network

Current **Tea Time Organizer**, Johns Hopkins Department of Biostatistics

2018-2019 **PhD Representative to Faculty Meetings**, Johns Hopkins Department of Biostatistics

Talks and Presentations

Invited Talks

2019 **Sensitivity Analysis for Unobserved Effect Modification when Generalizing Findings from Randomized Trials to Target Populations**
FCSM/WSS Workshop on Sensitivity Analysis with Integrated Data, *Washington, DC*.

Using Statistics and Data Science for Public Health and Social Good
Department of Global and Community Health, George Mason University, *Fairfax, VA*.
Invited talk for National Public Health Week 2019

Conference Talks and Posters

2020 **Generalizing Randomized Trial Findings to a Target Population using Complex Survey Population Data**
- *ENAR Spring Meeting*, *Virtual (due to COVID-19)*, Contributed Talk.
- *13th International Conference on Health Policy Statistics (ICHPS)*, *San Diego, CA*, Poster.
- *The Statistical and Applied Mathematical Sciences Institute (SAMSI) Program on Causal Inference Opening Workshop*, *Durham, NC*, Poster.

2019 **Calibrating Validation Samples when Correcting for Measurement Error in Intervention Study Outcomes**
- *Joint Statistical Meetings (JSM)*, *Denver, CO*, Topics Contributed Talk.
- *ENAR Spring Meeting*, *Philadelphia, PA*, Contributed Talk.

generalize: Statistical Software for Implementing Methods to Generalize Randomized Trial Findings to a Well-Defined Target Population
- *Society for Research on Educational Effectiveness (SREE) Spring 2019 Conference*, *Washington, DC*, Poster.
- *Institute of Education Sciences (IES) Annual PI Meeting*, *Washington, DC*, Poster.

2018 **Supporting Proactive Diabetes Screenings to Improve Health Outcomes**
Data Science for Social Good Data Fest, *Chicago, IL*, Speed Talk and Poster.

Sensitivity Analysis for an Unobserved Moderator in Trial-to-Target-Population Generalization of Treatment Effects
Society for Research on Educational Effectiveness (SREE) Spring 2018 Conference, Washington, DC, Contributed Talk.

Estimating Population Effects: Case Study of Generalizing Results of a Methamphetamine Dependence Trial
12th International Conference on Health Policy Statistics (ICHPS), Charleston, SC, Contributed Talk.

2017 **Characterizing the Burden of HIV and Specific Vulnerabilities among Transgender Women compared to Men who have Sex with Men across Eight Sub-Saharan African Countries**
- *Joint Statistical Meetings (JSM)*, Baltimore, MD, Contributed Talk.
- *Johns Hopkins LGBT Research Day*, Baltimore, MD, Talk.

2016 **Sensitivity Analysis for an Unobserved Moderator in RCT-to-Target-Population Generalization of Treatment Effects**
Joint Statistical Meetings (JSM), Chicago, IL, SPEED Talk and Poster.

2015 **Genetic Variants and IQ Among Individuals with Autism Spectrum Disorder**
- *6th Annual Undergraduate Conference in Public Health*, Baltimore, MD, Talk and Poster.
- *National Institutes of Health Summer Research Program Poster Day (2014)*, Bethesda, MD, Poster.

Teaching Experience

2018-2019 **Advanced Data Science I (Guest Lecturer)**, *JHSPH (25 graduate students)*
Professor: Dr. Stephanie Hicks, Dr. Roger Peng
Designed and led a 80-minute tutorial on creating R packages, Shiny apps and GitHub pages

Teaching Assistant

2018-2020 **Causal Inference in Medicine and Public Health I**, *JHSPH (60 graduate students)*
Professor: Dr. Elizabeth Stuart
TA. Held weekly office hours to review causal inference topics for both experimental and non-experimental studies, gave lecture on generalizability of randomized controlled trials.

2016-2019 **Public Health Biostatistics**, *JHU (225 undergraduate students)*
Professor: Dr. Margaret Taub, Dr. Leah Jager
Section Instructor. Reviewed introductory statistical concepts and R programming skills.

2017-2018 **Statistical Methods in Public Health III & IV**, *JHSPH (500 MPH students)*
Professor: Dr. Marie Diener-West, Dr. Leah Jager, Dr. Jim Tonascia
TA. Held weekly office hours for to review regression topics, provided assistance with STATA programming.

2013-2014 **Public Health Biostatistics**, *JHU (200 undergraduate students)*
Professor: Dr. Scott Zeger, Dr. Margaret Taub, Dr. Leah Jager
Learning Den Tutor and Guest Lecturer. Held biweekly small group review sessions.

Computing Skills

Languages	*Proficient:* R, SQL
	Intermediate: Python, SAS, Stata, SPSS
Markup	LaTeX, knitr, markdown, Sweave
Other	Git, Microsoft Word, Excel, PowerPoint, Google Documents