UNIVERSITY OF CALIFORNIA SAN DIEGO

Doubly robust causal inference from complex surveys using matching, with applications to the causal effect of e-cigarette use on smoking cessation

A dissertation submitted in partial satisfaction of the
requirements for the degree
Doctor of Philosophy

in

Biostatistics

by

Ruifeng Chen

Committee in charge:

Professor Karen Messer, Chair
Professor Tarik Benmarhnia
Professor John P. Pierce
Professor Xin Tu
Professor Xinlian Zhang

2022

The dissertation of Ruifeng Chen is approved, and it is acceptable in quality and form for publication on microfilm and electronically.

University of California San Diego

2022

DEDICATION

To my entire family: for your true love and unconditional support through all the years.

EPIGRAPH

The best thing about being a statistician,

is that you get to play in everyone's backyard

—John Tukey

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGEMENTS

I wish to take this opportunity to thank one of the greatest parts, the wonderful 5-year journey, in my life in the division of Biostatistics at University of California, San Diego (UCSD). This part of my life not only teaches me how to explore biostatistical methods, how to apply biostatistical knowledge to real applications to solve real problems, but more importantly, it finally imparts me how to live and learn. I sincerely thank all those who I had the chance to connect with, collaborate with and learn from during the entire time at UCSD.

First of all, I want to express my deepest gratitude and admiration to my advisor, my committee chair, Dr. Karen Messer. I really appreciate all the support and inspiration from her during the past 5 years. Dr. Messer has led me to so many interesting collaboration projects in different areas and taught me how to be a great collaborator with people have different background. Through those opportunities, I finally could feel how big the motivation and happiness are exploring and applying the biostatistical methods to solve real problems. In addition, Dr. Messer has taught me not only how to play with Biostatistics in real problems, but also how to initiate ideas and motivations from real data to improve and further explore existing biostatistics methods, and finally, she has guided me to the interesting causal inference area. Dr. Messer has also been always supportive and respectful of the choices I made about my career development. She provided great support and suggestions during the time I was seeking for summer internships and the full-time job. I am so appreciated of the time, the energy and the patience she spent on me. Last but not the least, I want to say Dr. Messer is not just a wonderful professor and a great researcher, she is also one of the greatest leaders I have ever came across with, who is always working hard to make our whole program better and better.

I also want to give my sincere gratitude to Dr. John Pierce, the giant and the greatest collaborator in tobacco control. I truly appreciate that I could have the chance to work with him and many thanks for his trust funding me for completing the study during the past a few years. Dr. Pierce has led me to know how important and complex understanding the tobacco control world is

and how hard people are trying to make real change in this field. I really admire his inspiration and hardworking and I have learned a lot from his genius way generating those simple tables to understand the complex real data and real problems. In addition, Dr. Pierce is very generous. He is always devoting so hard to make our entire tobacco control team here at UCSD moving fast and smoothly. I sincerely appreciate the opportunity having such a great collaborator and leader.

In addition, I want to give many thanks to my other committee members, Dr. Xin Tu, Dr. Xinlian Zhang and Dr. Tarik Benmarhnia. I really thank Dr. Xin Tu for his sincerity and enthusiasm which made me feel very warm and comfortable here at UCSD. Also, Dr. Tu has taught me a comprehensive causal inference class and has given me a great opportunity working with him on a project exploring the use of functional response models in the two active treatments setting, which is very interesting and has extended my knowledge a lot. Dr. Xinlian Zhang is not only a great young professor of me, but also one of my best friends at UCSD. She is always super patient about any of my question and has helped me a lot on multiple of my projects as well as assisting me during the preparation of seeking a job. Dr. Tarik Benmarhnia is one of my earliest collaborators in tobacco control projects who has taught me thoroughly of how to do both the design and the programming of projects in this field. I also thank him a lot for providing an interesting reproducibility class with me and collaborating with me on the matching methods project, which has also given me the chance to do a better literature review to better understand the matching area.

It is really my pleasure to work at Moores Cancer Center (MCC) at UCSD besides my PhD study. I want to thank Dr. Emily Pittman, who has helped me a lot and worked closely with me on several important clinical trials. Emily is always the great backup of me whenever I have clinical trial questions. I want to also thank Dr. Loki Natarajan, Dr. Liu, Dr. JingJing Zou, Minya Pu, Jing Zhang and Dr. Brian Kwan. All of them are my great colleagues and teachers at MCC who would never hesitate sharing their knowledge and support. I appreciate to have such a great team. Moreover, I want to thank my collaborator Dr. Assuntina Sacco at MCC. It was my great honor to be involved in her phase II pembrolizumab plus cetuximab trial which has ended up being promoted as a standard of care. I also want to thank Dr. Joel Baumgartner, Dr. Joseph Califano, Dr. Sandip Patel, Dr. Barbara Parker, Dr. Rebecca Shatsky, Dr. Kathryn Gold, Dr. Fanny Chapelin, Dr. Brinda Rana and all people who have collaborated with me closely on those many different high-quality clinical trials and projects.

Further, I want to thank the tobacco control team that I am working with at MCC. Many thanks to Martha White, Dr. Eric Leas, Sheila Kealey, Dr. Matthew Stone, Dr. Dennis Trinidad, Dr. Sara McMenamin and Dr. David Strong. It is really fun having meeting with, working with and learning from these guys in the fun team. Thank Dr. Strong for teaching me an important factor analysis class and instructing me on a follow-up project. I have also learned a lot from the graphic labeling cigarette pack project working with our group and I am very happy to have the chance being involved in it.

I also want to thank my friends and peers. Especially thanks to Dr. Yuqi Qiu, Dr. Siyun Chen and their baby Alayna Qiu, Yiran Zhang, Zhou Tang, Wei Xu, Wenyi Lin, Tuo Lin, for all their great support in different aspects over the past years. Life would be another story at UCSD without the company of these good friends. Meanwhile, I want to thank Jinyuan Liu, Kristen Hansen, Matt Chen, Tylor Vu, Junting Ren, Serene Jiang and all the peer students, for supporting and motivating each other all the time. Last but not the least, I want to thank my middle school classmate Shitong Zhu and my undergraduate classmate Zhiyong Hu, for their big support and communication during the most difficult times over the past a few years.

Finally, I want to thank my great family, my mom Siyao Du, my dad Jun Chen, my girlfriend Hsiao Yuan Wang, my grandma Mingxin Liu, my aunts Qiaoling Du and Qiongfang Du, and the entire big family of mine, for being my hardest backup over the past 5 years. I would never have been able to make it to the end without their support and company.

Chapter 2, in full, has been published and may be found as ”Chen, Ruifeng and Pierce, John P. and Leas, Eric C. and White, Martha M. and Kealey, Sheila and Strong, David R. and

Trinidad, Dennis R. and Benmarhnia, Tarik and Messer, Karen. Use of Electronic Cigarettes to Aid Long-Term Smoking Cessation in the United States: Prospective Evidence From the PATH Cohort Study, American Journal of Epidemiology, 189 (12), 1529-1537, 2020". The dissertation author was the primary author on this paper.

Chapter 3, in full, has been prepared for submission for publication as "Chen, Ruifeng; Messer, Karen. Doubly robust causal inference using the doubly-matched estimator, with application to the causal effect of e-cigarette use on smoking cessation and cigarette consumption reduction". The dissertation author was the primary author on this paper.

Chapter 4, in full, has been prepared for submission for publication as "Chen, Ruifeng; Messer, Karen. Large sample properties of the variance estimation of matching estimators in complex survey". The dissertation author was the primary author on this paper.

VITA

2016	Bachelor of Science in Statistics. Shandong University, China
2017	Master of Science in Statistics. University of Wisconsin-Madison, U.S
2017-2022	Graduate Student Researcher. University of California San Diego, U.S
2022	Doctor of Philosophy in Biostatistics. University of California San Diego, U.S

PUBLICATIONS

R. Chen, J.P. Pierce, E.C. Leas, T. Benmarhnia, D.R. Strong, M.M. White, M.D. Stone, D.R. Trinidad, S.B. McMenamin, K. Messer. *Effectiveness of e-cigarettes as aids for smoking cessation: evidence from the PATH Study cohort, 2017–2019* Tobacco Control, Published Online First: 2022.

R. Chen, J.P. Pierce, E.C. Leas, M.M. White, S. Kealey, D.R. Strong, D.R. Trinidad, T. Benmarhnia, K. Messer. *Use of Electronic Cigarettes to Aid Long-Term Smoking Cessation in the United States: Prospective Evidence From the PATH Cohort Study* American Journal of Epidemiology, 189 (12), 1529–1537, 2020.

J.P. Pierce, **R Chen**, S. Kealey, E.C. Leas, M.M. White, M.D. Stone, S.B. McMenamin, D.R. Trinidad, D.R. Strong, T. Benmarhnia, K. Messer. *Incidence of Cigarette Smoking Relapse Among Individuals Who Switched to e-Cigarettes or Other Tobacco Products* JAMA network open, 4 (10), e2128810, 2021.

J.P. Pierce, **R. Chen**, E.C. Leas, M.M. White, S. Kealey, M.D. Stone, T. Benmarhnia, D.R. Trinidad, D.R. Strong, K. Messer. *Use of E-cigarettes and Other Tobacco Products and Progression to Daily Cigarette Smoking* Pediatrics, 147 (2), e2020025122, 2021.

J.P. Pierce, E.C. Leas, T. Benmarhnia, S.B. McMenamin, D.R. Strong, **R. Chen**, K. Messer. *E-cigarettes and cessation: the introduction of substantial bias in analyses of PATH Study* Nicotine & Tobacco Research, 23 (5), 876-877, 2021.

ABSTRACT OF THE DISSERTATION

Doubly robust causal inference from complex surveys using matching, with applications to the causal effect of e-cigarette use on smoking cessation

by

Ruifeng Chen

Doctor of Philosophy in Biostatistics

University of California San Diego, 2022

Professor Karen Messer, Chair

Propensity score matching (PSM) is a statistical technique which is widely used in multiple disciplines to make causal inference. In this dissertation, we aim to explore a doubly robust matching method which improves PSM in certain circumstances. Moreover, we extend matching techniques to the setting of complex surveys, and investigate how to estimate the variance of the matching estimator in this setting. We apply these methodological investigations to data from the Population Assessment of Tobacco and Health (PATH) survey, and assess their performance in a real study.

The dissertation comprises three studies. In the first study, the main objective is to investigate whether the use of electronic cigarettes (e-cigarettes) aids long-term cigarette/ nicotine cessation among adult U.S. smokers who want to quit cigarette smoking, using data from the PATH survey. Caliper nearest neighbor PSM is conducted to match each e-cigarette user to one or more e-cigarette non-users. The weighted difference of cessation rates for cigarettes / nicotine between the two groups is calculated among the matched pairs, and the bootstrap is used to assess statistical significance . We find that e-cigarettes may not be an effective cessation aid for adult smokers in the US and, instead, may contribute to continuing nicotine dependence.

PSM depends strongly on the correctness of the propensity score (PS). The first study motivates us to explore whether we can improve the existing PSM method, in the case when the PS is incorrect. Hence, in the second study, we propose and study a doubly-matched (DMT) estimator, which matches simultaneously on both the PS and another 'balancing score', the prognostic score (PGS), to make doubly robust (DR) causal inference. Our simulation study demonstrates that the estimator is doubly robust; that is, even if either the PS or the PGS is incorrect, the DMT estimator performs well as long as the other is correct. Furthermore, under the simple random sampling design, the full bootstrap method of variance estimation, which resamples individuals, re-estimates the PS, and re-conducts PSM, works well although it is sometimes conservative. Finally, we apply the DMT estimator to the question of interest in the first study and end up with a consistent conclusion that e-cigarettes may not be effective for assisting cigarette smoking cessation.

In the third study, taking the PSM estimator as an example we explore approaches to variance estimates of the matching estimator which are appropriate for complex surveys with survey weights. Such complex surveys typically have a hierarchical sampling design, in which subjects are sampled within clusters, which are sampled within strata. We prove the large sample consistency of the jackknife estimate of variance and the balanced repeated replicate (BRR) estimate of variance in the case when both the number of sampling strata and the number of subjects within sampled clusters go to infinity. Simulation studies demonstrate that BRR and Fay's method, which is an adjusted BRR method commonly used in large population surveys, indeed outperform other commonly-used bootstrap methods. We also apply these variance estimates to PATH study data to investigate whether using counseling or self-help materials helps adult smokers reduce cigarette consumption in the long-term, and we end up with a negative conclusion. This study fills the gap in the variance estimation of the PSM/ general matching estimator in complex surveys.

Chapter 1

Introduction

E-cigarette use has become popular in recent years, with the sales of e-cigarette doubling in the US between 2015 and 2017. [37] E-cigarettes are now one of the most popular tobacco products . [7, 30, 53] However, in the US, e-cigarettes can deliver high doses of nicotine and experts have noted potential public health risks, including the potential for increased smoking initiation among minors, and for increased nicotine addiction among dual users of cigarettes and e-cigarettes. [63] Given these known risks, many arguments for a net public health benefit rely on the effectiveness of e-cigarettes in helping adult smokers to quit cigarette smoking for the long-term. [27, 100] Further studies are needed to quantify the size of any population-level benefit from smoking cessation using e-cigarettes. [86]

The effectiveness of e-cigarettes for smoking cessation is a causal question of great public health importance. Randomized studies are considered as the gold standard to provide causal evidence. However, these are not always feasible and can be costly in terms of time, money and effort. In addition, the population enrolled in randomized trials may not be representative of the total population. Thus observational studies also play a critical role in the tobacco control field. A causal model using a potential outcome framework and suitable for observational data was proposed by Rubin in 1974 [78]. Rosenbaum and Rubin [74] introduced the propensity score (PS), based on Rubin's potential outcome causal inference framework. The PS is defined as the probability of being assigned to the treatment group conditional on baseline covariates, assuming no unmeasured confounders. Since then, different PS-related methods have been developed for causal inference, including propensity score matching (PSM) [74, 39], propensity score stratification, [75] inverse probability weighting (IPW) [72] and others.

PSM is an ideal approach to investigate the effectiveness of e-cigarettes in our context. It is widely used to make causal inference when we have a target sub-population of interest, for example for estimating the average causal effect among the treated (ACET). In our circumstance, we are interested in estimating the effect among those who used e-cigarettes, and PSM can help avoid extrapolating to those who never used e-cigarettes. Compared to other approaches such as regression modeling, matching methods have the benefit of easy interpretation and simple visualization of the matching results. PSM is widely used in multiple disciplines including statistics, economics, epidemiology, sociology, and others [39, 79, 88, 17, 62]. In the PSM algorithm, we first estimate the PS for each subject, conduct the matching using this estimated PS, and then estimate the treatment effect among the matched pairs.

However, consistency of the PSM estimator of effect is very dependent on the consistency of estimated propensity score. In practice, unmeasured confounders and incorrectly specified functional models may lead to incorrect estimation of the PS, which may bias the resulting estimate of causal effect. A doubly robust (DR) approach was proposed and developed by Robins in 1994. [72] It combines an inverse probability weighting (IPW) estimator based on the PS with a regression estimator of the treatment effect to estimate the average causal effect among the entire population (ACE), and it has been shown to be consistent when either the PS model or the regression model is correctly specified. It is also called the augmented inverse probability weighting (AIPW) approach. Other DR approaches have been developed based on the AIPW form, mainly with the goal of variance reduction under certain circumstances. For example, Cao [18] and Tan [92] introduced improved DR methods such that when the propensity score model is correctly specified, these improved DR estimators are guaranteed to be as efficient as the IPW estimator. Vermeulen and Vansteelandt proposed a bias-reduced DR estimator, which is more robust to mild misspecification of the two models. [98] Doubly robust estimators are also proposed to estimate the ACET. [85, 93]

In addition to these well-known DR estimators, which combine an IPW estimator and a regression estimator, we think approaches for DR estimation using matching only are also needed, due to the benefits of the matching estimator such as easy interpretation and robustness to extrapolation. Hansen proposed the prognostic score (PGS), defined as the expected value of the potential outcome of a control subject given its covariates. [33] Similar to the PS, the PGS is also a balancing score, conditional on which the distributions of the covariates are the same for the treated subjects and the controlled subjects. We review [51] and propose the doubly-matched (DM) estimator; that is, we match treated subjects to control subjects based on both the PS and the PGS simultaneously, to achieve double robustness in estimating the ACET. Details are discussed in chapter three.

So far, we implicitly consider how to estimate the effect with a simple random sampling (SRS) design, that is, every subject in the population has an exactly equal chance of being selected. However, as mentioned earlier, large-scale population data is often needed to study the complex problems in tobacco control, and such data are often sampled from a more complex sampling design. Stratified sampling and cluster sampling [96] are commonly-used sampling designs for large-scale population surveys. Compared to SRS, stratified sampling can assure better representation of selected subgroups in a large-scale target population. It can also be more convenient to conduct and administer in practice, and it can be designed to reduce sampling variability as it reduces the within-strata variation. On the other hand, cluster sampling, such as sampling schools or households, can have the benefit of making the sampling process easier and less costly. [57] A large-scale complex survey often assembles all of SRS, stratified sampling and cluster sampling together to make the sampling design. [57] Causal inference made from such a complex survey design needs to account for the survey weights, which are attached to each subject in the sample and are used to obtain estimates of population parameters of interest (for example to account for the oversampling design in PATH study), as well as cluster based sampling and other survey design elements into consideration. [57, 44, 91, 56]

PSM is increasingly used in survey studies to make causal inference. [66, 20, 19] Theoretical methods have also been developed in this area, mainly focused on establishing the consistency of point estimates of the causal effect. Austin et al. have constructed a comprehensive simulation study to investigate the PSM estimator with survey weights. [11] Lenis et al. have demonstrated that weights are not needed in estimating the PS, and the controls included in matched pairs should use the weights of the matched treated subjects in follow-up analyses. [52] In addition, they have shown that the weights must be added in the outcome analysis after matching to make consistent estimates. They have proved their theoretical fundamentals following Austin's simulation setup. However, to our knowledge, there is no theoretical study investigating the variance estimation of the matching estimator in complex surveys, although the variance estimation of the matching method, especially the PSM, has been well discussed under the SRS design. [81, 4, 71, 80, 34, 12, 35, 5, 1, 3] We address gap in chapter 4 in which we discuss variance estimation of the matching estimator in complex surveys. We consider commonly-used approaches including the jackknife method, the balanced repeated replicate (BRR) method, Fay's method and the bootstrap. The large sample properties of the jackknife estimate and the BRR estimate are discussed, followed by a comprehensive simulation study and a case study to compare these estimates.

1.1 The organization of this dissertation

In chapter two, we present a recent study in tobacco control in which we assess whether the use of e-cigarettes helps long-term cigarette/ nicotine cessation among US smokers, a prospective study using the PATH Wave 2 (W2) to W4 survey data. We use caliper nearest neighbor PS matching to match non-e-cigarette users to e-cigarette users to assess the weighted difference of the follow-up cessation rates among the matched pairs. Bootstrap quantile confidence intervals are used to demonstrate whether the weighted difference is statistically significant. In the end, we find that e-cigarettes may not be an effective cessation aid for adult smokers. More importantly, e-cigarettes may contribute to continuing nicotine dependence. This study motivates us to further explore how to improve the PSM when the PS is incorrect in chapter three and how to estimate the variance of the matching estimator incorporating survey weights and the complex survey desing, in chapter four.

Chapter three presents our study of how to improve the PSM when the PS is incorrectly estimated. We construct and explicitly show the form of the DM estimator to make DR causal inference under the SRS design. The DM estimator matches on both the PS and the PGS and is consistent in estimating the ACET when either of the two scores is correctly estimated. The performance of the DM estimator is demonstrated in a simulation study. Further, approaches for estimating confidence intervals of the DM estimator, including the parametric approach and bootstrap approaches, are assessed. Simulation studies show that the full bootstrap method, which bootstrap resamples individuals and re-estimates the PS and re-conducts PSM, works consistently well although it is sometimes conservative, and is suggested to use in practice. Finally, the use of the DM estimator is demonstrated in a case study investigating whether the use of e-cigarette helps decrease smoking cessation rate and cigarette consumption respectively among US smokers.

In chapter four, we explore several variance estimation methods for the matching estimator, taking the PSM estimator as an example, in complex surveys. We show the asymptotic consistency of the jackknife estimate and the BRR estimate. These two estimates are equivalent when the estimator is a linear transformation of the outcomes and only 2 clusters are sampled in all strata. The consistency can be generalized to Fay's estimate, which is an improvement of the BRR estimate. The performance of the BRR and Fay's method are investigated in a simulation study, compared to the full bootstrap which bootstraps both clusters and subjects in a two step way, re-estimating the PS and re-conducting the PSM, as well as the conditional bootstrap which bootstraps the matched pairs directly from the original sample and neither the PS is re-estimated nor the PSM is re-conducted. Results show that both the BRR estimate and the Fay's estimate consistently work well, and the Fay's estimate performs the best with smaller variance compared to the BRR estimate. In addition, all 4 variance estimates are applied in a tobacco control case study using a complex survey, where we end up with the conclusion that using counseling or self-help materials among smokers who want to quit cigarette smoking doesn't help them reduce cigarette consumption in the long-term.

Chapter five finally summarizes the overall findings and discusses the future studies.

Chapter 2

Use of electronic cigarettes to aid long-term smoking cessation in the United States: prospective evidence from the PATH cohort study

2.1 Abstract

Electronic cigarettes (e-cigarettes) are the preferred smoking-cessation aid in the United States; however, there is little evidence regarding long-term effectiveness among those who use them. In this chapter, we used the Population Assessment of Tobacco and Health Study to compare long-term abstinence between matched US smokers who tried to quit with and without use of e-cigarettes as a cessation aid. We identified a nationally representative cohort of 2,535 adult US smokers in 2014–2015 (baseline assessment), who, in 2015–2016 (exposure assessment), reported a past-year attempt to quit and the cessation aids used, and reported smoking status in 2016–2017 (outcome assessment; self-reported $\geq$ 12 months continuous abstinence). We used propensity-score methods to match each e-cigarette user with similar nonusers. Among US smokers who used e-cigarettes to help quit, 12.9% (95% confidence interval (CI): 9.1%, 16.7%) successfully attained long-term abstinence. However, there was no difference compared with matched non–e-cigarette users (cigarette abstinence difference: 2%; 95% CI: 3%, 7%). Furthermore, fewer e-cigarette users were abstinent from nicotine products in the long term (nicotine abstinence difference: 4%; 95% CI: 7%, 1%); approximately two-thirds of e-cigarette users who successfully quit smoking continued to use e-cigarettes. These results suggest e-cigarettes may not be an effective cessation aid for adult smokers and, instead, may contribute to continuing nicotine dependence.

2.2 Introduction

Electronic cigarette (e-cigarette) sales doubled in the United States between 2015 and 2017. [37] In the United Kingdom and the United States, e-cigarettes are now the most popular product type used to aid smoking cessation, ahead of US Food and Drug Administration–approved products including nicotine replacement therapy (NRT) such as a nicotine patch or nicotine gum, and prescription medications, including buproprion and varenicline. Although many herald e-cigarettes as a harm-reduction device, [7, 53, 30] experts have noted potential public health risks, including the potential for increased smoking initiation among minors, and for increased nicotine addiction among dual users of cigarettes and e-cigarettes. [63] In the United States, e-cigarettes can deliver high doses of nicotine, and there is evidence of substantial uptake among nonsmoking minors. [60] Given these known risks, arguments for a net public health benefit rely on the effectiveness of e-cigarettes in helping adult smokers quit cigarette smoking for the long term. [27, 100]

Several national reports have considered the evidence on whether e-cigarettes increase long-term smoking cessation. [63, 59] The recent US Surgeon General's report [65] concluded that evidence remains inadequate to infer that e-cigarettes increase smoking cessation. Only

4 randomized trials, all conducted outside of the United States, have directly tested whether e-cigarettes are efficacious for smoking cessation with a follow-up of at least 6 months. The most promising of these randomly assigned attendees of UK National Health Service stop-smoking services (n = 866) to either e-cigarettes or NRT and reported that use of e-cigarettes as a cessation aid increased successful quitting 1 year later. [31] However, the importance of motivation was highlighted by a pragmatic trial conducted at 54 US businesses, which randomized 6,004 employees who smoked to provision of either free FDA approved cessation aids or to free e-cigarettes, as well as to 3 other arms. All arms received a brief communication intervention. As part of the primary analysis, the trial reported that provision of free e-cigarettes did not increase cessation as compared to provision of free FDA approved cessation aids. [32] There also have been several reports from nationally representative longitudinal studies in which smokers self-selected to use e-cigarettes to help quit smoking. Use of e-cigarettes for quitting in the Adult Tobacco Cohort was associated with short-term but not long-term cigarette abstinence. [90] There have been 5 reports using data from the US Population Assessment of Tobacco and Health (PATH) Study. [38] Two analyses [14, 41] had biased results because they included smokers who did not make a quit attempt only in the comparison group. [67] Authors of 1 of the analyses reported that use of e- cigarettes to quit was associated with increased short-term abstinence, measured at the same time e-cigarette use was assessed. [13] In the other analysis, [101] authors reported that substitution of e-cigarettes for cigarettes at wave 2 was not associated with sustained abstinence at wave 3, confirming an earlier report [21] that use of e-cigarettes after quitting was associated with increased relapse to smoking 1 year later. The authors of the latter 2 studies suggest nicotine abstinence after quitting cigarettes may be an important moderator of long-term abstinence from cigarette smoking.

In this chapter, we use more recent PATH data to address whether use of e-cigarettes to aid quitting contributed to increased successful smoking cessation in the US population (self-reported continuous abstinence of at least 12 months [28]). Many smokers use multiple cessation aids,

[43] thus, we focused on any e-cigarette use for quitting compared with no use. Furthermore, we include as a second comparison group those who used an approved pharmaceutical aid to quit but not an e-cigarette. The population of smokers who use e-cigarettes to quit is appreciably different from those who do not. [13] Thus, we identified a priori 24 potential confounders and used PS methods to match each e-cigarette user with up to 2 closely matched control respondents. We compared population-weighted abstinence rates in the matched samples. This approach estimates the causal effect of e-cigarette use explicitly among those who choose to use them as a cessation aid and is less dependent on modeling assumptions than regression-based approaches, which estimate average effects projected to the entire population. [9] However, we report regression-based approaches as sensitivity analyses.

2.3 Methods

2.3.1 Data source and sample

Data were from the restricted public use file of the PATH Study. [97] The surveys are conducted at approximately annual intervals (waves) with stratified oversampling for adult (aged 18 to 24 years) tobacco users, and Black adults. Response rates were as follows: initial household screener survey, 54%; in-depth adult interview at wave 1, 74.0%; annual follow-up, 83.1%, 78.4%, and 73.5% for waves 2, 3, and 4, respectively. Surveys included informed consent and the study is overseen by the Westat Institutional Review Board. Our sample was identified from 10,722 cigarette smokers at wave 2 (2014–2015, baseline assessment), of whom 2,852 reported a past-year quit attempt at wave 3 (2015–2016, exposure assessment), with 2,535 completing the wave 4 outcome assessment in 2017–2018. The data collection schema is provided in Figure A.1.

2.3.2 Measures

Tobacco and nicotine use. During each interview, after viewing an image of each tobacco product, participants were asked whether they had ever used that product and whether they currently used it every day or some days. Noncurrent users were asked ”In the past 30 days, have you smoked/used (product), even one or two puffs” and “In the past 12 months, have you smoked/used (product), even one or two puffs.” Ever-smokers were asked whether they had used the following NRT products in the past 12 months: a nicotine patch, gum, inhaler, nasal spray, lozenge, or pill. Our 2 outcome variables ($\geq$ 12 months’ abstinence from 1) cigarettes and 2) all nicotine products) were identified from these questions on the wave 4 survey. Nicotine use includes any use of cigarettes, e-cigarettes, NRT, cigars (traditional, cigarillo, and filtered), pipes, hookah, snus, or other smokeless products.

Use of e-cigarettes and pharmaceutical aids to quit smoking. Each survey asked smokers whether they had made a quit attempt within the past 12 months and which of the following products was used for their most recent quit attempt: e-cigarettes, NRT, varenicline (Chantix; Pfizer, Groton, Connecticut), or buproprion (Wellbutrin or Zyban; GlaxoSmithKline, London, UK). The primary exposure was reported use of e-cigarettes to quit at wave 3 (e-cigarette group, n = 427); comparison groups were those who did not use e-cigarettes to quit smoking (no–e-cigarette group, n = 2,108) and those who reported use of a pharmaceutical cessation aid at wave 3 (varenicline, buproprion, or NRT) but not e-cigarettes (n = 465).

Study covariates. A.1 presents details of 24 potential confounders, which we identified a priori. These include sociodemographic variables, cigarette-smoking history, duration of previous quit attempt reported prior to baseline, timing of most recent quit attempt from survey (assessed at wave 3); self-efficacy about quitting; interest in quitting cigarettes; exposure to smoking; perceived harm of cigarettes and e-cigarettes; daily e-cigarette use reported at current or prior surveys (”ever” daily use); nicotine dependence level (average agreement with a series of 15 statements on emotional and physical response to nicotine products, scaled from 0 to 100); [87]

and health-related covariates. All were assessed at wave 2, with the exception of timing of most recent quit attempt from the wave 3 survey, used to control potential recall bias associated with type of aid used. [16] Univariate distributions by cessation aid category are listed in Table A.1.

2.3.3 Statistical analyses

Estimates were weighted using the wave 1 through wave 4 longitudinal survey weights, which were adjusted for the sampling design, survey nonresponse, and longitudinal drop out. [69] Weighted percentages and Wilson confidence intervals for proportions were calculated. For confidence intervals and P values, the replicate survey weights were used with balanced repeated replication with Fay's adjustment ($\rho = 0.3$) [40] in R, version 3.5.3 (R Foundation for Statistical Computing, Vienna, Austria), except for the PS matched analyses, where bootstrap percentile confidence intervals were used.

For PSM, within each bootstrap sample for each participant we calculated a PS by estimating the probability of membership in the e-cigarette–use group using logistic regression. To obtain complete data for the 24 covariates, we used simple imputation (R package "mice"). To identify the optimal set of covariates among the 24 variables, we used a 10-fold cross-validated LASSO [95] procedure for each logistic regression model, with a tuning parameter selected from among the sequence 0–0.1 with step of 0.005 (R package "glmnet"), conducted without survey weights. We repeated this PS estimate for each bootstrap-resampled data set. Using the propensity score, we matched up to 2 controls for each case (nearest-neighbor matches using R package "Matchit") within the a priori caliper distance of 0.1 (if possible) or 0.2 (maximum allowed). [99] We chose the caliper that provided the lowest standardized difference averaged across all covariates after matching. Cases that did not have a match were omitted from the sample.

For each matched bootstrap sample, we used logistic regression with survey weights (R package "survey") to estimate the average risk difference between the 2 matched groups, for each outcome. The model included an indicator of the matched pair (or triple), the overall propensity score, and, to adjust for any remaining covariate imbalance, any covariate with median standardized difference between the 2 study groups larger than 0.10. [68, 35] We report the bootstrap mean estimate of risk difference and adjusted 95% bootstrap quantile confidence intervals, with Bonferroni adjustment [24] to account for the 2 abstinence outcomes studied. To identify a sufficient bootstrap sample size, we required a jackknife quality estimate [26] to be less than 0.1, resulting in 1,500 bootstrap samples for the comparison of e-cigarette use versus no e-cigarette use.

2.3.4 Sensitivity analyses

Sensitivity analyses included incorporating matching as random effects instead of fixed effects, and 1:1 rather than 1:2 PSM without additional covariate adjustment. We also used weighted multivariable logistic regression on the full sample; covariates included were age, sex, ethnicity, race, education, income, nicotine dependence, relative perceived harm of e-cigarettes, and previous daily e-cigarettes use, with simple imputation. Finally, we tested whether the results were robust to omission of adjustment for multiple comparisons.

As a post hoc exploratory sensitivity analyses, we used logistic regression to test whether the association of e-cigarette use with long-term cigarette abstinence and nicotine abstinence differed by baseline smoking status, nicotine dependence, age, sex, education level, and race/ethnicity. Statistical inference was based on 95% confidence intervals for interaction terms (uncorrected for multiple comparisons), and a stratified analysis was conducted when the boundary of the confidence interval was close to 1.

2.4 Results

2.4.1 Population rates of cigarette and nicotine abstinence at wave 4 follow-up

Among this representative sample of US smokers who reported a past-year quit attempt in 2015–2016 (wave 3), 17.4% used e-cigarettes to help quit smoking. Those who used e-cigarettes were younger, more nicotine dependent, more likely to be non-Hispanic White, and had higher income and level of education (Table 2.1).

Among US smokers who used e-cigarettes to quit, 12.9% (95% confidence interval (CI): 9.1%, 16.7%) achieved at least 12 months' abstinence from cigarettes at wave 4, compared with 11.3% (95% CI: 9.6, 13.0) among US smokers who did not use e-cigarettes to quit (Table 2.2). Among US smokers who used e-cigarettes to quit, the population-weighted estimate of at least 12 months of nicotine abstinence at wave 4 was 2.8% (95% CI: 0.9%, 4.8%), compared with 8.1% (95% CI: 6.5%, 9.7%) among those who did not use e-cigarettes to quit. Table 2.2 lists population abstinence rates among these smokers by baseline consumption level (daily or nondaily).

2.4.2 Propensity score–matched samples

We assessed appropriateness of the PSM by comparing kernel density estimates of the PS (i.e., the estimated probability of using e-cigarettes to quit on the index quit attempt). Comparing smokers who used e-cigarettes to quit and smokers who did not, the 2 density estimates were very different prior to matching (Figure A.2). In particular, there were few respondents with propensities greater than 0.6 in the no–e-cigarette population, indicating that some population subgroups are unlikely to use e-cigarettes. Matching, although restricted to respondents with propensity score less than 0.8, resulted in good overlap of the density estimates (Figure A.2). Matching used all the 427 available e-cigarette users, with a median sample size of 386 for the

Table 2.1: Sample characteristics of smokers[a] in 2014–2015 reporting a past-year quit attempt in 2015–2016, according to use or no use, of e-cigarettes to aid quitting, population assessment of tobacco and health study.

Sociodemographics	Used e-Cigarettes on Quit Attempt[b] (n = 427)			Did Not Use e-Cigarettes on Quit Attempt[b] (n = 2,108)			P Value
	No.	Weighted %	95% CI	No.	Weighted %	95% CI	
Age, years							< 0.001
18-34	218	46.8	41.1, 52.5	922	38.3	35.6, 41.0	
35-50	127	32.0	26.7, 37.3	546	28.5	26.0, 31.0	
≥ 50	82	21.2	16.9, 25.5	640	33.2	30.5, 35.9	
Sex							0.500
Male	202	50.7	44.8, 56.6	1,012	53.0	50.5, 55.5	
Female	225	49.3	43.4, 55.2	1,095	47.0	44.5, 49.5	
Education							0.006
Less than high school	89	19.4	14.9, 23.9	593	26.8	24.6, 29.0	
High school graduate	90	23.1	17.2, 29.0	502	27.5	25.0, 30.0	
Some college or higher	230	55.2	48.9, 61.5	944	43.7	41.2, 46.2	
Ethnicity							< 0.001
Hispanic	37	6.9	4.7, 9.1	334	15.1	13.1, 17.1	
Non-Hispanic	390	93.1	90.9, 95.3	1,732	82.9	80.7, 85.1	
Race							< 0.001
White	354	85.8	82.5, 89.1	1,400	69.1	66.6, 71.6	
Black	26	5.5	3.3, 7.7	433	19.5	17.5, 21.5	
Other	43	8.0	5.1, 10.9	223	9.0	7.6, 10.4	
Income (annually, US$)							< 0.001
$< 35,000$	220	47.5	42.2, 52.8	1,341	59.8	56.9, 62.7	
$\geq 35,000$	190	48.0	42.7, 53.3	633	34.0	31.3, 36.7	
Smoking-related diseases							0.178
Marked	201	47.0	41.7, 52.3	1,069	51.1	48.4, 53.8	
Not marked	226	53.0	47.7, 58.3	1,039	48.9	46.2, 51.6	
Nicotine dependence scale score							0.009
0–33.3	89	22.4	17.1, 27.7	571	28.5	25.6, 31.4	
33.4-66.7	172	38.6	33.9, 43.3	839	39.4	36.9, 41.9	
66.8–100	165	38.7	33.4, 44.0	648	29.6	27.4, 31.8	
Relative perceived harm of e-cigarettes							< 0.001
Less harmful	262	61.3	56.4, 66.2	726	34.7	32.3, 37.1	
More harmful	158	36.9	31.8, 42.0	1,306	61.4	58.9, 63.9	
e-Cigarette use before W2							0.001
Never	44	10.9	7.4, 14.4	949	48.1	45.9, 50.3	
Ever	383	89.1	85.6, 92.6	1,154	51.7	49.5, 53.9	
Daily e-Cigarette use before							< 0.001
Daily use at W1 or W2	106	24.5	20.0, 29.0	96	4.3	3.3, 5.3	
Not daily use at W1 or W2	321	75.5	71.0, 80.0	2,012	95.7	94.7, 96.7	

Abbreviations: W1, wave 1 of the study; W2, wave 2 of the study.

[a]. Weighted US population estimates.

[b]. e-Cigarette use for most recent quit attempt, among all those reporting quit attempts at wave 3.

Table 2.2: Long-term abstinence at follow-up[a,b] among US smokers who made a quit attempt in 2015–2016, according to use or no use, of e-cigarettes to aid quitting, population assessment of tobacco and health study.

Cigarette Smoking Status (W2) and e-Cigarettes Used to Aid		Cigarette Abstinence (W4)			Nicotine[d] Abstinence (W4)	
Quit Attempt?[c]	No.	Weighted % Abstinent	95% CI	No.	Weighted % Abstinent	95%
All current cigarette smokers						
Yes	427	12.9	9.1, 16.7	427	2.8	0.9, 4.8
No	2,108	11.3	9.6, 13.0	2,108	8.1	6.5, 9.7
Daily cigarette smokers						
Yes	290	13.7	8.8, 18.7	290	3.4	0.8, 6.1
No	1,455	9.5	7.7, 11.3	1,455	7.3	5.7, 9.0
Nondaily cigarette smokers						
Yes	137	11.1	5.7, 16.5	137	1.5	0.6, 3.7
No	653	15.1	11.7, 18.4	653	9.6	6.6, 12.6

Abbreviations: W2, wave 2 of the study; W4, wave 4 of the study.

[a]. Abstinence of $\geq$ 12 months, reported at wave 4.

[b]. Weighted US population estimates.

[c]. e-Cigarette use for most recent quit attempt, among all those reporting quit attempts at wave 3.

[d]. Nicotine use includes any of cigarettes, e-cigarettes, and nicotine replacement therapy.

matched sample. For each matching variable, we also plotted the standardized absolute mean difference between study groups across the 1,500 bootstrap re-samples, for the full sample and the matched samples (Figure A.3). The matched samples had a small between-group difference across all covariates with the exception of prior daily e-cigarette use. This variable was controlled for in the logistic regression comparing abstinence rates between the matched samples.

Figures assessing the quality of the match between the e-cigarette group and the matched US Food and Drug Administration–approved pharmaceutical aid group are presented in Figure A.4 and Figure A.5. The propensity scores were always positive, indicating that some respondents in each group were at least somewhat likely to belong to the other group. However, the fewer available respondents in the comparison group resulted in fewer successful matches: all 427 e-cigarette users were included in at least 1 matched sample but the median matched sample size was 244. We again used 1,500 bootstrap samples and the matching was improved in the between-group balance for all covariates. However, a residual difference remained for age, prior daily e-cigarette use, relative perceived harm of e-cigarettes, and smoking-related diseases, which we controlled for in the logistic regression.

2.4.3 Comparisons of abstinence rates between matched samples

There was no evidence for a difference in the proportion of persons who achieved long-term abstinence from cigarettes between those who used e-cigarettes to help quit smoking and the matched sample of those who did not use e-cigarettes as a cessation aid (risk difference: 0.02, 95% CI: 0.03, 0.07) (Figure 2.1). However, e-cigarettes users were less likely to be nicotine abstinent in the long term at follow up (risk difference: –0.04, 95% CI: 0.07, 0.01).

Comparing e-cigarette users with the matched sample of those who used pharmaceutical aids (but not e-cigarettes) to quit (Figure 2.1), there was no difference in the proportion who achieved either abstinence outcome (cigarette abstinence: risk difference: 0.02, 95% CI: 0.03, 0.08; nicotine abstinence: risk difference: –0.03, 95% CI: 0.07, 0.01).

Sensitivity analyses were consistent with these results (Figure A.6, Table A.2, A.2, A.3 and A.4). Exploratory analyses of interaction terms between e-cigarette use and baseline smoking status, nicotine dependence, age, sex, education level, and race/ethnicity revealed that all confidence intervals included 1, unadjusted for multiple comparisons (A.5). However, the interaction terms for the association of e-cigarettes with daily or nondaily smoking status, and with educational level appeared to be worth future exploration, and stratified analyses for these variables are presented in A.5.

2.4.4 US abstinence rates by product among those who successfully quit cigarettes

Table 2.3 lists population abstinence rates from various nicotine-containing products among all those who were long-term abstinent from cigarettes at wave 4. Among those who successfully used e-cigarettes to quit cigarette smoking, only approximately one-third were also

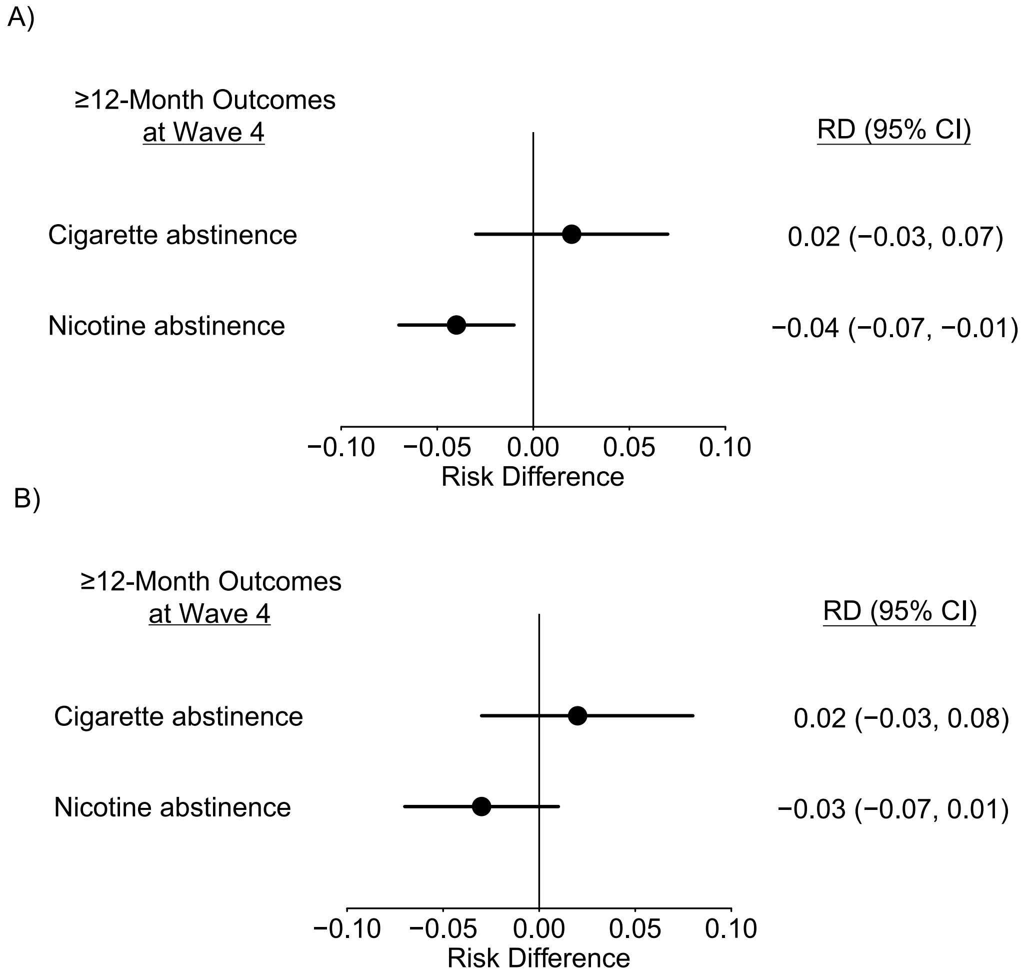


Figure 2.1: Differences in long-term abstinence rates from smoking cigarettes and from use of any nicotine-containing product, comparing the type of aid used for smoking cessation, 2014–2108, Population Assessment of Tobacco and Health (PATH) Study.

Note: A) e-Cigarettes used for cessation versus no e-cigarettes used for cessation. B) e-Cigarettes used for cessation versus pharmacotherapy but no e-cigarettes used for cessation. Weighted differences in rates of ≥ 12 months' abstinence between e-cigarette users and a matched sample of non–e-cigarette users, matched on 26 smoking-related characteristics and further adjusted by logistic regression. Bars represent Bonferroni adjusted 95% bootstrap confidence intervals (CI). Samples were drawn from 2,852 adult respondents to the PATH Study who reported smoking at wave 2 (2014–2015), reported a quit attempt and cessation aids used at wave 3 (2015–2016), and reported abstinence outcomes at wave 4 (2017–2018). RD, risk difference.

Table 2.3: Long-term abstinence[a,b] (≥ 12 months) from e-cigarettes, nicotine replacement therapy, and other tobacco products[c] among US smokers who were ≥ 12 months' cigarette abstinent at follow-up in 2016–2017, population assessment of tobacco and health study.

Products Abstained from for ≥ 12-Months at W4	e-Cigarettes Used To Quit[d] (n=49)		e-Cigarettes Not Used To Quit[d] (n=227)		Pharmaceutical Aid[e] Used To Quit[d] (n=45)	
	Weighted % Abstinent	95% CI	Weighted % Abstinent	95% CI	Weighted % Abstinent	95% CI
E-Cigarettes	31.7	16.4, 47.0	93.0	89.0, 96.9	96.1	89.7, 102.4
NRT	94.5	85.3, 103.8	91.9	87.4, 96.3	71.0	55.7, 86.4
Other tobacco products	82.2	70.0, 94.5	82.9	77.3, 88.5	93.1	85.2, 101.1
Combustible[f]	83.0	70.7, 95.2	86.1	80.5, 91.6	93.1	85.2, 101.1
Smokeless[g]	93.3	84.0, 102.6	95.6	92.6, 98.6	97.2	91.5, 102.9

Abbreviations: W4, wave 4 of the study; NRT, nicotine replacement therapy.

[a]. Abstinence of ≥ 12 months, reported at wave 4.

[b]. Weighted US population estimates.

[c]. Other tobacco products include cigars (traditional, cigarillo, and filtered), pipes, hookah, snus, or other smokeless products.

[d]. e-Cigarette use and pharmaceutical-aid status for most recent quit attempt, among all smokers reporting a quit attempt at wave 3.

[e]. Pharmaceutical aids include varenicline and buproprion.

[f]. Combustible products include cigars, pipes, and hookahs.

[g]. Smokeless products include snus, moist snuff, dip, and spit and chewing tobacco.

long-term abstinent from e-cigarettes at follow-up. Among those who successfully used approved pharmacotherapy to quit smoking, approximately 70% were abstinent from NRT. Among the larger group who successfully quit smoking without use of e-cigarettes (who may have used no aid or approved pharmacotherapy), greater than 90% were long-term abstinent from each of NRT and e-cigarettes at follow-up. Importantly, in each comparison group of cigarette-abstinent smokers, 7%– 17% were still using some form of combusted tobacco at follow-up. Overall, among US smokers in 2014–2015 who reported using e-cigarettes to quit in the following year, 8.4% (95% CI: 5.4%, 11.4%) had quit smoking and appeared to have substituted e-cigarettes for their cigarettes by 2016– 2017.

2.5 Discussion

We used the PATH survey to prospectively compare long-term cessation outcomes between a nationally representative sample of US smokers who tried to quit smoking with the help of e-cigarettes in 2016–2017 and a matched sample of US smokers who also tried to quit but without using e-cigarettes. We found that e-cigarette users did not have higher rates of long-term abstinence from cigarette smoking but did have lower rates of abstinence from nicotine than their matched peers. This difference appeared to be largely due to high rates of continuing use of e-cigarettes among those who quit smoking cigarettes. Two-thirds of those who successfully used e-cigarettes to attain long-term abstinence from cigarettes were still using e-cigarettes during the follow-up year. It would be important to assess eventual relapse rates among these groups. [22] We also compared abstinence rates among those who used e-cigarettes to quit and a matched sample of those who used US Food and Drug Administration–approved pharmaceutical cessation aids. Estimated effects were very similar, but confidence intervals were wider, likely due to the smaller matched sample sizes.

The low rates of nicotine abstinence found in our study are worth noting. We included in this measure e-cigarettes, other tobacco products, and NRT products. Long-term nicotine abstinence was well under 5% for US smokers who used e-cigarettes to quit, and less than 10% for those who did not. Our matched analysis attributes 4 percentage points of this difference to the use of e-cigarettes. Of particular concern is the high rate of continued smoking of other forms of tobacco among those who successfully quit cigarettes, ranging from 17% of those who successfully used e-cigarettes to quit to 7% among successful pharmaceutical aid users.

Smokers who used e-cigarettes to try to quit smoking were younger, more educated and affluent, had higher nicotine dependence levels, and were more likely to report mental health symptoms than smokers who tried to quit without e-cigarettes. We used propensity-score methods to match each e-cigarette user with up to 2 similar smokers who did not use e-cigarettes, and we compared the difference in abstinence rates for the matched samples. This procedure allowed us to estimate the average causal effect of e-cigarettes among the population of people who use them. [74] Alternatively, regression-based modeling can estimate average causal effects over the whole population, although at the risk of extrapolation to smokers who are unlikely to ever use e-cigarettes. Indeed, there were few non–e-cigarette users with a propensity score greater than 50%, whereas approximately 20% of e-cigarette users had a propensity score greater than 50%, indicating that such model-based extrapolation is needed to estimate a population-averaged effect. However, we used these types of model-based methods in our sensitivity analyses and obtained qualitatively similar results.

At the population level, we estimated that approximately 13% of US smokers who made a quit attempt using e-cigarettes achieved long-term smoking cessation success, as did approximately 11% of US smokers who tried to quit without use of e-cigarettes, similar to the propensity-score matched estimate of a difference of 2 percentage points in cessation rates. The 95% confidence interval for the matched difference in cessation rates was from 3 percentage points to 7 percentage points. These cessation rates observed in PATH are similar to those seen in other population studies. For example, the 2008 clinical practice guidelines for smoking cessation estimated that approximately 13% of US smokers who tried to quit smoking attained 6–12 months abstinence.

In our study, as in other population studies, daily smokers were less likely to quit successfully than nondaily smokers. Interestingly, the unadjusted observed association of e-cigarette use for cessation differed in direction between daily and nondaily smokers. In exploratory post hoc analyses, we used adjusted multivariable logistic regression to investigate interactions between the association of e-cigarette use and daily vs nondaily smoking, as well as with age, education, sex, and race ethnicity. All confidence intervals for these interaction terms included 1; however, estimated interactions for education and daily versus nondaily smoking appeared to be worth future investigation and are reported in A.5.

Our finding that e-cigarette use to quit smoking did not increase cigarette abstinence to 12 months or longer is similar to results from the Adult Tobacco Use Cohort, in which researchers found a cessation benefit for e-cigarettes at 6 months but not at 12 or 18 months. Using an earlier PATH Study cohort, [13] we reported that using an e-cigarette to quit was associated with short-term abstinence ($\geq$ 30 days); here, abstinence was reported contemporaneously with the report of use of e-cigarettes to quit. Thus, it is possible that e-cigarettes help short-term quitting but not sustained abstinence rates. These results are also consistent with those of a recent study using the PATH waves 1–3 database, [101] in which e-cigarette use among older smokers was associated with abstinence at wave 2 but relapse by wave 3.

Our results on substitution of e-cigarettes for cigarettes are qualitatively similar to the randomized trial of attendees to UK National Health Service stop-smoking services, in which 80% of successful quitters in the e-cigarette arm continued to use e-cigarettes at 1 year, compared with persistent use of NRT by only 9% of successful quitters in the NRT arm. [31] However, we did not replicate this trial's findings of a sustained cessation benefit from use of an e-cigarette to quit. The difference in our results may be related to the intensiveness of the UK intervention or to the lower level of nicotine in UK e-cigarettes. The motivation level of participants might also account for these differences: only 43% of those screened for the UK trial were randomly assigned to the study, whereas the PATH Study estimates are representative of the US population. Similar differences in conclusions between randomized trials and observational studies have been reported regarding use of NRT to quit. [13, 49] Our findings, however, are consistent with the lack of efficacy of e-cigarettes in the recent pragmatic randomized trial of provision of e-cigarettes to help cessation among employees at US workplaces who smoke. [32] It is possible that participants in the pragmatic trial more closely match the general population of US smokers who want to quit.

Strengths of this study include that data were drawn from a large representative sample of the US population who report tobacco use annually, that we used a prospectively assessed measure of 12-month abstinence, and we aimed to assess the causal effect of e-cigarettes for cessation as they are used in the US population. Results were robust to a variety of sensitivity analyses, and our propensity-score approach is relatively robust to modeling assumptions. [9] However, a limitation of all observational studies is the possibility of unmeasured confounding, such as differences in motivation level to quit smoking, in quitting history, or in self-efficacy to successfully quit smoking. The survey measures used are self-reported and, as such, may have measurement error. Although biomarkers of tobacco use are collected in the PATH Study, these were not available to validate the outcome at the time of writing. However, in an analysis of earlier PATH data, self-reported tobacco use was strongly associated with biomarker data. [77] In this study, the e-cigarette devices used were those that were generally available in 2015–2016 and the results may not generalize to the modifications in available products since that time.

In conclusion, in this chapter we compared long-term abstinence rates between a nationally representative cohort of US smokers who tried to quit smoking using e-cigarettes as a cessation aid, and a matched sample of smokers who tried to quit without using e-cigarettes. We found no evidence that e-cigarettes helped these smokers to successfully quit smoking. We estimated that approximately 8% of all adult US smokers who used an e-cigarette to quit cigarettes in 2015–2016 were able to successfully substitute e-cigarettes for cigarette smoking. However, our propensity score–matched results suggest these smokers would have been equally successful in quitting smoking without the use of e-cigarettes. Furthermore, our results suggest, these respondents were more likely to remain dependent on nicotine, largely due to continuing use of e-cigarettes.

2.6 Study funding

This work was funded by National Institutes of Health grant 1 RO1CA234539. The data were analyzed as part of a restricted-use file made available by the National Addiction & HIC Data Archive Program hosted by the Inter-university Consortium for Political and Social Research at University of Michigan. J.P.P. was supported by grants from National Cancer Institute during the conduct of the study; M.M.W. was supported by grants from National Institute of Health during the conduct of the study; and D.R.S. was supported by the National Cancer Institute (grant RO1CA234539) during the conduct of the study. D.R.T was funded on the Tobacco-Related Disease Program of the University of California, Office of the President (grant 28IR - 0066).

Conflict of interest: none declared.

2.7 Afterthoughts before the next Chapter

This chapter demonstrated some essential findings about the effectiveness of e-cigarette use on cigarette cessation, which may help regulate e-cigarettes more broadly in terms of tobacco control. During the course of this study, I had the chance to learn more about the use of causal inference in an important observational study, and I gained a comprehensive understanding of PSM. This motivated me to read further and to think about how to improve the existing well-studied PSM. As we have seen earlier in this chapter, PSM depends on the correctness of the PS which in turn requires correct specification of the PS model. In practice, it is hard to assume such a model is truly correct. In addition, in this chapter, we used bootstrap techniques to make statistical inference. However, so far, it was not well understood how to make correct inference for the matching estimator when using complex survey data. This chapter also motivated us to explore how to conduct correct variance estimation of matching estimators incorporating survey weights and respecting the complex survey design, which we will discuss in chapter 4.

In the following chapter, we will introduce the doubly-matched estimator, which is matched on both the PS and another balancing score, the prognostic score (PGS), to make doubly robust causal inference. This improves the PSM in terms of estimating the ACET, in that when the PS is incorrectly specified, there is still a chance to make the matching estimator consistent as long as the PGS is correctly specified.

This chapter, in full, has been published and may be found as ”Chen, Ruifeng; Pierce, John P.; Leas, Eric C.; White, Martha M.; Kealey, Sheila; Strong, David R.; Trinidad, Dennis

R.; Benmarhnia, Tarik; Messer, Karen. *Use of Electronic Cigarettes to Aid Long-Term Smoking Cessation in the United States: Prospective Evidence From the PATH Cohort Study*, American Journal of Epidemiology, 189 (12), 1529–1537, 2020”. The dissertation author was the primary author on this paper.

Chapter 3

Doubly robust causal inference using the doubly-matched estimator, with application to the causal effect of e-cigarette use on smoking cessation and cigarette consumption reduction

3.1 Abstract

Matching on the propensity score (PS) is a popular approach to causal inference in observational studies for estimating the average causal effect on the treated (ACET), because of its interpretability and robustness. There is also a large literature on doubly-robust (DR) estimators, however there is relatively little development as yet of matching approaches to DR estimation. Doubly-matched estimators match simultaneously on the PS, which models the probability of selecting treatment, and the prognostic score (PGS), which models the potential outcome. We

define and study the doubly-matched estimator DM for estimating the ACET, and show that the DM estimator is DR under standard assumptions; that is, it is consistent when either the PS or the PGS is correctly estimated. Performance of the estimator is compared by simulation with a more common DR approaches, PS matching followed by regression adjustment. Approaches to interval estimation are investigated, namely, full bootstrap percentile confidence intervals, conditional bootstrap confidence intervals, and a conditional parametric approach. Overall, The DM estimator which uses exact matching outperform other approaches when all predictors are categorical, while nearest-neighbor matching performs well under broader circumstances. As expected, double-matching without replacement is less variable than with replacement. Full bootstrap-percentile confidence intervals can be conservative in some circumstance, but can be a generally recommended approach; conditional intervals are computationally efficient and can work well as long as one of PS and PGS model is correct for the DM estimator, but can encounter under coverage in some circumstances. Use of the proposed estimators is demonstrated for a case study investigating the causal effect of e-cigarette use for smoking cessation and cigarette consumption reduction respectively among US smokers who used e-cigarettes to quit cigarette smoking, using data from a large nationally representative longitudinal survey.

3.2 Introduction

Matching estimators are a popular approach to causal inference because of their transparency and robustness, particularly for estimating the average causal effect of an exposure among the treated (ACET). [73] The most common approach takes the matching criterion to be the propensity score (PS), which models the probability of selecting treatment conditional on the confounders, often e.g. using logistic regression. For example, a popular approach to estimating the ACET matches each treated subject to a control with the same PS, or to the nearest control, and computes the ACET as the average difference in outcome between treated subjects and their matched controls. Advantages of the matching approach include that the resulting estimate is highly interpretable, and that it is easy to evaluate the success of the statistical adjustment by comparing the distribution of confounders between treated subjects and and their matched controls. It is also made explicit when there is a group of treated subjects for whom there are no matched controls, and thus for whom any estimates of treatment effect would necessarily rely on model-based extrapolation. There is a large literature on matching estimators, and with popular matching algorithms implemented in the highly cited software package in R. [89]

There is also an extensive literature on doubly-robust (DR) estimators (of both the ACET and the average causal effect among the population, the ACE), which is largely separate from the matching literature. Popular approaches to DR estimators include augmented inverse probability weighting (AIPW) and related estimators. [72, 42, 18, 92, 98, 85, 82, 93] DR estimators use both a PS and a prognostic score (PGS, i.e. a regression model which models the potential outcome conditional on covariates), and have the property that they are consistent when either model is correctly specified. [42] To date, these two different approaches to causal inference have remained fairly distinct, in that matching approaches to DR inference have received comparatively less attention. [48] In this chapter, we investigate the use of doubly-matched estimators, which match on both the PS and the PGS, to make DR causal inference for estimating the ACET.

3.2.1 The motivation

We are motivated by two of our recent studies which used propensity score matching (PSM) to investigate whether e-cigarette use, under current use patterns in the US population, would help smokers to quit smoking. [66, 20] This question is of interest to the US Food and Drug Administration (FDA) to inform its regulation of e-cigarettes. The public health question is to balance a potential benefit to existing smokers, who might switch from conventional cigarettes to less harmful e-cigarettes, with the potential harm from widespread co-marketing and use of e-cigarettes, which may potentially increase traditional cigarette smoking particularly among young people. Our objective was to assess whether use of e-cigarettes as a smoking cessation aid was causally associated with future abstinence from cigarette smoking.

The study population was US smokers who actually used e-cigarettes in a quit attempt; thus the ACET was the estimand. The data were from multiple waves of the Population Assessment of Tobacco and Health (PATH) study, a large nationally representative survey series jointly sponsored by the FDA and the US National Institutes of Health. The approach was to match each e-cigarette user with a comparable non-e-cigarette user using nearest-neighbor caliper PSM, and then to compare the abstinence rate between the matched pairs, at one year after the reported quit attempt. The PS for e-cigarette use was estimated by logistic regression, using more than 20 measures which are known predictors of both e-cigarette use and smoking cessation, including sociodemographic variables, cigarette smoking history, nicotine dependence, quitting history, timing of LQA, relative perceived harm (cigarettes, e-cigarettes), social variables and other covariates. To further reduce potential confounding, PS measures were assessed prior to the exposure (i.e. the index quit attempt). The lasso [95] was used to select the best model for the PS, before conducting the matching algorithm. After matching, any residual imbalance in predictors was adjusted for by incorporating these covariates into the PGS, which was then used to compare the matched pairs. For inference, the bootstrap was used to construct interval estimates.

The PSM in these two studies highly depends on the model specification, indicating that a DR estimator would be desirable. This was achieved in our case by inclusion of any imbalanced predictors in the outcome model. [89] However, given the desirable properties of matching estimators described above, we were motivated to investigate the properties of a fully matching-based DR estimator. In addition, it was unclear from the literature how to best construct a confidence interval from such a DR matching estimator.

3.2.2 Doubly-matched estimators

The doubly-matched estimators we investigate here involve simultaneous matching on both the PS and the PGS. As is well-known, the PS has the balancing property that, conditional on the PS, the covariates and the indicator of treatment are independent, and is the coarsest such score. [74] Similarly, the PGS as defined by Hansen [33] is any function of covariates which, within a given exposure group, enables conditional independence of the potential outcome and the covariates. Under assumptions, the PGS can be taken as the expected value of the outcome of a treated or a control case, conditional on covariates. Conditioning on the PGS provides an additional method to eliminate covariate imbalance between two treatment groups of interest.

Relatively few studies have investigated doubly-matched estimators. Leacy [48] used simulation to study a variety of methods for matching on the PS and / or the PGS to estimate the ACET, and noted the empirical double robustness of several approaches. Lee [51] et al. were the first to give explicit conditions under which a DR estimator for the ACET can be obtained by matching on both the PS and the PGS simultaneously, with additional conditions for DR estimation of the ACE. Antonelli [8] proved the consistency of a doubly-matched estimator using a Lasso approach for variable selection in a high dimensional setting, and investigated an ad hoc parametric method for constructing confidence intervals.

How to best obtain a confidence interval for a PS matching estimator, or more generally a matching estimator, is an active area of research. [50, 39, 34, 2, 12, 4, 8, 15] The details of the matching algorithm appear to be important. Considering matching on the PS with replacement, Hill [34] showed that the bootstrap works well to construct confidence intervals. Abadie [4] derived an expression for the asymptotic variance of the PS matched estimator, and proposed a corresponding ”plug in” variance estimate for matching with replacement, and they also demonstrate that sampling without replacement is more efficient than sampling with replacement. Bodory et al [15] conducted a comprehensive simulation study of nearest neighbor (NN) caliper matching, including matching followed by regression adjustment, and confirm that plug in variance estimates which do not adjust for the estimated propensity score typically produce conservative confidence intervals. They also show that bootstrap methods generally perform better than the Abadie approach. Considering matching without replacement, Austin [12] performed a simulation study of confidence interval methods, including the bootstrap both re-estimating the PS and conditional on the original PS, and parametric estimators of variance which do and do not account for matching. Parametric estimators conditional on the matching performed well. Normal theory confidence intervals performed well for caliper matching. Abadie [5] proved that ignoring the matching step in a post matching regression still results in valid standard error estimation of the model coefficients when the regression model is correct and matching is done without replacement. In addition, conditional approaches were given to make correct inference when the regression model is incorrectly specified. However, the recommended approaches under various circumstances remain unclear, and we are not aware of studies that investigated interval estimation for doubly-matched estimators.

3.2.3 Organization of the chapter

In this chapter, we study the doubly-matched estimator of the ACET, and give an elementary demonstration of their double robustness. We also study the performance of confidence interval approaches for the doubly-matched estimator. We compare the doubly-matched estimator with more common estimators such as the ordinary least square estimator, the PSM estimator, and a PSM estimator followed by regression adjustment. Confidence interval approaches considered for the doubly-matched estimator include a full bootstrap, a bootstrap conditional on the matching, and conditional parametric approaches. The chapter is organized as follows: In 3.3, we define the doubly-matched estimator and demonstrate its double robust properties. We also define algorithms for confidence interval construction. In 3.4, performance of the point estimation of the doubly-matched methods is investigated by simulation. In 3.5, simulation studies are conducted to assess the interval estimators for doubly-matched estimators. In 3.6, these methods are applied to the question of effectiveness of e-cigarette use. 3.7 summarizes results and explores future work. 3.8

provides the afterthoughts before next chapter.

3.3 Doubly-matched estimators

3.3.1 Notation

Let the data consist of n i.i.d. subjects with outcome Y_i, covariates W_i, and a Bernoulli treatment indicator R_i. We follow the potential outcome framework, [54] where the outcome $Y_i = (Y_{i1}, Y_{i0})$, and where potential outcome Y_{i1} is observed if $R_i = 1$, and Y_{i0} is observed if $R_i = 0$. One and only one of Y_{i1}, Y_{i0} is observed according to whether the case is in the treated or the control condition. The predictors $W_i = (W_{i,1}, \ldots, W_{i,p})$ are fully observed. The PS is defined as $\pi(W_i) = E(R_i|W_i)$, and the PGS will be taken to be $m_0(W_i) = E(Y_{i0}|W_i)$, the outcome of a control case conditional on covariates. We sometimes write π_i for $\pi(W_i)$, and similarly for m_{i0}. The ACET is defined as

$$\text{ACET} = E[Y_1 - Y_0|R = 1] \tag{3.1}$$

and the population ACE as $\text{ACE} = E[Y_1 - Y_0]$ (not of main interest in our work).

3.3.2 Elementary properties of the prognostic and propensity scores

Throughout we assume that the basic causal assumptions hold: 1) The treatment must occur before outcome; 2) $P[R = 1|W] > 0$; 3) there are no unmeasured confounders, so that $(Y_0, Y_1) \perp R|W$; and 4) potential outcomes of any subject are independent of potential outcomes of other subjects. Then, conditioning on the PS deconfounds Y and R , [74, 54, 6] that is,

$$Y \perp R|\pi(W). \tag{3.2}$$

We further assume that $m_0(W) = E(Y_0|W)$ and $m_1(W) = E(Y_1|W)$ are prognostic scores, that is, that $Y_0 \perp W|m_0(W)$ and $Y_1 \perp W|m_1(W)$. In this case, conditioning on the appropriate prognostic score also deconfounds Y_0 and Y_1 with (R, W) [33]

$$Y_0 \perp (R,W)|m_0(W). \tag{3.3}$$

and

$$Y_1 \perp (R,W)|m_1(W). \tag{3.4}$$

3.3.3 A doubly-matched estimator of the ACET

We first consider one-to-one double matching without replacement. For each treated case i (i.e. a case with observed $R_i = 1$) let $\bar{Y}_{i0}$ be the outcome for a case randomly chosen with equal probability from among the set of cases j with observed $R_j = 0$, $\pi(W_j) = \pi(W_i)$ and $m_0(W_j) = m_0(W_i)$. Then the doubly-matched estimator of the ACET (DM) is

$$\hat{\Delta}_{DM} = \frac{1}{n\hat{\pi}} \sum_{i=1}^{n} R_i(Y_{i1} - \bar{Y}_{i0}) \tag{3.5}$$

where $\hat{\pi} = \sum_{i=1}^{n} \frac{R_i}{n}$.

The DM estimator is DR, as can be seen from the following: suppose either the propensity model is correct, i.e. formula 3.2 holds, or the prognostic model is correct, i.e. formula 3.3 holds.

Then, in either case, R and Y_0 are conditionally independent given $\pi(W)$ and $m_0(W)$, and we have

$$\begin{aligned}
E[\bar{Y}_{i0}|R_i=1] &= E_{\pi,m_0|R=1}[E[\bar{Y}_{i0}|\pi(W_i),m_0(W_i),R_i=1]] \\
&= E_{\pi,m_0|R=1}[E[Y_{j0}|\pi(W_i),m_0(W_i),\hat{\pi}(W_j)=\hat{\pi}(W_j), \\
&\quad \hat{m}_0(W_j)=\hat{m}_0(W_i),R_j=0, j \text{ is selected}, R_i=1]] \\
&\xrightarrow{p} E_{\pi,m_0|R=1}[E[Y_{j0}|\pi(W_i),m_0(W_i),\pi(W_j)=\pi(W_j), \\
&\quad m_0(W_j)=m_0(W_i),R_j=0, j \text{ is selected}, R_i=1]] \\
&= E_{\pi,m_0|R=1}[E[Y_{i0}|\pi(W_i),m_0(W_i),R_i=1]] \\
&= E[Y_{i0}|R_i=1] \qquad (3.6)
\end{aligned}$$

Note that under mild regularity conditions, $E[\frac{RY}{\hat{\pi}}] \to^p E[\frac{RY}{\pi}] = E[Y|R=1]$, hence $\hat{\Delta}_{DM}$ is consistent for the ACET. This algorithm describes exact matching with replacement, suitable for categorical covariates; for continuous covariates we use nearest neighbor matching. [76]

The doubly-matched estimator of the population average causal effect can be formed in a similar way, which is constructed from two-step matching: first each treated case is matched with a control using the algorithm above, then the algorithm is re-run, matching each control to a treated case. Consistency of this population estimator can also be proved.

3.3.4 Interval estimation for doubly-matched estimators

We consider confidence interval methods for the doubly-matched estimator DM, constructed using three different approaches.

The full bootstrap

The full bootstrap generates B bootstrap samples from the original sample of individual subjects, with replacement. Within each bootstrap sample, we re-estimate the PS and PGS and carry out the matching algorithm, then compute the mean difference between the matched pairs. A bootstrap percentile confidence interval for the mean difference is then constructed.

The conditional bootstrap

The conditional bootstrap is conditional on the original sample of m matched pairs obtained by running the matching algorithm on the data: generate B bootstrap samples of matched pairs, by resampling with replacement from this original set of m matched pairs. Note that there is no re-estimation of PS or PGS, and no re-matching, as the pairs are already matched. For each bootstrap sample of pairs, calculate the mean difference between pairs, and then construct a bootstrap percentile confidence interval for the mean difference.

A conditional parametric approach

The conditional parametric confidence interval also takes the matched pairs as given, and assumes the sample of differences of the matched pairs are i.i.d. draws from an appropriate parametric distribution. Then standard statistical theory can be used to construct a confidence interval, based on sample statistics such as the estimated mean difference and its SE.

3.4 Simulation performance of the doubly-matched point estimators

In this section we use simulation to compare the performance of the doubly-matched DR estimator DM to a more standard model-based DR counterpart, namely PSM followed by regression adjustment (PSM-OLS) [89]. As a benchmark to evaluate the degree of confounding by treatment selection in our simulation scenarios, we include the ordinary least squares (OLS) estimator, which is consistent (and efficient) when the PGS model is correct but not otherwise. We also include the PSM estimator which is consistent (but not efficient) when the PS model is correct but not otherwise. In addition, we add the naive estimator which computes the mean difference between observed cases and controls. We study performance under both correct and incorrect PGS and PS models. Importantly, we have set an interaction between a covariate and the treatment indicator to make the true ACET differ from the true population ACE. The performance of the corresponding interval estimators (i.e. confidence intervals) is presented in 3.5.

3.4.1 Simulation details

We compare the bias, efficiency, and RMSE of these estimators under four situations, according to whether the PS / PGS models are correct / incorrect. The specific performance metrics are: bias, the average difference between the estimate and the true treatment effect; standard deviation (SD), the standard deviation of the estimate across the simulation samples; bias / SD, the bias expressed as a percentage of the SD [42]; RMSE, square root of the sum of the bias squared and the SD squared; and, for the matching methods, the proportion of successful matches. These metrics are generally computed within each simulated sample and then averaged across the simulation.

The simulation size is 1000 and the sample size is 500. We consider two generative simulation models, one with categorical covariates, where we use exact matching (Scenarios 1 and 2), and one with continuous covariates, where we use nearest-neighbor matching (Scenario 3). In Scenario 1, mis-specified models lead to incorrect matching while in Scenario 2 mis-specified models still allow for correct matching, in order to demonstrate a robustness property of matching methods. In Scenario 3, the matching is approximate. In each scenario, the mean proportion of treated subjects is around 20%. The matching algorithms use 1:1 matching without replacement. In these scenarios, matching of 1:2 or higher would be expected to reduce the variance, and matching with replacement to increase the variance, in the case of exact matching without affecting the bias. Matching with and without replacement is considered in 3.5 on confidence interval estimation.

Scenario 1: Exact matching, with severe model mis-specification alternatives

For i = 1, ..., 500, we generate i.i.d. Bernoulli variables $(w_{i1}, w_{i2}, w_{i3})^T$, each taking the value of 1 or 2 with probability 0.5. The treatment indicator R_i is generated as an independent Bernoulli trial with probability

$$\pi_i = expit(0.8 - 2w_{i1} + 0.3w_{i2}) \qquad (3.7)$$

The outcome is generated by

$$y_i = 5 + 5r_i + 10r_i w_{i1} + w_{i1} + w_{i2} + w_{i3} + 2w_{i1}w_{i2} + 4w_{i1}w_{i3} + 6w_{i2}w_{i3} + \varepsilon \qquad (3.8)$$

where ε is from an independent standard normal distribution. We conduct exact matching without replacement, on both the estimated PS, obtained by logistic regression, and the estimated PGS for the control group, obtained by OLS. Subjects for whom there is not an exact match on both scores are omitted from the matched sample. For the form of the estimated PS and PGS models, we use both the correct specification (corresponding to each generative model, equation 3.7 and 3.8), a mis-specified PS model which omits covariate w_2, and a mis-specified PGS model which omits w_2, all two-way interactions and the interation between treatment indicator and w_1. Under these circumstances, matching on a mis-specified estimated model will result in an incorrect match with high probability.

Scenario 2: Exact matching, with mild model mis-specification alternatives

Scenario 2 is identical to Scenario 1, except that in the mis-specified PS model, w_1 and w_2 are included but the intercept is excluded. In this scenario, when the PS model is incorrectly specified the match is still likely to be correct.

Scenario 3: Nearest-neighbor matching

For i = 1, ..., 500, we generate i.i.d. normal variables $(w_{i1}, w_{i2}, w_{i3})^T$, with the same variance 1 and means 1, 2, and 3, respectively. The treatment indicator R_i is generated as an independent Bernoulli trial with probability

$$\pi_i = expit(-1.5 - 0.5w_{i1} + 0.3w_{i2}) \tag{3.9}$$

and the outcome is generated by

$$y_i = 5 + 5r_i + 10r_i w_{i1} + w_{i1} + w_{i2} + w_{i3} + 2w_{i1}w_{i2} + 4w_{i1}w_{i3} + 6w_{i2}w_{i3} + \varepsilon \tag{3.10}$$

where ε is from standard normal distribution. We conduct nearest-neighbor matching without replacement, on both the estimated PS, and the estimated PGS. Subjects for whom there is not a match on both scores are omitted from the matched sample. Correct and incorrect models are defined as in Scenario 1.

3.4.2 Results evaluating the doubly matched point estimators

Scenarios 1 and 2: Exact Matching

Simulation results for exact matching, the most favorable setting for the matching estimators, are given in Table 3.1 and Table 3.2. The high selection bias in these scenarios is demonstrated by the naive estimator, which computes the mean difference between observed cases and controls. The correctly specified OLS estimator is unbiased and efficient for the ACET, providing a benchmark. Under the correctly specified models, the doubly robust DM estimator and the doubly robust PSM-OLS estimator perform similarly to each other and are nearly as efficient as OLS, although they pay a tiny RMSE penalty for their double robustness. The correctly specified PSM estimator is inefficient but it's unbiased for estimating the ACET. Almost every treated subject is matched with a control subject.

When either the PS model or the PGS model is incorrect (severely mis-specified in Table 3.1, less severely in Table 3.2), all DR methods vastly outperform the non-DR methods (OLS and PSM), which are unacceptable in these circumstances. When the PGS model is correct but PS model is incorrect, the two DR estimators DM and PSM-OLS have similar RMSE. However, when the PGS model is incorrect but the PS model is correct, the doubly-matched estimator DM substantially outperform the regression modeling approach PSM-OLS in terms of RMSE. The DM estimator is around 15% more efficient than the PSM-OLS estimator. It is likely that these estimators would perform even better using 1-2 matching. Finally, when neither the PGS nor the PS model is correctly specified, all estimators perform similarly, and none is acceptable.

Scenario 2 (Table 3.2) is the same as Scenario 1, except that the form of the PS model mis-specification differs, and here does not necessarily cause a mis-match. In this case the advantage of the double-matched estimator over the competing regression-based approach is very pronounced. This example illustrates the potential for the doubly-matched estimator to enjoy extra robustness against mis-specification of the model, compared to the regression-based estimator.

Scenario 3: Nearest Neighbor Matching

Scenario 3 has continuous covariates, and so exact matching is no longer possible. Nearest neighbor matching is used and all subjects in the treatment group have matches in the control group. Results are summarized in Table 3.3. As before, the uncorrected naive estimator shows a high percent bias and RMSE. The degree of confounding in this simulation scenario is lighter compared to that in the discrete case (-2.42 vs -3.32). The correctly specified OLS estimator is unbiased and efficient, providing a benchmark. Overall, results are quite consistent with the results in Scenario 1, as expected. The doubly-matched estimator DM is slightly less efficient than the model based PSM-OLS estimator when the PGS model is correct (1.09 vs 0.96 when PS is correct; 1.02 vs 0.97 when PS is incorrect). However, when the PS is correct and the PGS is incorrect, the

Table 3.1: Point estimate performance: categorical covariates with exact matching; severe model mis-specification.

Method	Estimand	PS	PGS	Bias	SD	Bias/SD	RMSE
		model	model				
Naive	-	-	-	-3.322	1.385	-2.399	3.599
		Correct	Correct				
OLS	ACET			0.008	0.397	0.021	0.397
PSM	ACET			0.023	1.196	0.019	1.196
PSM-OLS*	ACET			0.009	0.404	0.021	0.404
DM*	**ACET**			**0.010**	**0.408**	**0.025**	**0.408**
		Incorrect	Correct				
OLS	ACET			0.001	0.397	0.002	0.397
PSM	ACET			0.864	1.441	0.600	1.681
PSM-OLS*	ACET			0.003	0.405	0.007	0.405
DM*	**ACET**			**-0.001**	**0.404**	**-0.002**	**0.404**
		Correct	Incorrect				
OLS	ACET			1.359	0.886	1.534	1.623
PSM	ACET			0.009	1.216	0.008	1.216
PSM-OLS*	ACET			0.002	0.465	0.004	0.465
DM*	**ACET**			**0.000**	**0.403**	**0.000**	**0.403**
		Incorrect	Incorrect				
OLS	ACET			1.344	0.877	1.533	1.604
PSM	ACET			0.843	1.467	0.575	1.692
PSM-OLS*	ACET			0.861	0.977	0.881	1.303
DM*	**ACET**			**0.868**	**0.956**	**0.908**	**1.291**

*Doubly robust method; **Bold** indicates the doubly-matched estimator

OLS: ordinary least squares; PSM: propensity score matched

PSM-OLS: PSM followed by OLS; DM: doubly-matched

Table 3.2: Point estimate performance: categorical covariates with exact matching; mild model mis-specification.

Method	Estimand	PS model	PGS model	Bias	SD	Bias/SD	RMSE
Naive	-	-	-	-3.322	1.385	-2.399	3.599
		Correct	Correct				
OLS	ACET			0.008	0.397	0.021	0.397
PSM	ACET			0.023	1.196	0.019	1.196
PSM-OLS*	ACET			0.009	0.404	0.021	0.404
DM*	**ACET**			**0.010**	**0.408**	**0.025**	**0.408**
		Incorrect	Correct				
OLS	ACET			0.007	0.407	0.018	0.407
PSM	ACET			0.007	1.149	0.006	1.149
PSM-OLS*	ACET			0.004	0.411	0.010	0.411
DM*	**ACET**			**0.008**	**0.412**	**0.020**	**0.412**
		Correct	Incorrect				
OLS	ACET			1.359	0.886	1.534	1.623
PSM	ACET			0.009	1.216	0.008	1.216
PSM-OLS*	ACET			0.002	0.465	0.004	0.465
DM*	**ACET**			**0.000**	**0.403**	**0.000**	**0.403**
		Incorrect	Incorrect				
OLS	ACET			1.363	0.904	1.509	1.635
PSM	ACET			0.006	1.185	0.005	1.185
PSM-OLS*	ACET			0.016	0.476	0.034	0.476
DM*	**ACET**			**0.000**	**0.406**	**0.000**	**0.406**

*Doubly robust method; **Bold** indicates the doubly-matched estimator

OLS: ordinary least squares; PSM: propensity score matched

PSM-OLS: PSM followed by OLS; DM: doubly-matched

DM estimator significantly outperforms the PSM-OLS estimator (RMSE: 1.10 vs 1.85). All DR methods now pay a more substantial price in terms of increased variance, compared to the discrete covariate case.

3.5 Simulation performance of the double matching interval estimators

We next use simulation to compare the performance of the three approaches to confidence interval estimation for the doubly-matched estimator DM, namely a full bootstrap percentile approach, a conditional bootstrap approach, and a conditional parametric approach. In these comparisons we consider matching both with and without replacement. For the comparison estimators, details of the interval estimators are as follows: bootstrap percentile approaches are used for interval estimation for the naive and the OLS estimators; the same three interval estimation approaches we use for the DM estimator are also used for the PSM estimator; and the full bootstrap percentile approach and the conditional bootstrap approach are used for assessing the confidence interval for the PSM-OLS estimator.

Performance metrics for each confidence interval method are: the coverage probability; the mean width; and the interval score [29] given by $S = (u-l)+2/\alpha(l-x)I(x<l)+2/\alpha(x-u)I(x>u)$, where x is the true value of the estimand, u is the upper confidence limit and l is the lower confidence limit. We consider simulation Scenarios 1 with categorical covariates in which we do exact matching, and we assess the interval estimation performance in cases where at least one of the PS and the PGS models is correct. Both matching with and without replacement are considered. The sample size is 300, the bootstrap size is 1000 and the simulation size is 500.

Table 3.3: Point estimate performance: continuous covariates and nearest neighbor caliper matching.

Method	Estimand	PS model	PGS model	Bias	SD	Bias/SD	RMSE
Naive	-	-	-	-2.417	4.156	-0.582	4.808
		Correct	Correct				
OLS	ACET			-0.049	0.953	-0.051	0.955
PSM	ACET			-0.145	3.551	-0.041	3.554
PSM-OLS*	ACET			-0.046	0.954	-0.049	0.955
DM*	**ACET**			**0.017**	**1.085**	**0.015**	**1.085**
		Incorrect	Correct				
OLS	ACET			-0.043	0.964	-0.044	0.965
PSM	ACET			5.591	3.794	1.474	6.757
PSM-OLS*	ACET			-0.042	0.967	-0.043	0.968
DM*	**ACET**			**0.152**	**1.013**	**0.150**	**1.024**
		Correct	Incorrect				
OLS	ACET			6.994	2.532	2.762	7.438
PSM	ACET			-0.018	3.388	-0.005	3.388
PSM-OLS*	ACET			0.264	1.828	0.144	1.846
DM*	**ACET**			**0.092**	**1.099**	**0.084**	**1.103**
		Incorrect	Incorrect				
OLS	ACET			31.278	4.630	6.755	31.619
PSM	ACET			30.260	8.276	3.656	31.371
PSM-OLS*	ACET			30.468	5.900	5.164	31.034
DM*	**ACET**			**30.266**	**5.962**	**5.076**	**30.848**

*Doubly robust method; **Bold** indicates the doubly-matched estimator

OLS: ordinary least squares; PSM: propensity score matched

PSM-OLS: PSM followed by OLS; DM: doubly-matched

3.5.1 Simulation results for the interval estimators

Simulation results for matching with and without replacement are summarized in Table 3.4 for the case when both the PS and the PGS models are correct. Table 3.5 is for the case when the PS is incorrect but the PGS model is correct, and Table 3.6 represents the case in which the PS is correct but the PGS model is incorrect.

In Table 3.4, the bootstrap quantile confidence interval performs well for the OLS estimator when both the PS and the PGS models are correct, and it shows the OLS estimator is the most efficient one (smallest mean CI width and interval score), consistent with the results in Table 3.1. For matching without replacement, the full bootstrap performs well and conservative for the PSM estimator, and performs well for the DM estimator and the PSM-OLS estimator. The conditional bootstrap and the conditional parametric approaches perform well for all estimators. The interval estimation for the DR estimators (DM and PSM-OLS) perform better than those using PSM, which is again consistent with our previous findings. In summary, given the large computational advantage, in the setting of discrete covariates and matching without replacement, the conditional bootstrap or corresponding conditional parametric methods can be recommended.

Compared to matching without replacement, the matching estimators with replacement are slightly less efficient, as expected. The full bootstrap again works quite well for all estimators and conservative for the PSM estimator. However, the conditional methods of confidence interval estimation end up with under coverage when matching with replacement. Conditional variance estimators which account for the replicate use of subjects or pairs need to be considered when using matching with replacement. Notice that in our case, the coverage probability which comes from the conditional methods is slightly smaller than 95%. However, as the proportion of replication of matching increases, the coverage probability will decrease as well. Appropriate bootstrap procedures such as the wild bootstrap [15], similar to those suggested in the literature for single matching estimators, may be considered to correct for this under coverage.

When the PS model is incorrect and the PGS model is correct (Table 3.5). The full

Table 3.4: Confidence interval performance: categorical covariates with exact matching; PS and PGS models are correct.

Estimator	Interval Method	Bias	RMSE	Coverage	Mean CI Width	Interval Score**
Naive	Bootstrap	-3.376	3.642	0.341	5.423	43.613
OLS	Bootstrap	-0.011	0.392	0.952	1.525	1.863
	Conditional parametric			0.946	1.537	1.894
Matching without replacement						
PSM	Full bootstrap	-0.019	1.149	0.972	4.617	4.986
	Conditional bootstrap			0.954	4.607	5.330
	Conditional parametric			0.957	4.644	5.356
PSM-OLS*	Full bootstrap	-0.010	0.402	0.955	1.556	1.866
	Conditional bootstrap			0.945	1.546	1.872
	Conditional parametric			0.942	1.569	1.949
DM*	Full bootstrap	-0.010	0.396	0.956	1.558	1.864
	Conditional bootstrap			0.951	1.546	1.881
	Conditional parametric			0.948	1.563	1.920
Matching with replacement						
PSM	Full bootstrap	-0.030	1.289	0.979	5.027	5.198
	Conditional bootstrap			0.915	4.583	6.471
	Conditional parametric			0.920	4.631	6.465
PSM-OLS*	Full bootstrap	-0.013	0.407	0.956	1.579	1.873
	Conditional bootstrap			0.950	1.548	1.894
	Conditional parametric			0.940	1.567	1.990
DM*	Full bootstrap	-0.010	0.405	0.957	1.579	1.873
	Conditional bootstrap			0.943	1.545	1.904
	Conditional parametric			0.946	1.564	1.928

*Doubly robust method

**Smaller is better

Bold indicates the doubly-matched estimator

OLS: ordinary least squares; PSM: propensity score matched; PSM-OLS: PSM followed by OLS

bootstrap still works for the PSM estimator although the PS model is incorrect. In this case, the conditional methods no longer work for the PSM estimator, even though matching without replacement. For the other matching estimators (PSM-OLS and DM), the results do not change significantly compared to the previous case when both models are correct, except, the full bootstrap performs slightly conservative for the PSM-OLS and the DM estimators when the PS model is incorrect. Again, the conditional methods perform well for the PSM-OLS and the DM estimators when matching without replacement, and the coverage is lower using these conditional methods when matching with replacement.

In the case when the PS model is correct and the PGS model is incorrect, Table 3.6 summarizes that the bootstrap no longer performs well for the OLS estimator. For the matching estimators (PSM, PSM-OLS and DM), results are as expected and similar compared to the results in table 3.4, except, the full bootstrap performs conservative for the PSM-OLS estimator, no matter the matching is with or without replacement. This is consistent with our previous results that the incorrect PGS has more effect on the performance of the PSM-OLS estimator compared to the DM estimator in our simulation setting. In addition, the conditional methods end up with slightly under coverage for the DM estimator when matching without replacement. However, the mean CI interval and the interval score do not get diminished, thus this might actually come from the limited simulation sample size. Notice that when the PGS model is incorrect, conditional bootstrap approach still performs very well for the PSM-OLS estimator when the PS model is correct, which is also proved and shown by Abadie [5]. In addition, in the previous Table 3.5 we have shown that the conditional methods perform well for the PSM-OLS and the DM estimators when matching without replacement as long as the estimators are correctly estimated even the PS model is incorrect.

3.5.2 Summary results

In summary, for matching without replacement, conditional methods of confidence interval estimation have a large computational advantage and are recommended for all matching estimators when the estimator is correctly estimated. For matching with replacement, the full bootstrap works well and sometimes it performs conservative, although it may have a computational burden. Conditional methods have under coverage, and would need to adjust for replication. The phenomenon of under coverage gets more severe as the replication rate increases.

3.6 PATH study

We applied the doubly-matched DM estimator to the question of the population effectiveness of e-cigarette use for smoking cessation, using a publicly available longitudinal sample of US smokers from the PATH survey wave 2-4 database. [38] Here, we are interested in the ACET which answers the question whether e-cigarettes help smokers to quit among those who choose to use them as a cessation aid. Among US smokers, the most popular method of smoking cessation is no use of any aid; the next most popular method is use of e-cigarettes; use of FDA approved cessation aids is less popular. Thus the question is of public health and regulatory relevance. Our published work in this area used PSM estimators of the same estimand, the ACET; [66, 20] here we also include for comparison the regression, as well as the regression-based doubly-robust estimator PSM-regression. For the matching, we used more than 20 mixed categorical and continuous variables. We used nearest neighbor matching (without replacement), with full bootstrap confidence intervals as recommended by our investigations, while recognizing that these might have conservative coverage. We study two outcomes, assessed at wave 4: a binary indicator of cigarette abstinence of 12+ months, and a continuous measure of change in cigarette consumption level between wave 4 and wave 2 (reduction in consumption is sometimes considered as a secondary outcome measure).

Table 3.5: Confidence interval performance: categorical covariates with exact matching; PS model is incorrect and PGS model is correct.

Estimator	Interval Method	Bias	RMSE	Coverage	Mean CI Width	Interval Score**
Naive	Bootstrap	-3.376	3.642	0.340	5.427	43.567
OLS	Bootstrap	-0.011	0.392	0.948	1.525	1.861
	Conditional parametric			0.950	1.549	1.921
Matching without replacement						
PSM	Full bootstrap	0.959	1.781	0.935	5.732	6.882
	Conditional bootstrap			0.875	5.693	9.013
	Conditional parametric			0.883	5.753	8.839
PSM-OLS*	Full bootstrap	-0.008	0.396	0.954	1.558	1.863
	Conditional bootstrap			0.954	1.551	1.865
	Conditional parametric			0.944	1.579	1.986
DM*	Full bootstrap	-0.008	0.401	0.953	1.559	1.859
	Conditional bootstrap			0.944	1.549	1.872
	Conditional parametric			0.943	1.567	1.921
Matching with replacement						
PSM	Full bootstrap	0.964	1.853	0.962	6.210	6.995
	Conditional bootstrap			0.860	5.660	9.597
	Conditional parametric			0.868	5.720	9.443
PSM-OLS*	Full bootstrap	-0.010	0.408	0.959	1.578	1.863
	Conditional bootstrap			0.945	1.548	1.916
	Conditional parametric			0.953	1.581	1.953
DM*	Full bootstrap	-0.009	0.402	0.957	1.579	1.867
	Conditional bootstrap			0.943	1.545	1.857
	Conditional parametric			0.946	1.565	1.903

*Doubly robust method

**Smaller is better

Bold indicates the doubly-matched estimator

OLS: ordinary least squares; PSM: propensity score matched; PSM-OLS: PSM followed by OLS

Table 3.6: Confidence interval performance: categorical covariates with exact matching; PS model is correct and PGS model is incorrect.

Estimator	Interval Method	Bias	RMSE	Coverage	Mean CI Width	Interval Score**
Naive	Bootstrap	-3.376	3.642	0.341	5.423	43.613
OLS	Bootstrap	1.428	1.675	0.646	3.481	11.935
	Conditional parametric			0.606	3.325	13.348
Matching without replacement						
PSM	Full bootstrap	-0.019	1.149	0.972	4.617	4.986
	Conditional bootstrap			0.954	4.607	5.330
	Conditional parametric			0.957	4.644	5.356
PSM-OLS*	Full bootstrap	-0.007	0.466	0.976	1.856	2.074
	Conditional bootstrap			0.960	1.849	2.165
	Conditional parametric			1.000	3.800	3.800
DM*	Full bootstrap	-0.010	0.396	0.956	1.558	1.864
	Conditional bootstrap			0.951	1.546	1.881
	Conditional parametric			0.948	1.563	1.920
Matching with replacement						
PSM	Full bootstrap	-0.030	1.289	0.979	5.027	5.120
	Conditional bootstrap			0.915	4.583	6.471
	Conditional parametric			0.920	4.631	6.465
PSM-OLS*	Full bootstrap	-0.006	0.479	0.977	1.928	2.103
	Conditional bootstrap			0.946	1.851	2.189
	Conditional parametric			1.000	3.798	3.798
DM*	Full bootstrap	-0.010	0.405	0.957	1.579	1.873
	Conditional bootstrap			0.943	1.545	1.904
	Conditional parametric			0.946	1.564	1.928

*Doubly robust method

**Smaller is better

Bold indicates the doubly-matched estimator

OLS: ordinary least squares; PSM: propensity score matched; PSM-OLS: PSM followed by OLS

As in our published work, the study population was all respondents who reported current smoking at wave 2, reported a quit attempt at wave 3, and reported subsequent smoking status at wave 4. There was around one year of separation between any two waves. Baseline covariates for the matching were assessed at wave 2; the exposure of interest was e-cigarette use as a cessation aid on the wave 3 quit attempt. Baseline covariates included demographic characteristics, (age, sex, ethnicity, race, education, income, health insurance status), comorbidities (symptoms of mental health problems and smoking-related health diagnoses), and measures of smoking behavior (nicotine dependence, daily cigarette use status at W2, duration and timing of most recent previous quit attempt, pack-years of smoking, age started smoking fairly regularly, self-efficacy about quitting, interest in quitting cigarettes, smoking-free home, exposure to other smokers and perceived harm of cigarettes and e-cigarettes). For the first outcome, the binary indicator of cigarette abstinence of 12+ months, we had one more covariate adjusted, the baseline cigarette consumption per day.

There were 2535 subjects who met our inclusion criteria for our primary binary outcome, 427 of whom used e-cigarettes to aid their quit attempt and who 2108 did not. For the secondary outcome, 2009 of these subjects reported cigarette consumption at both time points, of whom 310 used e-cigarettes to help quit and 1699 did not. The PATH study provides sampling weights which can be used to provide estimates representative of the US population, however the question of how to appropriately incorporate these sampling weights into matching estimators is complex and thus been listed as our future work. In this example we provide unweighted estimates which pertain to the cohort but are not representative of the US population. Simple imputation (R package *mice*) was used to deal with missing values in the baseline covariates. The full bootstrap confidence intervals also incorporate this imputation step.

Table 3.7 summarizes the PATH study results. Considering the binary cessation outcome first, all the comparison estimators including the naive estimator suggested a small positive but non-significant effect for use of e-cigarettes on the long-term smoking cessation among those

Table 3.7: Estimated effect sizes for PATH study data.

		Long-term cigarette cessation		**Change in cigarette consumption**	
Method	Estimand	Risk difference	95% CI	Risk difference	95% CI
Naive		0.01	(-0.03,0.04)	0.19	(-0.84,1.22)
Regression	ACET	0.02	(-0.02,0.05)	0.18	(-0.80,1.16)
PSM	ACET	0.01	(-0.05,0.06)	0.16	(-1.55,2.31)
PSM-regression*	ACET	0.01	(-0.05,0.06)	0.12	(-1.66,2.18)
DM*	ACET	0.01	(-0.05,0.05)	0.19	(-1.66,2.08)

*Doubly robust method; **Bold** indicates doubly-matched estimators
Regression: logistic regression for the binary outcome and ordinary least squares for the continuous outcome
PSM: propensity score matched; PSM-regression: PSM followed by regression

who used e-cigarette to help them quit cigarette. The point estimations were very similar to each other, indicating that there was almost no difference between e-cigarette users and non-e-cigarette users on the long-term cigarette cessation (RD: 1-2%). The results also implied that there was no difference between the whole population and the target cohort of e-cigarette users. It would be interesting if there is a case that the treated subjects differ a lot from the population. Remember that we used simple imputation to deal with the missing covariates and we didn't include the survey weights in our estimation, which might cause differences between our example and the real situation. The bootstrap 95% quantile confidence intervals were used for the naive and regression estimates, and the full bootstrap 95% quantile confidence intervals were applied for the matching methods.

Considering the reduction in cigarette consumption, again, all the comparison estimators including the naive estimator suggested a positive but non-significant effect for use of e-cigarettes on the reduction of cigarette consumption among those who were likely to use e-cigarette to help them quit. The risk differences were far from significant. For both of the two outcomes, the regression method was the most efficient one, and the PSM-regression and DM estimators performed similarly and gave slightly narrower 95% CI than the PSM estimator. As before, no significant result was found assessing e-cigarette use on the cigarette consumption reduction in this cohort.

3.7 Discussion

In this chapter we studied the performance of a doubly-matched DR estimator based on matching both the PS score and the PGS score simultaneously for estimation of the ACET. We compared performance of the doubly-matched estimator of the ACET to a more usual regression-based DR estimator, PSM-OLS (PSM-regression more generally), and the commonly-used OLS and PSM estimators. We compared both point estimators and methods for confidence interval estimation.

In our simulation studies, the two DR approaches performed well when both the PS the PGS models are consistent, paying only a very small to a modest penalty in efficiency for their double-robustness. As expected these DR estimators vastly outperformed non-DR approaches (the OLS and PSM estimators) when either the PS or PGS model are incorrectly specified. In our simulated data, one of our covariates is omitted in predicting the PS, compared to the situation where all covariates are simulated to be related to the PGS, thus the PSM is less efficient than the OLS estimator when both of the models are correct. Importantly, an interaction between one of the covariates and the treatment indicator has been added, which makes the ACET differ from the ACE (the difference is around 15% to 25% under different scenarios).

Comparing the DR doubly-matched estimator, DM, to the DR regression-based approach PSM-OLS, when exact matching is possible (e.g. with sparse categorical predictors, a case which is not uncommon in practice), in our examples the DM outperforms the PSM-OLS estimator substantially in cases where the PGS model is incorrect but the matching is correct. In addition, the DM estimator is as good as the PSM-OLS estimator when the PGS model is correct, no matter whether we use exact matching or nearest neighbor matching. We use 1:1 matching; performance would be improved even further with 1: m matching in this setting of exact matching, where there is no bias penalty to pay.

An interesting finding in our study is the extra robustness to model mis-specification that is possible for matching estimators. In certain circumstances, exact matching is still possible even though the distance measure in the PGS or PS model is incorrect. This allows for extra robustness in the incorrect model specification case for matching estimators as compared to their regression-based counterparts. Methods which are even more independent of the functional form of these models have been proposed [36] and are worth further study. However, there is also the trade-off for the matching estimators when a non-negligible proportion of treated subjects needs to be discarded after matching, which is not the case in our study but is a realistic circumstance.

When considering methods for confidence interval construction, the details of the matching algorithm are important. For matching without replacement, we recommend constructing a confidence interval conditional on the matched sample. In our examples, both the conditional bootstrap (which resamples matched pairs) and a conditional parametric approach (e.g. maximum likelihood conditional on the matched pairs) worked well, performed similarly to each other, and have a large computational advantage over full bootstrap approaches. Importantly, as long as the estimator is correctly estimated, the conditional methods will work. This is consistent with the recent publication of Abadie [5], which shows that the conditional methods work well for inference regarding regression coefficients in the PSM-OLS estimator (without replacement), no matter whether the post PSM regression model is correct or not. We have also shown that the conditional methods work for the DM estimator for estimating the ACET. In addition, we have shown that when the matching is incorrect but the regression is correctly specified, the conditional methods still support correct inference for the PSM-OLS and DM estimators.

Considering matching with replacement, theory suggests and our simulations confirm that it has higher variance than matching without replacement. In case of large covariate imbalance between treatment groups, matching with replacement may reduce bias, however. For estimators incorporating matching with replacement, conditional methods of confidence interval estimation are not recommended. Both the conditional bootstrap and conditional parametric approaches encountered under coverage in our simulation scenarios. As the replication rate increases, the under coverage becomes more severe. The full bootstrap had conservative coverage for some of the matching estimators, and it became more conservative when matching with replacement, consistent with previous findings [34, 12]. The full bootstrap can be considered for use, as Abadie [4] shows that matching on the estimated PS adds a non-positive adjustment factor to the asymptotic variance of the estimator which matches on the true PS. This lends theoretical support to the idea that the full bootstrap will always be conservative for matching with replacement. As a correction, Bodory [15] proposes the wild bootstrap, in which the sample covariates are fixed, and within each bootstrap sample the random treatment indicators are generated based on the sample estimated PS, followed by re-estimating the PS within the bootstrap sample. The final bootstrap approximation is constructed by perturbating the martingale representation for matching estimators using those re-estimated PS within each bootstrap sample. They show that the wild bootstrap comes closer to the nominal size than the full bootstrap for matching with replacement. Future work in this area will be to find a similar adjusted approach for the DM estimator.

Our application of these estimators to data from the PATH study suggests that, among those who used them, e-cigarette use for smoking cessation is not causally related to either reduction in consumption of cigarettes or to long-term abstinence from smoking in the study cohort. However, this is a controversial public policy topic and we did not weight our estimates to the US population or conduct the kind of robust sensitivity analyses we have used in our prior published work on this topic. The missing data in covariates are based on simple imputation here also. Results are not significant in this demonstration study, which used conservative full bootstrap confidence intervals. We have noticed that the ACET didn't differ from the ACE in this study, however, there is the possibility that the estimate of the ACET differs in direction among the treated and untreated groups, an effect which cannot be detected without using appropriate methods such as we employ here. In practice, we recommend that several robust and non-parametric approaches to causal inference should be routinely incorporated, as robust sensitivity analyses.

Finally, our studies in this chapter point the way to future work. PSM is often used with survey data, as in our example, but there are many details regarding how to properly incorporate the survey weights in matched estimators. [52] These approaches may potentially be extended to doubly-matched estimators. The question of how to obtain appropriate confidence intervals for matched estimators using complex survey data remains open.

3.8 Afterthoughts before next Chapter

In this chapter, we introduced and explicitly studied a DM estimator which can correctly estimate the ACET even when the PS is incorrectly specified, in order to make DR causal inference. Using a comprehensive simulation study, the performance of the proposed DM estimator was compared with other commonly-used estimators, as well as with second DR estimator. We also studied several methods for variance estimation, including a parametric method, the full bootstrap, and the conditional bootstrap. These variance estimators were investigated using a simulation study under the simple random sampling design. However, how to estimate the variance of the matching estimator with more complex designs such as complex surveys requires further discussion.

In the next chapter, we will explore variance estimation of a matching estimator, taking the PSM estimator as an example, in complex surveys with survey weights. We consider some existing methods including the jackknife method, balanced repeated replicate (BRR), Fay's method, and the bootstrap approaches. We will explore the large sample properties of the jackknife method as well as the BRR and investigate the performance of these methods compared to the bootstrap methods. PATH survey study data will be used as an example.

This chapter, in full, has been prepared for submission for publication as ”Chen, Ruifeng; Messer, Karen S. *Doubly robust causal inference using the doubly-matched estimator, with application to the causal effect of e-cigarette use on smoking cessation and cigarette consumption reduction*”. The dissertation author was the primary author on this paper.

Chapter 4

Large sample properties of variance estimation for matching estimators in complex surveys

4.1 Abstract

To study complex causal questions in tobacco control, a large-scale population data is needed which is often sampled from a complex survey design. Besides simple random sampling, stratified sampling and cluster sampling are also taken into account in survey designs. Causal inference made from such survey designs needs to consider the survey weights which are attached to survey subjects to account for different survey design such as oversampling of certain subpopulation. Only in this way estimates of population parameters of interest can be obtained. Propensity score matching (PSM), proposed by Rosenbaum and Rubin in 1983, or more general matching methods, are a commonly-used approach in causal inference. PSM has the advantage of easy interpretation and it can be used to avoid extrapolation. There are many examples in the literature using PSM to make inference in the setting of complex survey designs. Although

previous studies have shown how to construct consistent point estimators using PSM with survey data, the variance estimation for the PSM estimator still remains as an open research area. In our study, we want to fill this gap. Specifically, we provide the large sample properties of the jackknife variance estimate and the balanced repeated replication variance estimate. The results can be extended to the Fay's method, an improvement of the BRR method. We use simulation to assess the performance of the commonly-used BRR method and the Fay's method, and compare them to the full bootstrap, as well as the conditional bootstrap, which bootstraps the matched pairs directly from the original sample. Our simulation results show that the BRR and the Fay's methods consistently perform well when the number of primary sampling units (PSU) is 2. When we have a large number of PSU, both the adjusted BRR and the adjusted Fay's method (creating pseudo sub-strata followed by using each of these two methods) and the two bootstrap methods work well. The full bootstrap and the conditional bootstrap work similarly when matching without replacement. The variance estimators are applied in a tobacco control case study using a complex survey where we end up with a consistent conclusion that using counseling or self-help materials among smokers who want to quit cigarette smoking didn't help them reduce cigarette consumption in the long-term.

4.2 Introduction

In observational studies, when making causal inference, it is important that the sample we collect is representative of the target population that we are interested in so that we can make correct inference about the entire target population. Different designs for sampling from the target population include simple random sampling (SRS), stratified sampling and cluster sampling. [96] Stratified sampling generally reduces the sample variability as it reduces the within strata variation. In addition, it is more convenient to conduct and administer in practice. Another sampling approach, cluster sampling, often takes the benefits of making sampling process easier and cost less. [57] A large-scale representative sampling design often needs to combine several of these sampling approaches. The large-scale complex survey is widely-used to collect sample from the target population in practice. The complex survey often assembles all of SRS, stratified sampling and cluster sampling together into the final sampling design. [57] Specifically, it stratifies the population into different strata, followed by constructing and sampling clusters within each stratum. Finally individuals are drawn using SRS within the sample clusters. For instance, the well-known PATH tobacco survey study [38] used a stratified multistage sampling to make survey design. It stratified the US population to 92 strata within which 156 clusters were sampled. Many strata included 2 sample clusters. Individuals were finally drawn within sample clusters. Causal inference made from such a complex survey design needs to account for the survey weights, which are attached to survey samples to account for specific survey design such as oversampling of certain sub-population to obtain estimates of population parameters of interest, as well as other survey design elements. [57, 44, 91, 56] The parameters which we are interested in will not be consistent with the parameters in the full population without considering survey weights due to selection bias. The survey weights themselves are derived using the inverse probability weighting approach, considering the sampling design information as well as other factors such as non-response. [55]

After the sample is collected, different approaches can be used for estimation and inference. Based on Rubin's potential outcome causal inference framework, [78] Rosenbaum and Rubin [74] introduced the propensity score (PS), defined as the probability of being assigned to the treatment group conditional on baseline covariates, assuming there are no unmeasured confounders. Since then, different PS-related methods have been developed for causal inference, including propensity score matching (PSM) [74, 39], propensity score stratification, [75] and inverse probability weighting (IPW) [72]. These PS based methods can be compared with regression modeling approaches to make causal inference from the data.

Propensity score matching and similar matching methods are frequently used in multiple disciplines including statistics, economics, epidemiology, sociology, and etc. [39, 79, 88, 17, 62] Compared to other methods, matching has the benefit of easy interpretation and simple visualization of the matching results. Also, it can avoid extrapolation in certain circumstances when we have a sub-population of interest, for example for estimating the average causal effect among the treated (ACET). Austin [11] and Lenis [52] have extended PSM into the complex survey setting and have shown how to construct consistent PSM point estimators. However, to our knowledge, the variance estimation of the PSM estimator remains as an open research area.

4.2.1 Motivating study

In our previous studies in the field of tobacco control, in which we investigated whether e-cigarette use helped those who wanted to quit cigarette successfully quit cigarette smoking, [66, 20, 19] we used caliper nearest neighbor PSM (with the caliper set to 0.1) to match a non-cigarette user to each of those who used e-cigarettes to help them quit. The matched pairs were used to estimate the mean difference between the two groups of interest in the cessation rate 2 years post- baseline. In detail, in the 3 separate tobacco control studies, we got 2443, 2535 and 3578 (with additional refreshed sample) baseline smokers respectively who had made at least one quit attempt in the follow-up year and had completed survey assessment 1 or 2 years later. In those studies, around 13% to 23% of participants had used e-cigarette to help them quit during their last quit attempt. The PS was consistently estimated based on a lasso -selected set of covariates, from among around 20 initial baseline covariates, including age, sex, ethnicity, race, education, income, nicotine dependence score, baseline cigarette consumption, duration of last quit attempt, timing from last quit attempt, smoking-related health problems, smoking pack-years, age started smoking fairly regularly earlier than 18, perceived harm of cigarettes, self-efficacy about quitting cigarettes, interest in quitting cigarettes, smoking-free home, exposure to other smokers, internal/ external mental health problems and health insurance status. Logistic regression models were used to estimate the PS, followed by PSM. The difference in the weighted cessation rates were then calculated for the matched samples.

To assess whether the difference in cessation rates achieved statistical significance, we used the bootstrap to estimate the variance of the PSM estimator. Specifically, we generated substantial numbers of bootstrap samples in each of the studies and the bootstrap sample was generated by re-sampling the survey individuals. In each bootstrap sample, we re-estimated the PS and re-conducted the PSM to form bootstrap matched pairs. Ninety-five percent bootstrap quantile confidence intervals were used to make inferences. Using the bootstrap method across 3 studies resulted in a consistent conclusion that e-cigarette use didn't help people who wanted to quit cigarettes successfully quit cigarette smoking in the long-term.

The bootstrap method that we used, which re-samples the survey individuals, is commonly used in practice. Several studies have used it to draw their conclusions. [61, 94] Others have also used parametric approaches to estimate the variance on the post-matching sample. [23, 47] However, to our knowledge, so far there's no theoretical discussion about how to make variance estimation of the matching estimators for complex survey data along with survey weights. This motivates the present study which explores how to construct consistent estimators of the PSM estimator variance.

4.2.2 Literature review

To date, there does not appear to be a gold standard approach to variance estimation for matching estimators, even with the SRS design. Matching is complex and can be with or without replacement and different variance estimation methods behave differently with different matching designs. The true variance of the PSM estimator is also different depending on matching either on the estimated PS or on the true PS. [81, 4, 71, 80] With the SRS design, Hill et al. [34] have shown that the conditional parametric approach, which focuses on the variance of the difference of matched pairs, works well for scenarios where the 2 groups are similar to each other, or for matching without replacement. However, in most other cases it underestimates the variance for matching with replacement. Austin and Small [12] have shown by simulation that the conditional bootstrap method, which bootstraps the matched pairs to form the bootstrap samples and neither re-estimates the PS nor re-conducts the PSM, performs well when matching is without replacement. The full bootstrap method appears to be conservative in their simulation. In addition, the results of their conditional bootstrap approach are quite similar to the results of the conditional parametric approach. When matching with replacement, the conditional bootstrap usually underestimates the variance, as does the conditional parametric method. In addition to the conditional parametric methods and the bootstrap methods, Ho et al. [35] have studied the variance derived from a post-matching regression model. Abadie and Spiess [5] have extended Ho's idea to a more general setting where either the matching or the regression model could be wrong when matching without replacement and they have studied the large sample properties of the post-matching regression adjusted estimator. Even if the matching is incorrect, as long as the regression model is correct, the standard errors are asymptotically valid when matching without replacement. Abadie and Imbens have also studied large sample properties of PSM estimators when matching with replacement and proposed a method to estimate the variance in such circumstances. [1] They have also proposed a bias correction which renders general matching estimators root-n consistent. [3]

Previous work on PSM in complex surveys has been developed mainly for the point estimation. Austin et al. have constructed a comprehensive simulation study to investigate PSM estimators with survey weights. [11] Lenis et al. [52] followed Austin's simulation setup and they have demonstrated that weights are not needed in estimating the PS, and that controls, which remain in matched pairs should use the weights of the matched treated subjects. In addition, the weights must be added in the outcome analysis after matching to make a consistent estimation. In terms of general approaches to variance estimation in complex surveys, well-known methods include the Taylor expansion or linearization, the jackknife, balanced repeated replication (BRR) and bootstrap methods. Shao and Tu [84] have given a detailed investigation of these methods. Fay's method, an adjusted BRR method, [40] improves the original BRR method in certain circumstances for example in estimating ratios. [70] The PATH study used the Fay's method with the adjustment parameter $\rho = 0.3$ to form the replicate weights to reflect the variance in survey weights. The large sample properties of these variance estimation methods in the general setting have been discussed in Krewski and Rao. [45]

4.2.3 Organization of the chapter

In this chapter, we study the variance estimation for the matching estimator in complex survey data. Specifically, we investigate the following variance estimation methods: the jackknife, the BRR, the Fay's method, the full bootstrap and the conditional bootstrap. We provide the explicit proof of the large sample properties of the jackknife method and the BRR method for matching estimators with data from a complex survey sampling design. These two methods are equivalent when the estimator is a linear transformation of the outcomes and the number of sample clusters is 2 across all strata. Fay's method is commonly-used in practice in survey data and the asymptotic consistency can be extended to the case where the Fay's method is applied. We assess the performance of the variance estimators using both the BRR and the Fay's method using the PSM estimator as an example, and compare them with bootstrap methods, including both the full bootstrap and the conditional bootstrap.

The chapter is organized as follows: in 4.3, we introduce the framework and assumptions as well as the variance estimation methods, and we provide the proof of the consistency of the variance estimation by the jackknife and the BRR methods. Consistency can be extended to include Fay's method. In 4.4, we follow Austin's simulation setup [11] to compare the results of variance estimation by the BRR and the Fay's methods and the two bootstrap methods. In 4.5, we apply the variance estimation methods to a case study using The Population Assessment of Tobacco and Health (PATH) survey data [38] to assess whether accepting counseling or self-help materials helped smokers reduce cigarette consumption in the long-term. In 4.6, we summarize the results and interesting findings and discuss potential directions of the future work. Finally in

4.7, we provide the summary thoughts after the completion of this chapter.

4.3 Methods

4.3.1 Notation

We consider estimating the ACET, using data sampled from a population following a hierarchical sampling design. Suppose that the population is divided into into H strata, and within stratum $h(h = 1 \ldots H)$, there are M_h total clusters. Cluster $hl(hl = 1 \ldots M_h)$ in stratum h contains N_{hl} subjects. Let N_h denote the total number of subjects in stratum h, and N represent the total number of subjects in all strata, that is, $N_h = \sum_{l=1}^{M_h} N_{hl}$ and $N = \sum_{h=1}^{H} N_h$. Let $W_{hls} = 1/N$ denote the weight for any subject s (same for all subjects in the population), $W_{hl} = N_{hl}/N$ denote the weight for cluster hl in stratum h over the population and $W_h = N_h/N$ denote the weight for stratum h over the population.

We follow the potential outcome framework proposed by Rubin [78]. Suppose each subject in the population belongs to either the treatment group or the control group. Let R_{hls} denote the treatment group indicator for subject s in cluster hl in stratum h, where $R_{hls} = 1$ if subject hls is assigned to the treatment group and $R_{hls} = 0$ if it is assigned to the control group. Two potential outcomes $\{Y_{hls0}, Y_{hls1}\}$ could be observed for each subject, depending on which group the subject is assigned to. One and only one of the two potential outcomes is observed, and $Y_{hls} = Y_{hls0}$ if subject hls is assigned to the control group and $Y_{hls} = Y_{hls1}$ otherwise. The p dimensional covariate vector $X_{hls} = (X_{hls,1}, X_{hls,2}, \ldots, X_{hls,p})$ is fully observed for subject hls. Both the treatment indicator and the potential outcomes are associated with the same set of covariates X. The estimand of interest is the ACET which is defined as:

$$\text{ACET} = E[Y_1 - Y_0 | R = 1] \tag{4.1}$$

We are interested in how to conduct variance estimation for the matching estimator for estimating the ACET. As a special example, we consider the variance estimation of the PSM estimator. It can be easily generalized to other matching estimators.

We assume subjects are sampled as follows: first, a sample of clusters are drawn with replacement in each stratum, followed by sampling subjects with replacement within these sample clusters. Suppose m_h clusters are drawn with replacement followed by drawing n_{hi} $(hi = 1 \ldots m_h)$ subjects within these sample clusters in each stratum h. We use $y_{hij} = \{y_{hij1}, y_{hij0}\}$ to denote the observed sample subject outcome where j is used as the subject level subscript in the sample. Let $\bar{y}_{hi}$ represent the observed sample cluster mean, $\bar{y}_h$ represent the observed sample strata mean, and $\bar{y}$ represent the observed overall sample mean. Different subscripts are used in the survey sample to differentiate from subscripts used in the survey population above.

Non-response commonly occurs in survey data. Hence we consider a missing at random non-response mechanism in this study and suppose the non-response indicator Z is associated with the same set of covariates X. $Z_{hls} = 0$ if subject hls is missing and $Z_{hls} = 1$ otherwise. $W_h^{'}$, $W_{hi}^{'}$ and $W_{hij}^{'}$ are used to denote the weights for the sample stratum, the sample cluster and the sample subject respectively.

4.3.2 Assumptions for large sample properties

In the following sections, in order to show the large sample properties, we assume both the number of strata H and the number of sample subjects n_{hi} in any sample cluster hi go to infinity. We put additional assumption on the sample cluster size n_{hi} that we assume $c_1 s_H < n_{hi} < c_2 s_H$ as $s_H \rightarrow \infty$. In addition, we assume the number of sample clusters m_h within any stratum h is fixed and there are at least 2 clusters sampled in any stratum, that is, $2 \leq m_h \leq c_0$, where c_0 is a finite positive number.

Throughout, we assume the sampling probability of the cluster is proportional to the cluster size, thus the weights of sample subjects are proportional to the weights of subjects in the population, without considering the non-response mechanism. After considering the non-response mechanism, we have that $W_{hij'} \propto W_{hij}\frac{1}{p(Z_{hij}|X_{hij})}$. As the cluster size hi goes to infinity, we also have that $W_h^{'} \propto W_h$, $W_{hi}^{'} \propto W_{hl}$ if hi and hl indicate the same cluster. Later after scaling, W_h and W_{hl} are used to represent the sample strata and the sample cluster weights respectively. For simplicity, we further assume the cluster size is the same across different clusters in later proof.

Suppose μ_{hi} is the unknown true mean of the PSM estimate in sample cluster hi and μ_h is the unknown true mean of the PSM estimator in stratum h. To the end, we are going to show the asymptotic behavior of the variance estimator. In order to do it, we suppose the variance of the observed outcome y_{hij} for any subject hij is bounded by finite constant: $c_3 < Var(y_{hij}) = \sigma_{hij}^2 < c_4$. Further, we assume that the fourth moment of $\bar{y}_{hi}$ is bounded, $\frac{c_5}{n_{hi}^2} < E\bar{y}_{hi}^4 < \frac{c_6}{n_{hi}^2}$ and the third moment of $\bar{y}_{hi}$ is also bounded, $\frac{c_7}{n_{hi}^{\frac{3}{2}}} < E[\bar{y}_{hi}] < \frac{c_8}{n_{hi}^{\frac{3}{2}}}$.

We make use of Serfling [83] in following sections to prove the consistency of the PSM estimator and the consistency of the variance estimation of the PSM estimator.

Theorem: Serfling (1.9.2 & 1.9.3) [83] has shown that: let $\{X_{nh} : 1 \leq h \leq H; n = 1, 2 \ldots\}$ be a double array with independent random variables within rows (with subscript n). Suppose that, for some $\delta > 0$, $\sum_{h=1}^{H} E|X_{nh} - \mu_{nh}|^{2+\delta} = o(A_n^{2+\delta})$, $n \longrightarrow \infty$. Then $\sum_{h=1}^{H} X_{nh}$ is $AN(\sum_{h=1}^{H} \mu_h, A_n^2)$, where $\mu_{nh} = E(X_{nh})$.

From now on we will add the subscript n to denote the index for the row of the double array. Two assumptions are needed respectively to satisfy the condition in the theorem that $\sum_{h=1}^{H} E|X_{nh} - \mu_{nh}|^{2+\delta} = o(A_n^{2+\delta})$. This condition is used in turn to show that commonly-used Lindeberg condition [83] is satisfied, and hence that consistency can be achieved. To prove the consistency of the PSM estimator, we need the following assumption 1:

Assumption 1: We assume

$$\sum_{h=1}^{H} E|(Hs_H)^{\frac{1}{2}} \frac{1}{m_{nh}} \sum_{nhi=1}^{m_{nh}} (\bar{d}_{nhi} - \mu_{nhi}) W_{nh}|^{2+\delta} \xrightarrow{p} 0$$

holds as n goes to infinity, from which the consistency of the PSM estimator in complex surveys over the entire sample will follow.

We need the following Assumption 2 to prove the consistency of the variance estimation of the PSM estimator.

Assumption 2: We assume

$$\sum_{h=1}^{H} E|(H^3 s_H^2)^{\frac{1}{2}}(\frac{1}{m_{nh}(m_{nh}-1)}\sum_{nhi=1}^{m_{nh}}((\bar{d}_{nhi}-\bar{d}_{nh})^2-(\mu_{nhi}-\mu_{nh})^2)W_{nh}^2)|^{2+\delta} \xrightarrow{p} 0$$

holds as n goes to infinity, from which will follow the consistency of the variance estimation of the PSM estimator using either the jackknife method or the BRR method in complex surveys.

4.3.3 Consistency of PSM estimators on the cluster level

We first show the consistency of the PSM estimator for estimating the ACET in complex surveys. PSM is conducted with replacement on the cluster level: the control subjects and the treated subjects can be matched only if they are in the same cluster. We start by investigating the PSM estimator on the cluster level, after which we can show the consistency of the PSM estimator over the entire sample using the asymptotic theory in double array setting in the following section.

We follow 4 causal assumptions [78, 74] to show the consistency of the point estimation: 1) The treatment must occur before observing the outcome; 2) every subject can be assigned to either the treatment group or the control group; 3) the strongly ignorable treatment assumption: $(Y_0, Y_1) \perp R|X$, that is, there are no unmeasured confounders; and 4) potential outcomes of any subject are independent of potential outcomes of other subjects. Then, conditioning on the propensity score $\pi(X)$ deconfounds the outcome Y and the treatment indicator R, [74] that is,

$$Y \perp R|\pi(X) \tag{4.2}$$

For any sample cluster *hi* in any stratum *h* in the n_{th} row of the double array, the PSM estimator can be written as:

$$\hat{D}_{nhi} = \frac{1}{\sum_{j=1}^{n_{nhi}} R_{nhij} \frac{W_{nhij}}{p(Z_{nhij}|X_{nhij})}} \sum_{j=1}^{n_{nhi}} R_{nhij}(Y_{nhij1} - \bar{Y}_{nhij0}) \frac{W_{nhij}}{p(Z_{nhij}|X_{nhij})} \tag{4.3}$$

where $\bar{Y}_{nhij0}$ is the matched control for subject hij in the n_{th} row. To show the consistency, we'd like to show the expectation of equation 4.3 converges in probability to the true ACET. W_{nhij} are the same for all samples in the same row and could be cancelled out. After removing W_{nhij}, equation 4.3 can be split to 2 parts:

i) the weighted mean outcome of the matched treated subjects,

$$\frac{1}{\sum_{j=1}^{n_{nhi}} R_{nhij} \frac{1}{p(Z_{nhij}|X_{nhij})}} \sum_{j=1}^{n_{nhi}} R_{nhij} Y_{nhij1} \frac{1}{p(Z_{nhij}|X_{nhij})}$$

,

and ii) the weighted mean outcome of the matched control subjects,

$$-\frac{1}{\sum_{j=1}^{n_{nhi}} R_{nhij} \frac{1}{p(Z_{nhij}|X_{nhij})}} \sum_{j=1}^{n_{nhi}} R_{nhij} \bar{Y}_{nhij0} \frac{1}{p(Z_{nhij}|X_{nhij})}$$

.

For i), we can find its expectation under the regularity condition $\sum_{j=1}^{n_{nhi}} R_{nhij} \frac{1}{p(Z_{nhij}|X_{nhij})} \xrightarrow{p}$

$n_{nhi}\pi_{\frac{1}{p(Z|X)}}$ as $n_{nhi} \to \infty$, as follows:

$$
\begin{aligned}
& E\left[\frac{1}{\sum_{j=1}^{n_{nhi}} R_{nhij}\frac{1}{p(Z_{nhij}|X_{nhij})}}\sum_{j=1}^{n_{nhi}} R_{nhij}Y_{nhij1}\frac{1}{p(Z_{nhij}|X_{nhij})}\right] \\
= \quad & E\left[\frac{1}{\sum_{j=1}^{n_{nhi}} R_{nhij}\frac{1}{p(Z_{nhij}|X_{nhij})}}\sum_{j=1}^{n_{nhi}} R_{nhij}Y_{nhij1}\frac{1}{p(Z_{nhij}|X_{nhij})}|R_{nhij}=1\right] \\
= \quad & E\left[E\left[\frac{1}{\sum_{j=1}^{n_{nhi}} R_{nhij}\frac{1}{p(Z_{nhij}|X_{nhij})}}\sum_{j=1}^{n_{nhi}} R_{nhij}Y_{nhij1}\frac{1}{p(Z_{nhij}|X_{nhij})}|X_{nhij},R_{nhij}=1\right]\right] \\
\xrightarrow{p} \quad & E\left[\frac{1}{n_{nhi}\pi_{\frac{1}{p(Z|X)}}}E\left[\sum_{j=1}^{n_{nhi}} R_{nhij}Y_{nhij1}\frac{1}{p(Z_{nhij}|X_{nhij})}|X_{nhij},R_{nhij}=1\right]\right] \\
= \quad & E\left[E\left[Y_{nhij1}|X_{nhij},R_{nhij}=1\right]\right] \\
= \quad & E\left[Y_{nhij1}|R_{nhij}=1\right] \qquad (4.4)
\end{aligned}
$$

The expectation of ii) in formula (4.3) can be shown in a similar way as:

$$
\begin{array}{rcl}
ii) & = & E\left[-\frac{1}{\sum_{j=1}^{n_{nhi}} R_{nhij}\frac{1}{p(Z_{nhij}|X_{nhij})}}\sum_{j=1}^{n_{nhi}} R_{nhij}\bar{Y}_{nhij0}\frac{1}{p(Z_{nhij}|X_{nhij})}\right] \\
 & = & -E\left[\frac{1}{\sum_{j=1}^{n_{nhi}} R_{nhij}\frac{1}{p(Z_{nhij}|X_{nhij})}}\sum_{j=1}^{n_{nhi}} R_{nhij}\bar{Y}_{nhij0}\frac{1}{p(Z_{nhij}|X_{nhij})}|R_{nhij}=1\right] \\
 & = & -E\left[E\left[\frac{1}{\sum_{j=1}^{n_{nhi}} R_{nhij}\frac{1}{p(Z_{nhij}|X_{nhij})}}\sum_{j=1}^{n_{nhi}} R_{nhij}\bar{Y}_{nhij0}\frac{1}{p(Z_{nhij}|X_{nhij})}|X_{nhij},R_{nhij}=1\right]\right] \\
 & = & -E\left[E\left[\frac{1}{\sum_{j=1}^{n_{nhi}} R_{nhij}\frac{1}{p(Z_{nhij}|X_{nhij})}}\sum_{j=1}^{n_{nhi}} R_{nhij}Y_{nhij'0}\frac{1}{p(Z_{nhij}|X_{nhij})}|X_{nhij},R_{nhij}=1,\right.\right. \\
 & & \left.\left. R_{nhij'}=0,\hat{\pi}_{nhij}=\hat{\pi}_{nhij'},j'selected\right]\right] \\
 & \xrightarrow{p} & -E\left[E\left[\frac{1}{\sum_{j=1}^{n_{nhi}} R_{nhij}\frac{1}{p(Z_{nhij}|X_{nhij})}}\sum_{j=1}^{n_{nhi}} R_{nhij}Y_{nhij'0}\frac{1}{p(Z_{nhij}|X_{nhij})}|X_{nhij},R_{nhij}=1,\right.\right. \\
 & & \left.\left. R_{nhij'}=0,\pi_{nhij}=\pi_{nhij'},j'RandomlySelected\right]\right] \\
 & = & -E\left[E\left[\frac{1}{\sum_{j=1}^{n_{nhi}} R_{nhij}\frac{1}{p(Z_{nhij}|X_{nhij})}}\sum_{j=1}^{n_{nhi}} R_{nhij}Y_{nhij0}\frac{1}{p(Z_{nhij}|X_{nhij})}|X_{nhij},R_{nhij}=1]]\right.\right. \\
 & \xrightarrow{p} & -E[\frac{1}{n_{nhi}\pi_{\frac{1}{p(Z|X)}}}E[\sum_{j=1}^{n_{nhi}} R_{nhij}Y_{nhij0}\frac{1}{p(Z_{nhij}|X_{nhij})}|X_{nhij},R_{nhij}=1\Big]\Big] \\
 & = & -E[E[Y_{nhij0}|X_{nhij},R_{nhij}=1]] \\
 & = & -E[Y_{nhij0}|R_{nhij}=1]
\end{array}
\qquad (4.5)
$$

Thus we have that the PSM estimator $\hat{D}_{nhi}$, for any sample cluster hi of any stratum h in n_{th} row, is asymptotically consistent for estimating the ACET. The estimated propensity score $\hat{\pi}$ can be replaced with the estimation of other measurement scores, or estimations of multiple other measurement scores, thus the results can be generalized to any general matching estimator as long as the regularity condition $\sum_{j=1}^{n_{nhi}} R_{nhij}\frac{1}{p(Z_{nhij}|X_{nhij})} \xrightarrow{p} n_{nhi}\pi_{\frac{1}{p(Z|X)}}$ as $n_{nhi}\to\infty$, is satisfied.

4.3.4 Consistency of PSM estimators over the entire sample

In this subsection, We want to show that the PSM estimator overall the entire population in the n_{th} row of the double array, with the observed value $\bar{d}_n = \sum_{h=1}^{H} W_{nh}\bar{d}_{nh}$, where $\bar{d}_{nh} = \frac{1}{\sum_{nhi=1}^{m_{nh}}}\sum_{nhi=1}^{m_{nh}} \bar{d}_{nhi}$, is asymptotically consistent for estimating the ACET. Remember we have assumed that the cluster size is bounded by s_H, $c_1 s_H < n_{nhi} < c_2 s_H$, with which we have $M_{nh}c_1 s_H = \sum_{nhl=1}^{M_{nh}} c_1 s_H < N_{nh} = \sum_{nhl=1}^{M_{nh}} n_{nhl} < \sum_{nhl=1}^{M_{nh}} c_2 s_H = M_{nh}c_2 S_H$, so the stratum weight $W_{nh} = N_{nh}/N = N_{nh}/\sum_{h=1}^{H} N_{nh}$ is bounded, and $\frac{c_1}{c_2}H^{-1} < W_{nh} < \frac{c_2}{c_1}H^{-1}$ for any stratum h in the n_{th} row of the double array. With $H \longrightarrow \infty$, we have $W_{nh} \longrightarrow 0$.

We construct the double array variable $X_{nh} = (Hs_H)^{\frac{1}{2}}\frac{1}{m_{nh}}\sum_{hi=1}^{m_{nh}} \bar{d}_{nhi}W_{nh}$, which is the scaled weighted mean estimation for any stratum h in the n_{th} row of the double array. With the assumption that the variance of subject outcome y_{nhij} for any subject hij in the n_{th} row is bounded, we can show that the variance of the cluster mean, $Var(\bar{y}_{nhi}) = \frac{1}{n_{nhi}}Var(y_{nhi})$, is bounded, $\frac{c_3}{c_2 s_H} < Var(\bar{y}_{nhi}) < \frac{c_4}{c_1 s_H}$. Further, the variance of $\bar{d}_{nhi}$, the mean difference in cluster hi in the n_{th} row, $Var(\bar{d}_{nhi})$, is also bounded, $\frac{2c_3}{c_2 s_H} < Var(\bar{d}_{nhi}) < \frac{2c_4}{c_1 s_H}$. Then we have the variance of the constructed variable X_{nh} as following:

$$Var(X_{nh}) \quad = \quad Hs_H\frac{1}{m_{nh}^2}\sum_{nhi=1}^{m_{nh}} Var(\bar{d}_{nhi})W_{nh}^2$$

With those assumptions we have just discussed, we can derive the following:

$$\frac{2c_3c_1^2}{m_{nh}c_2^3}H^{-1} < Var(X_{nh}) < \frac{2c_4c_2^2}{m_{nh}c_1^3}H^{-1}$$

Let $A_n^2 = Var(\sum_{h=1}^{H} X_{nh})$ to denote the variance of the scaled PSM mean over the entire sample in the n_{th} row of the double array, we have:

$$\frac{2c_3c_1^2}{c_0c_2^3} < A_n^2 = Var(\sum_{nh=1}^{H} X_{nh}) < \frac{2c_4c_2^2}{2c_1^3}$$

where we use the assumption we have previous assumed in 4.3.2 that $2 \leq m_{nh} \leq c_0$. Thus we have shown A_n^2 is a constant. Following Assumption 1 in 4.3.2, we have that $\sum_{h=1}^{H} E|X_{nh} - \mu_{nh}|^{2+\delta}/A_n^{2+\delta} = \sum_{h=1}^{H} E|(Hs_H)^{\frac{1}{2}} \frac{1}{m_{nh}} \sum_{nhi=1}^{m_{nh}} (\bar{d}_{nhi} - \mu_{nhi}) W_{nh}|^{2+\delta}/(A_n^2)^{(\frac{2+\delta}{2})} \xrightarrow{p} 0$. In other words, the Lindeberg-Feller condition is satisfied. Following Serfling's theorem cited above, we have proved that $\sum_{h=1}^{H} X_{nh}$ is consistent, from which we finally have that the PSM estimator over the entire sample, scaled by both the strata size and the cluster level sample size simultaneously, is consistent as n goes to infinity: $\bar{d}_n = \sum_{h=1}^{H} W_{nh}\bar{d}_{nh} = (Hs_H)^{-\frac{1}{2}} \sum_{h=1}^{H} X_{nh}$. Further, it follows the asymptotically normally (AN) distribution shown below:

$$AN(E(\sum_{h=1}^{H} (Hs_H)^{\frac{1}{2}} \frac{1}{m_{nh}} \sum_{nhi=1}^{m_{nh}} \bar{d}_{nhi} W_{nh}), A_n^2) \tag{4.6}$$

4.3.5 Variance estimation

In the current section, we discuss the variance estimation of the PSM estimator in complex surveys. We consider commonly used variance estimation methods in survey data including the jackknife method, the BRR method, the Fay's method and the bootstrap methods. In many large scale population surveys, including the Census Bureau's Current Population Survey and the PATH study, the BRR method and Fay's method are implemented by the use of replicate weights to which re-weight original sample, either omitting (BRR) or down-weighting (Fays' method) observations which are left out of a particular re-sampled replicate sample. The main goal of our study is to investigate the large sample property of the variance estimate using these replicate weight implementations. Bootstrap methods which draw sample with replacement from the original sample are also frequently used in practice for estimating the variance in complex surveys. We consider two bootstrap methods in this study. One is the full bootstrap, in which we bootstrap the survey subjects followed by re-estimating the PS and re-conducting the PSM. The other approache is the conditional bootstrap, in which we bootstrap the PSM matched pairs in the original sample and we neither re-estimate the PS nor re-conduct the PSM. Details are introduced in later sections.

The jackknife method

The jackknife method [25] involves a leave-one-out strategy to create replicate samples. As originally proposed it studies the variability of replicates which leave one subject out with the SRS design, and it can be extended to other general primary sampling units (PSU) such as groups or clusters. In our setting, for each stratum in the sample, we leave one of the clusters *hi* out to form a replicate sample and calculate the variability of estimates across all replicate samples. We consider a general form of the jackknife estimate, which is written as follows:

$$\hat{Var}_{jackknife}(\bar{d}) \quad = \quad \sum_{h=1}^{H} \frac{m_h - 1}{m_h} \sum_{hi=1}^{m_h} (\bar{d}^{(hi)} - \bar{d})^2 \tag{4.7}$$

where $\bar{d}^{(hi)}$ is the PSM estimate leaving cluster *hi* in stratum h out.

The BRR method

The BRR method is a special case of the jackknife method, constructed in the special case where the number of PSU $m_h = 2$ for all strata. This is a commonly used method of variance estimation in large scale surveys which use resampling-based estimators. Instead of leaving one sample cluster in each stratum out, the BRR leaves half of the sample clusters out (select one out of two sample clusters in all strata) each time to create replicates. Suppose there are R_W replicates in total. These replicates need to satisfy the following condition:

$$\sum_{r_w=1}^{R_W} \beta_{r_w h_1} \beta_{r_w h_2} \quad = \quad 0 \quad \text{for all} \quad h_1 \neq h_2 \tag{4.8}$$

where R_W is the total number of replicate samples and $\beta_{r_w.}$ is a vector with the length the same as the number of strata, and taking elements either 1 or -1 to indicate whether the first sample cluster is selected in the current replicate or not. h_1 and h_2 represent any two strata. With these constructed replicates, the original BRR estimate can be written as:

$$\hat{Var}_{BRR}(\bar{d}) \quad = \quad \sum_{r_w=1}^{R_W} \frac{(\bar{d}^{(r_w)} - \bar{d})^2}{R_W} \tag{4.9}$$

where $\bar{d}^{(r_w)}$ is the PSM estimate using the r_w replicate weights.

Krewski and Rao [46] have shown that when the estimator is a linear transformation of the outcomes, the jackknife estimate can be re-written to the following form:

$$\hat{Var}_{rep}(\bar{d}) \quad = \quad \sum_{h=1}^{H} W_h^2 \frac{1}{m_h} \frac{\sum_{hi=1}^{m_h} (\bar{d}_{hi} - \bar{d}_h)^2}{(m_h - 1)} \tag{4.10}$$

Furthermore, when the number of sample cluster is 2 across all strata, the BRR estimate reduces to the same form in 4.10. In our case considering the variance estimation of the PSM estimator when, the condition about the linear transformation of the outcomes is satisfied. Later, we will study the large sample properties of $\hat{Var}_{rep}(\bar{d})$.

Fay's method

Rao and Shao [70] have noted that the BRR method might lead to problems in certain situations for example for estimating ratios, as it puts 0 weights on half of the sample clusters in each replicate sample. Fay's method is an alternative which makes improvement in these circumstances and is used in several large population surveys, including the Current Population Survey and the PATH study. Specifically, Fay's method introduces an adjustment parameter ρ on top of the BRR method by oversampling one of the two sample clusters in each stratum. Compared to the BRR method which assigns weights 0 and 1 to the two sample clusters in each stratum in each replicate, the Fay's method assigns weights ρ and $1 - \rho$ to the two sample clusters, with $0 < \rho < 1$. In other words, none of the two sample clusters is removed in any replicate. The variance estimate using the Fay's method can be shown as:

$$\hat{Var}_{Fay}(\bar{d}) \;=\; \sum_{r_w=1}^{R_W} \frac{(\bar{d}^{(r_w)} - \bar{d})^2}{R_W}(1-\rho)^2 \tag{4.11}$$

where ρ is the adjustment parameter ranging from 0 to 1. In the PSM scenario, the Krewski's form $Var_{rep}(\bar{d})$ can be modified to the variance estimation using the Fay's method as well.

The full bootstrap

The full bootstrap method constructs bootstrap samples with replacement from the original sample in two steps. In the first step, we bootstrap clusters within each stratum. After that in the second step, we bootstrap individual subjects within each bootstrapped cluster. The PS is re-estimated for each bootstrap sample followed by re-conducting the PSM using the re-estimated PS. Again, the PSM is conducted within each cluster: only those in the same cluster can be matched to each other. Finally in each bootstrap sample, we calculate the weighted mean difference based on which the full bootstrap variance is estimated.

The conditional bootstrap

The conditional bootstrap is conditional on the matched pairs in the original sample. Similarly, it is implemented in two steps. The first step is the same as before in which we bootstrap clusters within each stratum. In the second step, we bootstrap the matched pairs that matched with replacement in the original sample in each bootstrap cluster. Notice that we neither re-estimate the PS, nor do we re-conduct the PSM. We calculate the weighted mean difference of these bootstrap samples and finally estimate the conditional bootstrap variance.

4.3.6 Consistency of variance estimates

To prove the consistency of variance estimations using either the jackknife method or the BRR method, $\hat{Var}_{rep}(\bar{d})$, we construct another variable $Z_{nh} = (H^3 s_H^2)^{\frac{1}{2}} \frac{1}{m_{nh}} \frac{1}{m_{nh}-1} \sum_{nhi=1}^{m_{nh}} (\bar{d}_{nhi} - \bar{d}_{nh})^2 W_{nh}^2$, which is the scaled variance estimate for stratum h in the n_{th} row of the double array. We again follow Serfling's theory to prove the consistency. First, we calculate the variance of the constructed variable as following:

$$
\begin{aligned}
Var(Z_{nh}) &= (H^3 s_H^2) \frac{1}{m_{nh}^2 (m_{nh}-1)^2} \sum_{nhi=1}^{m_{nh}} W_{nh}^4 Var((\bar{d}_{nhi} - \bar{d}_{nh})^2) \\
&= (H^3 s_H^2) \frac{1}{m_{nh}^2 (m_{nh}-1)^2} \sum_{nhi=1}^{m_{nh}} W_{nh}^4 Var(\bar{d}_{nhi}^2 - 2\bar{d}_{nhi}\bar{d}_{nh} + \bar{d}_{nh}^2) \\
&= (H^3 s_H^2) \frac{1}{m_{nh}^2 (m_{nh}-1)^2} \sum_{nhi=1}^{m_{nh}} W_{nh}^4 E(\bar{d}_{nhi}^2 - 2\bar{d}_{nhi}\bar{d}_{nh} + \bar{d}_{nh}^2 - E(\bar{d}_{nhi}^2) \\
&\quad +2E(\bar{d}_{nhi}\bar{d}_{nh}) - E(\bar{d}_{nh}^2))^2 \qquad (4.12)
\end{aligned}
$$

We have that $E\bar{d}_{nh}^2 = E(\frac{1}{m_{nh}} \sum_{nhi=1}^{m_{nh}} \bar{d}_{nhi})^2$, and $E\bar{d}_{nhi}\bar{d}_{nh} = E(\bar{d}_{nhi} \frac{1}{m_{nh}} \sum_{nhi=1}^{m_{nh}} \bar{d}_{nhi})$. Remember previously we have assumed that both the fourth moment of $\bar{y}_{nhi}$, and the third moment of $\bar{y}_{nhi}$ are bounded, and $\bar{d}_{nhi}$ is a linear transformation of $\bar{y}_{nhi}$. We finally can get that $E(\bar{d}_{nhi}^2 - 2\bar{d}_{nhi}\bar{d}_{nh} + \bar{d}_{nh}^2 - E(\bar{d}_{nhi}^2) + 2E(\bar{d}_{nhi}\bar{d}_{nh}) - E(\bar{d}_{nh}^2))^2$ is bounded:

$$\frac{c_9}{s_H^2} < E(\bar{d}_{nhi}^2 - 2\bar{d}_{nhi}\bar{d}_{nh} + \bar{d}_{nh}^2 - E(\bar{d}_{nhi}^2) + 2E(\bar{d}_{nhi}\bar{d}_{nh}) - E(\bar{d}_{nh}^2))^2 < \frac{c_{10}}{s_H^2}$$

with which we derive that:

$$\frac{c_1^4 c_9}{m_{nh}(m_{nh}-1)^2 c_2^4} H^{-1} < Var(Z_{nh}) < \frac{c_2^4 c_{10}}{m_{nh}(m_{nh}-1)^2 c_1^4} H^{-1}$$

Similarly, let $B_n^2 = Var(\sum_{nh=1}^{H} Z_{nh})$, we have:

$$\frac{c_1^4 c_9}{c_0(c_0-1)^2 c_2^4} < B_n^2 < \frac{c_2^4 c_{10}}{2c_1^4}$$

where we use the assumption that $2 \leq m_{nh} \leq c_0$. Thus B_n^2 is a constant.

Following Serfling's theorem in 4.3.4 and assuming Assumption 2 there holds, that is, $\sum_{h=1}^{H} E|(H^3 s_H^2)^{\frac{1}{2}}(\frac{1}{m_{nh}(m_{nh}-1)}\sum_{nhi=1}^{m_{nh}}((\bar{d}_{nhi}-\bar{d}_{nh})^2-(\mu_{nhi}-\mu_{nh})^2)W_{nh}^2)|^{2+\delta} \xrightarrow{p} 0$, then we derive that $\sum_{h=1}^{H}(H^3 s_H^2)^{\frac{1}{2}}(\frac{1}{m_{nh}(m_{nh}-1)}\sum_{nhi=1}^{m_{nh}}(\bar{d}_{nhi}-\bar{d}_{nh})^2 W_{nh}^2)$, which is the variance estimate by either the jackknife or the BRR over the entire sample, scaled by $(H^3 s_H^2)^{\frac{1}{2}}$, is consistent and follows the AN distribution as shown below:

$$AN(E(\sum_{h=1}^{H}(H^3 s_H^2)^{\frac{1}{2}}\frac{1}{m_{nh}(m_{nh}-1)}\sum_{nhi=1}^{m_{nh}}(\bar{d}_{nhi}-\bar{d}_{nh})^2 W_{nh}^2), B_n^2) \tag{4.13}$$

4.4 Simulation study

4.4.1 Population generating model

We use simulation to study the performance of the methods listed in 4.3 to estimate the variance of the PSM estimator in complex surveys. We have shown that the jackknife and the BRR estimators can be reduced to the same form $\hat{Var}_{rep}(\bar{d})$ when the number of sample clusters is 2 across all strata; thus we only need to show the performance of one of them. The jackknife estimate costs a substantially longer time since it creates more replicates. Furthermore, Fay's method works similar to the BRR method and performs better in certain situations, thus we also consider this estimator. The full bootstrap and the conditional bootstrap are included as comparison methods. The simulation setup closely follows Austin. [11]

To generate the population, we consider splitting the entire population into 4 strata. Within each stratum, we simulate 20 iid clusters and we generate 5000 iid subjects within each cluster. One population is generated. In total, the simulated population has 400,000 subjects. We generate 6 normally distributed covariates $X_{hls,k}$ ($k = 1,2,\ldots 6$) for each subject hls in the population. We assume that there are no unmeasured covariates. These 6 covariates are simulated from different normal distributions to reflect the fact that different strata and different clusters can have subjects with different characteristics. In detail, for each covariate, the stratum mean is simulated by $\mu_{h,k}^{stratum} \sim N(0, \tau^{stratum})(h = 1\ldots 4; k = 1\ldots 6)$, where $\tau^{stratum}$ is a fixed value reflecting the variability across strata. We will compare scenarios with different values of $\tau^{stratum}$ in 4.4.5. The cluster-specific covariate mean is simulated by $\mu_{hl,k}^{cluster} \sim N(0, \tau^{cluster})(hl = stratum1, cluster1; \ldots stratum4, cluster20; k = 1\ldots 6)$. Similarly, $\tau^{cluster}$ is fixed and we will compare scenarios with different values of $\tau^{cluster}$ in 4.4.5. For different strata, the cluster specific covariate means are different. After we simulate both the stratum mean and the cluster mean, the subject covariate is simulated by $X_{hls,k} \sim N(\mu_{h,k}^{stratum} + \mu_{hl,k}^{cluster}, \tau^{covariates})(k = 1,2,\ldots 6)$, for each of the 6 covariates. $\tau^{covariates}$ is fixed, reflecting the variance of the covariates within each cluster.

The treatment indicator R is associated with all 6 covariates X as follows:

$$logit(p_{hls}^{r}) = a_0 + a_1 x_{hls,1} + a_2 x_{hls,2} + a_3 x_{hls,3} + a_4 x_{hls,4} + a_5 x_{hls,5} + a_6 x_{hls,6} \quad (4.14)$$

After we construct the true probability for each subject of being assigned to the treatment group, the binary variable treatment indicator is simulated from the Bernoulli distribution: $r_{hls} \sim Bernoulli(p_{hls}^{r})$. We chose the coefficients to make the proportion of treated subjects around 30%, namely $a_0 = log(0.43)$, $a_1 = log(5)$, $a_2 = log(6)$, $a_3 = log(7)$, $a_4 = log(8)$, $a_5 = log(9)$, $a_6 = log(10)$.

The non-response mechanism is also considered in the simulation study. The non-response indicator is assumed to be associated with the same set of covariates $X_{hls,k}(k = 1,2,\ldots 6)$. It is represented by the variable Z, where $Z = 1$ indicates a response subject and $Z = 0$ represents a non-response or dropout subject. Similar to the treatment indicator, we generate the true probability of being a response using the following relationship:

$$logit(p^z_{hls}) \quad = \quad b_0 + b_1 x_{hls,1} + b_2 x_{hls,2} + b_3 x_{hls,3} + b_4 x_{hls,4} + b_5 x_{hls,5} + b_6 x_{hls,6} \qquad (4.15)$$

and we generate the non-response indicator for each subject *hls* from $z_{hls} \sim Bernoulli(p^z_{hls})$. The coefficients in 4.15 are set as $b_0 = -log(0.11)$, $b_1 = -log(1.10)$, $b_2 = -log(1.25)$, $b_3 = -log(1.50)$, $b_4 = -log(1.75)$, $b_5 = -log(2.00)$, $b_6 = -log(3.50)$. With these given coefficients, the non-response rate is around 10%.

Finally, we simulate the subject outcomes. For the potential outcome of the control for subject *hls*, we used the following linear model to generate y_{hls0}:

$$y_{hls0} \quad = \quad c_0 + c_1 x_{hls,1} + c_2 x_{hls,2} + c_3 x_{hls,3} + c_4 x_{hls,4} + c_5 x_{hls,5} + c_6 x_{hls,6} + \varepsilon_{hls} \qquad (4.16)$$

where we set $c_0 = 1$, $c_1 = 2.5$, $c_2 = -2$, $c_3 = 1.75$, $c_4 = -1.25$, $c_5 = 1.5$, $c_6 = 1.1$ and generate $\varepsilon_{hls} \sim N(0, \sigma^{outcome})$. $\sigma^{outcome}$ is fixed reflecting the variability of potential outcomes of the control. The potential outcome of the treated is simulated to be associated with the same set of covariates $X_{hls,k} (k = 1, 2, \ldots 6)$. The overall treatment effect is set to 1. In addition, $y_{hls,1}$ is also associated with interactions between each of the covariates $x_{hls,1}$, $x_{hls,2}$, $x_{hls,3}$ and the treatment indicator r_{hls}, as well as a cluster specific treatment effect Δ_{hl}. These interactions are added to allow the ACET to differ from the average causal effect on the population (ACE), as in this case the PSM estimator could help avoid extrapolating the results of the target population to a broader population. The cluster specific treatment effect is added to mimic the case where the weights really need to be taken into account, that is, where the sample ACET will differ from the population ACET. Finally, the true potential outcome of the treated for each subject *hls* is simulated as following:

$$\begin{aligned} y_{hls1} \quad = \quad & c_0 + c_1 x_{hls,1} + c_2 x_{hls,2} + c_3 x_{hls,3} + c_4 x_{hls,4} + c_5 x_{hls,5} + c_6 x_{hls,6} + \varepsilon_{hls} \\ & \Delta_{treatment} + c_{1r} x_{hls,1} + c_{2r} x_{hls,2} + c_{3r} x_{hls,3} + \Delta_{cluster_{hl}} \end{aligned} \qquad (4.17)$$

where we set $\Delta_{treatment} = 1$, $c_{1r} = 10$, $c_{2r} = -8$ and $c_{1r} = 7$. The cluster specific treatment effect is generated from the normal distribution $\Delta_{cluster_{hl}} \sim N(0, \sigma^{cluster})$. After the two potential outcomes are generated for each subject, the observed outcome is given by $y_{hls} = r_{hls}y_{hls1} + (1 - r_{hls})y_{hls0}$. With this setup, we have that the true ACET in the entire population is given by $1 + 10E(X_1|R = 1) - 8E(X_2|R = 1) + 7E(X_3|R = 1) + E(\Delta_{cluster}|R = 1)$. Using this relation, the ACET can be empirically estimated from the simulated population. A single generated population is fixed for the following simulation components.

4.4.2 Sampling schema

After the fixed population is generated, we use Monte Carlo simulation to assess the performance of the variance estimation methods. In each Monte Carlo simulation, we first draw a random sample from each stratum in the population: in each stratum, we randomly draw 2 out of 20 clusters with replacement. Then, in each of the sample clusters, we sample an equal sample size of 100 subjects, with replacement. Thus, in total we draw $4 * 2 * 100 = 800$ subjects in each simulation sample. The initial weight before considering non-responses for sample subject is calculated as $w_{hij}^{initial} = \frac{1}{(\frac{2}{20} * \frac{100}{5000})}$. As mentioned, we also consider the missing at random non-response mechanism in our simulation. Using the non-response mechanism introduced in the previous section, we have $\sim 10\%$ non-response subjects in the sample. We estimate the predicted value of being a response $\hat{p}_{hij}^{z}$ by modeling the binary indicator r_{hij} in the simulation sample on all 6 covariates $x_{hij,k}$. The final weights for each sample subject after considering the non-response mechanism is given by $w_{hij}^{'} = w_{hij}^{initial} \frac{1}{\hat{p}_{hij}^{z}}$.

4.4.3 Propensity score matching

For each Monte Carlo sample, we estimate the unweighted PS using the logistic regression fitting the treatment indicator on the 6 covariates X_1 to X_6. Survey weights are not included in the logistic regression model to estimate the PS. [52] After we obtain the estimated PS in each simulation sample, we use it to conduct PSM. The matching is conducted on the cluster level, that is, for any treated subject, the match can only be implemented when the control is in the same cluster. Since we consider the continuous covariates case, nearest neighbor caliper PSM is conducted, using a single matched control subject without replacement, [10] with caliper set to 0.1. For each treated subject, we match it with a control in the same cluster whose PS is closest and differs no more than 0.1 to the PS of the treated subject. We do this matching for each treated subject in the sample. Finally, we calculate the weighted mean difference between the matched pairs to derive the mean estimate of ACET using the PSM estimator. Depending on the simulation scenarios, generally, not all treated subjects will successfully be matched with controls each time, thus the true estimand in practice is actually the expected ACET among those treated subjects conditional on existence of a successful match.

4.4.4 Variance estimation

We considered 4 variance estimators, namely the BRR, the Fay's method, the full bootstrap and the conditional bootstrap respectively, to estimate the variance of the PSM estimator of the ACET. We empirically estimated the true variance of the PSM estimator directly from the Monte Carlo simulation, that is, as the variance of the PSM estimator across the Monte Carlo samples. To implement the BRR and Fay's method of variance estimation, we use the R 'survey' package [58] to create the replicate weights. The adjustment parameter ρ in the Fay's method is set to 0.3. These replicate weights are used as follows: for BRR (4.9), for each of the replicate weights, we obtain the weighted PSM estimator of the ACET. The variance of these PSM estimators obtained across the set of replicate weights is calculated as the final variance estimate of the PSM estimator. For the Fay's method (4.11), it further adjusts the variance estimator by the adjustment parameter ρ. For the full bootstrap method, we run 1000 bootstrap samples to study the variance of the PSM estimator. We first draw bootstrap clusters with replacement in each stratum followed by drawing bootstrap subjects with replacement within each bootstrap cluster to get the final bootstrap sample. In the bootstrap sample, we re-estimate the PS, using the same unweighted logistic regression model, and we re-conduct the PSM. Finally, we calculate the weighted mean difference among re-matched pairs. Same as before, the variance of PSM estimators from the 1000 full bootstrap samples is calculated as the variance estimate of the PSM estimator. For the conditional bootstrap, we neither re-estimate the PS nor re-conduct the PSM. Instead, we draw bootstrap clusters with replacement in each stratum similar to what we do for the full bootstrap, then we bootstrap the matched pairs in the original simulation sample within each bootstrap cluster. Afterwards we calculate the weighted mean difference among the re-matched pairs. Finally, we derive the variance estimate of the PSM estimator from studying the variance of PSM estimators from the 1000 conditional bootstrap samples.

As performance metrics, the mean value, the median value, the 25% percentile and the 75% percentile of the Monte Carlo simulation estimates by each of the 4 methods are reported and shown in boxplots. We compare the empirically estimated true variance to the 4 variance estimates to assess the performance of these variance estimation methods.

4.4.5 Simulation scenarios

Scenario one: different covariate distributions

Varying the distributions of the covariates will change the heterogeneity in characteristics. It will affect the PSM. In addition, covariates generated from different distributions will result in different non-responses and different population effects (due to interactions between the treatment and covariates). In short, varying the covariate distribution results in different heterogeneity in characteristics. It changes the true variance of the PSM estimator, and affects performance of the variance estimators. Three varying parameters are considered in this scenario, in order to change the covariate distribution: 1) the stratum mean $\tau^{stratum}$; 2) the cluster specific covariate mean $\tau^{cluster}$ and 3) the variability of the covariates within the cluster $\tau^{covariates}$. We vary one of the 3 parts and fix the other two each time. The possible values of each of the three varying parts are set to 0.1, 0.2 and 0.3. At the same time, $\sigma^{cluster}$ and $\sigma^{outcome}$ are fixed and are set to 1.

Scenario two: different effect distributions

We also vary the distributions of the potential outcomes to vary the effect distributions. This variation changes neither the matching mechanism nor the non-response mechanism. However, it may change the variability of the final estimated ACET. We hypothesize that this change would have less effect on the variance estimation compared to that in the previous section. There are 2 parts that could change the effect distribution: 1) the variation of the cluster specific treatment effect $\sigma^{cluster}$ (which leads to different heterogeneity in cluster specific treatment effect) and 2) the variation of the error term of the outcome $\sigma^{outcome}$. Similar to the above, we change one of them each time, and the potential values of both of them are set to 1 and 2. In this scenario, $\tau^{stratum}$, $\tau^{cluster}$ and $\tau^{covariates}$ are fixed and are set to 0.1.

Scenario three: increased number of strata

We set the number of strata to a relatively small number, four, in all scenarios we have introduced above. However, different numbers of strata could also change the variability of the true variance. Especially when the true variance of the PSM estimator is large, a relatively large number of strata could help reduce the true variance to a reasonable value. Increasing the number of strata is equivalent to increasing the number of replicate weights, either for the jackknife method or for the BRR method or for Fay's method. We investigate a scenario where we increase the total number of strata from 4 to 10. In this scenario, $\tau^{stratum}$, $\tau^{cluster}$ and $\tau^{covariates}$ are fixed to 0.1. $\sigma^{cluster}$ and $\sigma^{outcome}$ are set to 1.

Scenario four: increased sample clusters

So far we have considered scenarios where the number of PSU's sampled, i.e the number of clusters samples, is 2 in all strata. However, in practice there also could be more than 2 clusters drawn in each stratum. In addition, for both the full bootstrap and the conditional bootstrap, sampling with replacement from 2 clusters is problematic, as in those situations the bootstrap population size on the stratum level is 2 which is too small. In this scenario, we consider drawing 10 clusters, instead of 2 clusters, with replacement to form the sample. Correspondingly, we increase the number of clusters in each stratum in the population from 20 to 50. However, both the BRR and Fay's method require that there are only 2 sample clusters in all stratum. To solve this issue, we proceed in 2 different ways. In the first, we combine the 10 sample clusters to 2 big clusters. The method used for combination is kmeans clustering where k is set to 2 based on the estimated PS. Combining multiple clusters together is frequently used in practice. For example, the PATH study combines 3 clusters into 2 in those strata which have more than 2 sampled clusters. After combining the clusters, we follow the same process we have discussed before to estimate the variance. This is a straightforward approach but it ignores the between cluster variation, which matters in survey design. In the second approach, we randomly split the 10 clusters in each stratum into 5 groups. Each group contains 2 out of 10 randomly selected sample clusters. We call the randomly formed groups pseudo sub-strata, thus we create 5 pseudo sub-strata within each stratum in the sample. In this way, we artificially create and increase the number of strata. Again, we follow the same process with these constructed sub-strata for follow-up analysis afterwards.

4.4.6 Simulation results

We are most interested in investigating the bias of the variance estimators. In Figure 4.1-4.3, we increase the value of the 3 varying parts, $\tau^{stratum}$, $\tau^{cluster}$ and $\tau^{covariates}$ respectively each time. Three panels are included in each figure, representing different values of the parameter under consideration, with the other two parameters fixed at the value of 0.1. The corresponding values are shown in the label on top of each panel. The horizontal dashed line represents the true variance empirically estimated from the Monte Carlo simulation, and the dot in the boxplot indicates the empirically estimated mean using the corresponding variance estimator. In Figure 4.1, when we increase the standard deviation of the stratum mean from 0.1 to 0.3, the true variance of the PSM estimator increases. This is as expected, because the subjects are more heterogeneous in this case. Notice that the magnitude of the increase is not big. Among the 4 variance estimation methods, BRR and Fay's method perform the best and are similar to each other. Although the variance of these two estimators are larger than the 2 bootstrap methods, the expected values of the BRR or the Fay's method are very close to the true variance. The percent bias, defined as the percentage of the absolute difference between the estimate and the true variance over the true variance, is 9% for the BRR estimate and 10% for the Fay's estimate, as shown in panel 3. The 2 bootstrap methods perform similarly to each other, and both underestimate the true variance. In Figure 4.2, we increase the variability of the cluster specific covariate means $\tau^{cluster}$. As this variability increases, we notice a significant increase in the true variance of the PSM estimator (panel 3). In addition to the increased heterogeneity of subject characteristics, as we have seen when we increase $\tau^{stratum}$, more importantly here there is less chance that subjects can successfully match each other within the same cluster. This has a huge effect on the PS estimation as well as the followup PSM estimator. In this case, the percent bias is 4% for the BRR estimate and 2% for the Fay's estimate. Meanwhile, the Fay's method results in a smaller standard deviation. Each of these two methods still out-performs the bootstrap methods which consistently underestimate the true variance. In Figure 4.3, we show the summary results when we increase the covariates variability $\tau^{covariates}$, from the left panel to the right panel. Increasing $\tau^{covariates}$ has the least effect on the variance estimation compared to increasing $\tau^{stratum}$ or $\tau^{cluster}$. The increased variability of the covariates within the clusters doesn't affect the PSM much.

In the second scenario summarized in Figure 4.4, we consider varying the 2 parameters

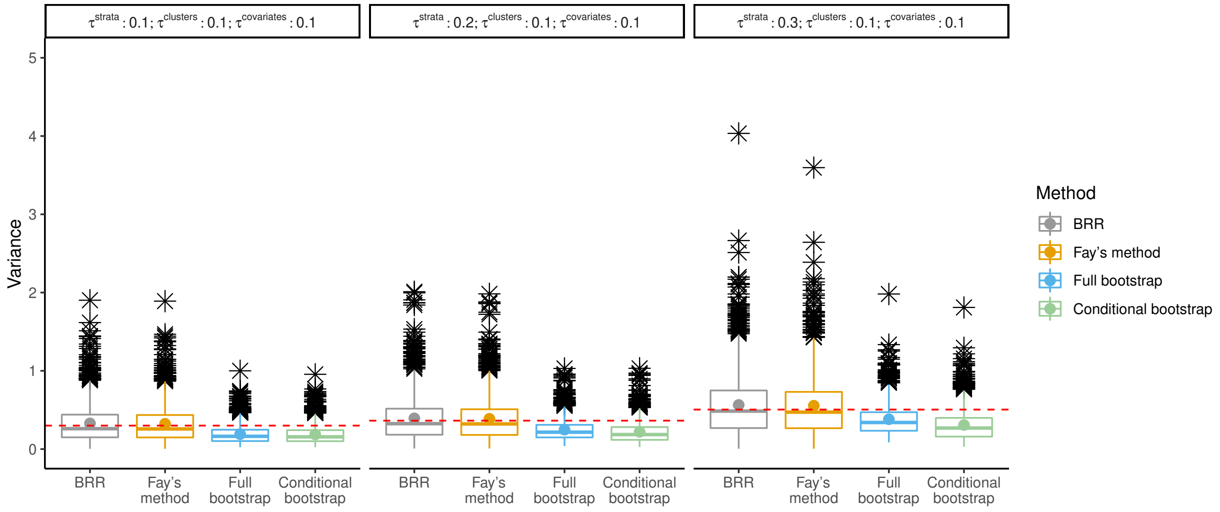


Figure 4.1: Scenario one results (varying $\tau^{stratum}$).

Scenario one with fixed $\tau^{cluster} = 0.1$, $\tau^{covariates} = 0.1$, $\sigma^{cluster} = 1$ and $\sigma^{outcome} = 1$. The number of strata is 4 and the number of clusters in all strata is 20. Varying $\tau^{stratum}$ from 0.1, 0.2 to 0.3.

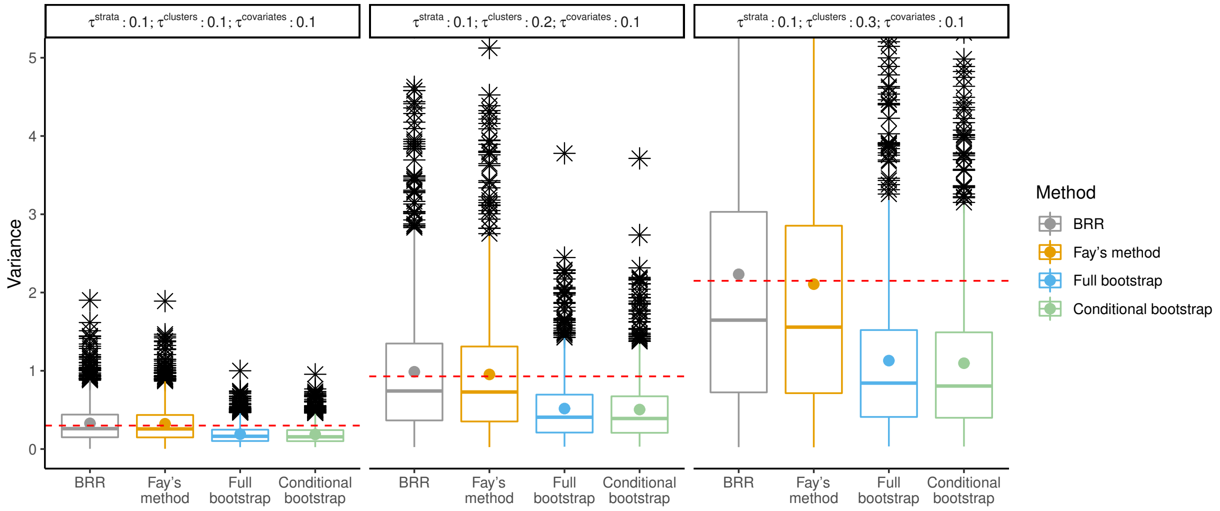


Figure 4.2: Scenario one results (varying $\tau^{cluster}$).

Scenario one with fixed $\tau^{stratum} = 0.1$, $\tau^{covariates} = 0.1$, $\sigma^{cluster} = 1$ and $\sigma^{outcome} = 1$. The number of strata is 4 and the number of clusters in all strata is 20. Varying $\tau^{cluster}$ from 0.1, 0.2 to 0.3.

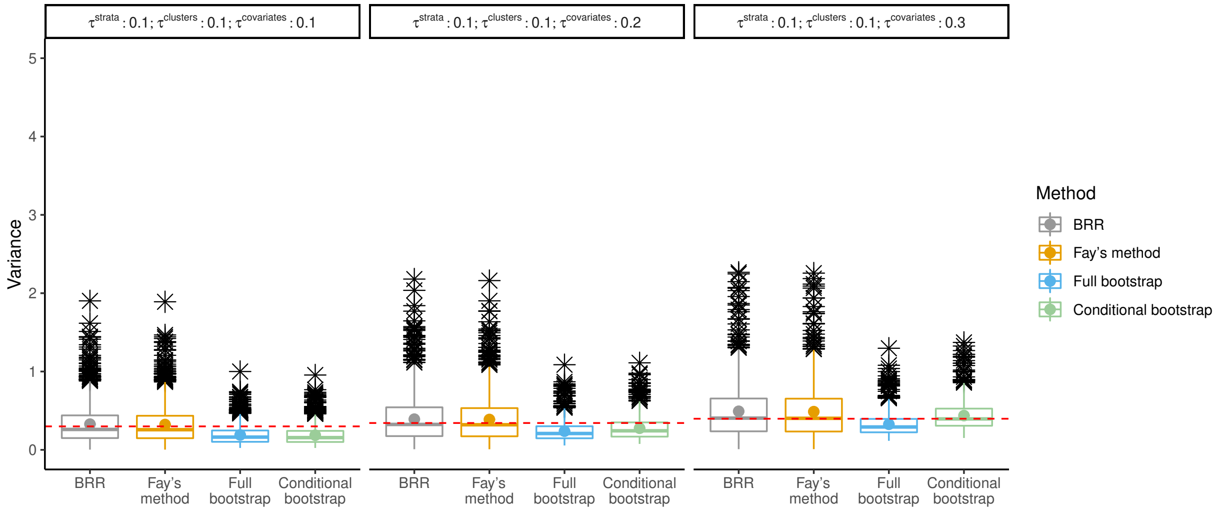


Figure 4.3: Scenario one results (varying $\tau^{covariates}$).

Scenario one with fixed $\tau^{stratum} = 0.1$, $\tau^{cluster} = 0.1$, $\sigma^{cluster} = 1$ and $\sigma^{outcome} = 1$. The number of strata is 4 and the number of clusters in all strata is 20. Varying $\tau^{covariates}$ from 0.1, 0.2 to 0.3.

$\sigma^{cluster}$ and $\sigma^{outcome}$ which are related to the effect distribution. These parameters have no effect on either the PS estimation or the non-response mechanism. The PSM is conducted within each cluster, thus as the second panel in Figure 4.4 shows, increasing the variability of the cluster specific treatment effect doesn't affect the true variance of the PSM estimator as long as the matching proportion is high enough. In the 3rd panel in Figure 4.4, we see that increasing the effect variability by increasing the variability of the error term increases the true variance of the PSM estimator of ACET, similar to the scenario we have shown when we increase $\tau^{stratum}$. In all scenarios we have seen so far, both the BRR estimate and the Fay's estimate work well and the Fay's estimate performs slightly better with smaller standard deviation. In scenario 1 & scenario 2, we consider cases with a comparably small number of strata, which is 4. In scenario 3, we increase the number of strata from 4 to 10. As Figure 4.5 shows, the variance is reduced when the number of strata increases, as expected. Again, the Fay's estimate ends up with a smaller percent bias as well as a smaller standard deviation compared to the BRR estimate.

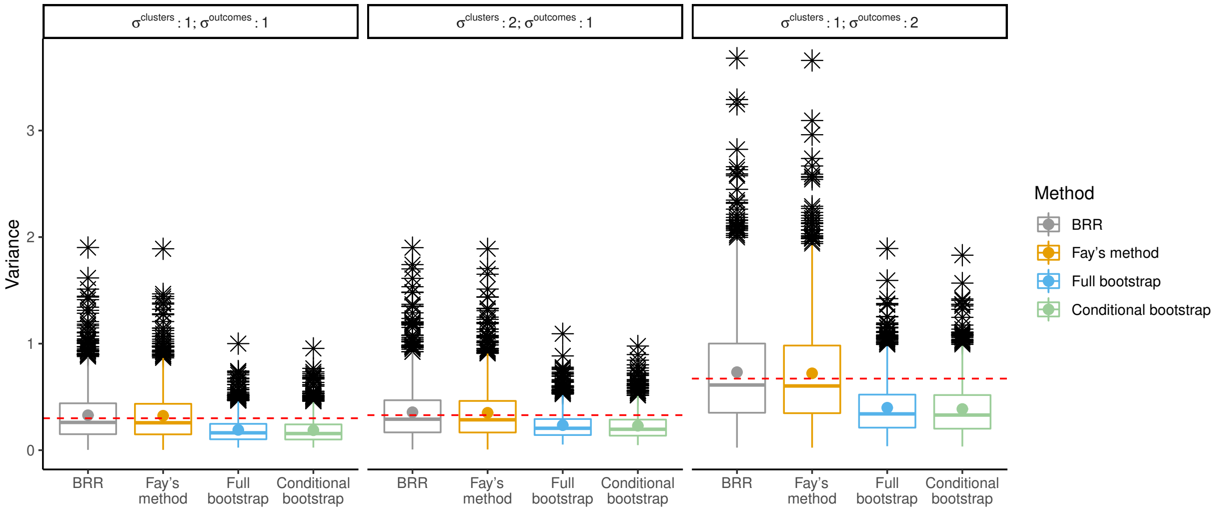


Figure 4.4: Scenario two results.

Scenario two with fixed $\tau^{stratum} = 0.1$, $\tau^{cluster} = 0.1$ and $\tau^{covariates} = 0.1$. The number of strata is 4 and the number of clusters in all strata is 20. Varying $\sigma^{cluster}$ and $\sigma^{outcome}$ respectively from 1 to 2.

In scenario 4, the number of strata remains at 4, and we increase the number of clusters within each stratum in the population from 20 to 50. Now, instead of drawing 2 sample clusters of 4 in each stratum as previously, we draw 10 out of the 50 clusters, with replacement. Figure 4.6 summarizes the findings. Combining multiple clusters to 2 PSU is a commonly-used strategy in practice to handle the dilemma where we have more than 2 PSU. [38, 64] However, in this simulation scenario, we seen that it has poor performance, in that combining clusters followed by using either the BRR method or the Fay's method significantly overestimate the variance of the PSM estimator of ACET. Both of the two methods result in a percent bias of 207%. However, creating pseudo sub-strata followed by using these two methods works very well (with percent biases 1% and 9% respectively). An interesting finding in this scenario is that both the full bootstrap and the conditional bootstrap perform very well as the number of sample clusters increases (with percent biases 5% and 4% respectively). This implicitly demonstrates the reason why the bootstrap methods failed in previous scenarios. It is actually due to the limited number

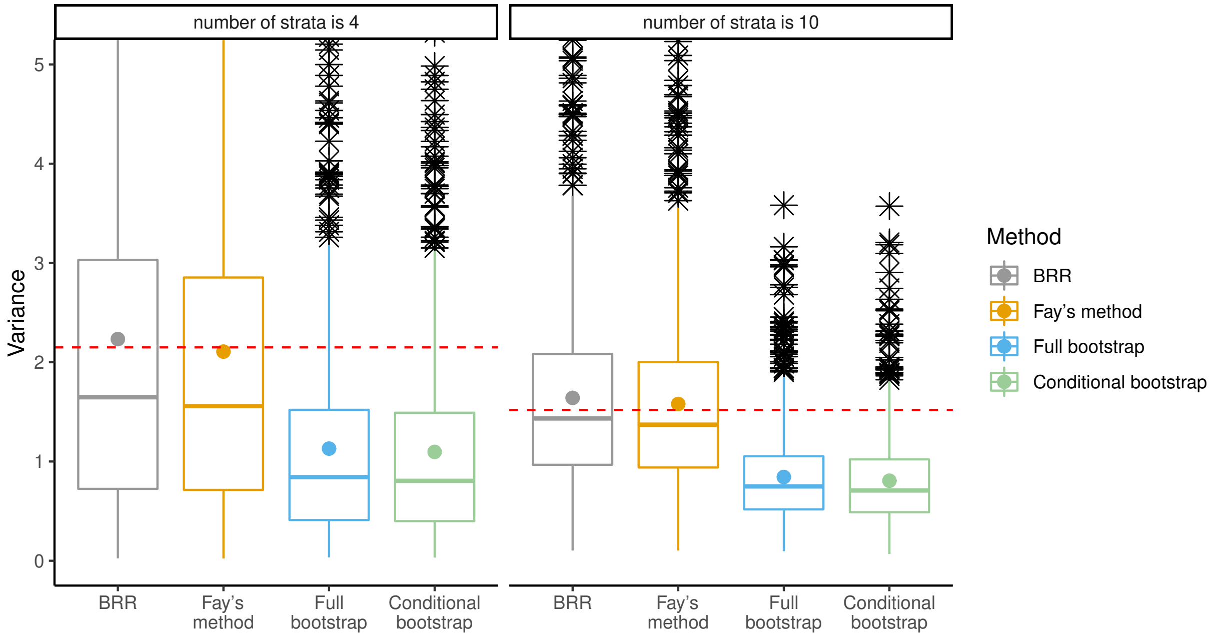


Figure 4.5: Scenario three results.

Scenario three with fixed $\tau^{stratum} = 0.1$, $\tau^{cluster} = 0.3$, $\tau^{covariates} = 0.1$, $\sigma^{cluster} = 1$ and $\sigma^{outcome} = 1$. The number of clusters in all strata is 20. Increasing the number of strata from 4 to 10.

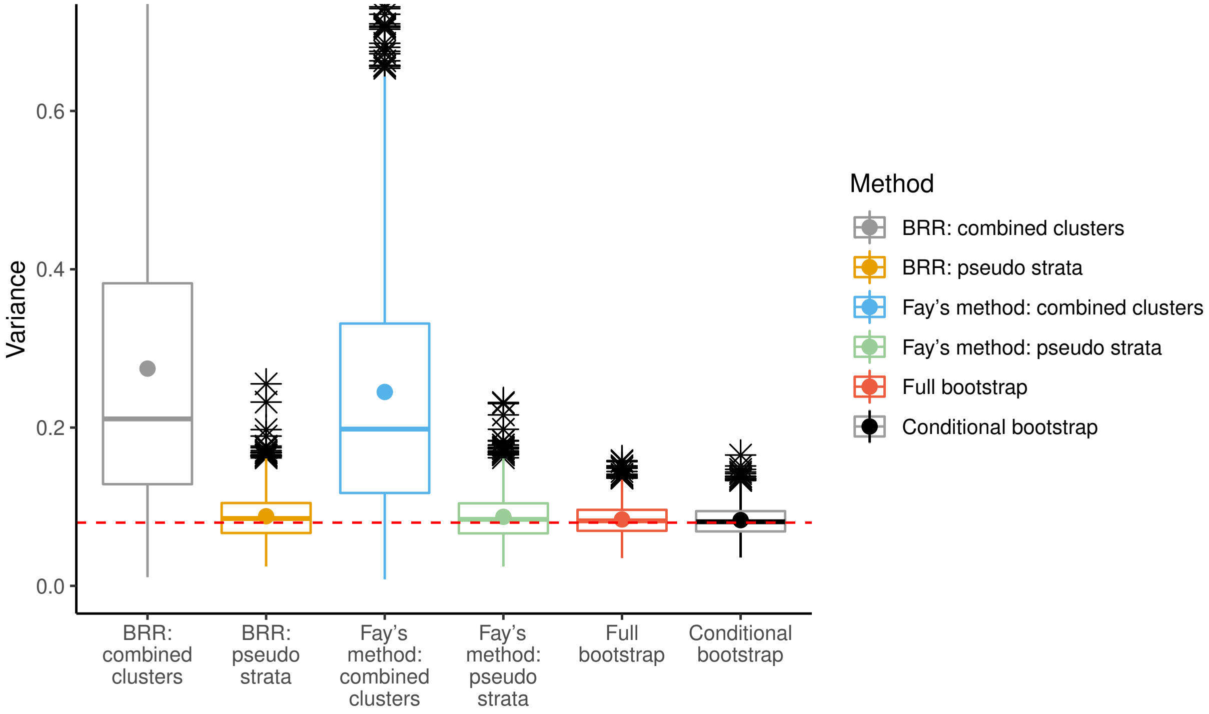


Figure 4.6: Scenario four results.

Scenario four with fixed $\tau^{stratum} = 0.1$, $\tau^{cluster} = 0.1$, $\tau^{covariates} = 0.1$, $\sigma^{cluster} = 1$ and $\sigma^{outcome} = 1$. The number of strata is 4. Increasing the number of clusters from 20 to 50 and sample 10 out 50 clusters in all strata.

of clusters, in other words, the bootstrap population is too small and drawing bootstrap sample clusters from the limited bootstrap population with replacement is not valid. When the number of sample clusters from the population is big enough, the bootstrap methods would finally outperform the BRR and the Fay's method. In addition, the variability of the bootstrap estimators are smaller than the variability of the BRR and the Fay's method estimators. Again, Fay's method slightly outperforms the BRR method and the full bootstrap and the conditional bootstrap consistently perform similar to each other, which is also as what we expected because the PSM is conducted without replacement.

4.5 Case study

Finally, we apply the variance estimation methods of the PSM estimator discussed in the simulation study to a real study to assess their performance using actual survey data. We use the data from the PATH study, [38] a large national longitudinal study aiming to assess tobacco use and how it affects the health of people in the United States. Data are collected at approximately annual intervals (Waves). The question of interest here is whether, among baseline smokers who made quit attempt within the past year, using counseling or self-help materials in the quit attempt helped them reduce cigarette consumption in the long-term. Comparing to those who didn't use such assistance to help them quit, we hypothesize that those who accepted counseling or self-help materials would smoke less in the long-term. In addition to using counseling or self-help materials, participants could use e-cigarettes, other tobacco products such as cigars, pipe, hookah, snus, smokeless tobacco, as well as nicotine replacement therapy and pharmaceutical aids to help them quit cigarette smoking, or none of these. To test our hypothesis, we use the PSM to match those who didn't use any counseling or self-help materials to those who used either counseling or self-help materials in their last quit attempt, followed by calculating the weighted mean difference in cigarette consumption reduction among the matched pairs. Our main focus in this real study is to assess the variance estimate of this PSM estimate of ACET.

In detail, we focus on Wave 2 (W2) to W4 PATH survey data. We include survey participants who were baseline cigarette smokers at W2 and made at least one quit attempt between W2 to W3. Whether they used counseling or self-help materials to help quit smoking during their last quit attempt is the exposure of interest and it was assessed at W3. Finally, we calculate the outcome, cigarette consumption reduction, as the difference between W4 cigarette consumption and W2 cigarette consumption. We use PSM to form the matched pairs as we mentioned earlier. Specifically, we select 20 baseline covariates which were potentially related to either the exposure or the outcome to estimate the PS using a logistic regression model. The covariates are mixed which include both categorical covariates and continuous covariates. In detail, they are age, sex, ethnicity, race, education, income, nicotine dependence score, duration of last quit attempt, timing from last quit attempt, smoking-related health problems, smoking pack-years, age started smoking fairly regularly earlier than 18, perceived harm of cigarettes, self-efficacy about quitting cigarettes, interest in quitting cigarettes, smoking-free home, exposure to other smokers, health insurance status and internal/ external mental health problems. The PATH study used a stratified multistage sampling, same as what we have discussed in this chapter. Two clusters were sampled as PSU in all strata followed by sampling individual participants from each of the sample PSU. The non-responses were also adjusted by multiplying the inverse probability of being a response to the original survey weights for all participants. The PATH study used the Fay's method with the adjustment parameter $\rho = 0.3$ to form the replicate weights, thus the BRR is not considered in this case study. In total, one hundred replicate weights were generated and provided in the survey data. We use these 100 replicate weights directly to do the following analysis. We calculate the variance of the PSM estimator using the Fay's method, the full bootstrap and the conditional bootstrap respectively and report the results.

In total, we get 2454 baseline smokers who either smoked daily or non-daily at W2, and have answered both W2 and W4 cigarette consumption questions. Among those participants, 237 used counseling or self-help materials to help them quit during their last quit attempt. The other 2217 participants either used other products such as e-cigarettes, tobacco products to help them quit, or they didn't use anything during that time. Among the 20 mixed baseline covariates, participants have the choice and can not answer specific questions or they forget certain questions. Because of this, many covariates have missing values. We use simple imputation to impute the missing covariates before we conduct the PSM and the final analysis. The imputation is conducted with R package 'mice'. [**?**] The variation of the imputation is not taken into account as it is not of the main interest in this chapter. Instead, we fix one imputed data set for further analysis.

Those 2454 participants came from 98 unique strata. In each stratum, we have 2 PSU except for one specific stratum in which we have only one sample cluster remain. In the analysis, we find that the mean estimate, the cigarette consumption reduction between the 2 groups of participants, by the PSM, is -1.98. That is, those who didn't use counseling or self-help materials to help them quit reduced around 2 more cigarettes per day compared to those who used counseling or self-help materials. The estimated variance of this PSM estimator by the 3 variance estimation methods are: 1) the Fay's method: 8.13; 2) the full bootstrap: 13.52 and 3) the conditional bootstrap: 9.32.

The results of the case study show that using counseling or self-help materials among PATH W2 smokers didn't help them reduce cigarette consumption in the long-term compared to those who didn't use counseling or self-help materials. The results are far from achieving statistical significance. The variance estimation by the Fay's method, the full bootstrap and the conditional bootstrap do not really differ to each other in this real study.

4.6 Discussion

In this chapter, we have investigated variance estimation for the matching estimator of the ACET, in the context of a complex survey sample, with survey weights. We take the PSM estimator as an example, however our results can be easily generalized to other matching estimators. In 4.3, we have shown that under mild regularity conditions when the cluster size goes to infinity, the PSM estimator is consistent in estimating the ACET within clusters. Based on this property along with other assumptions given in 4.3, we have shown that the mean estimate of the PSM estimator is asymptotically consistent over the entire sample when both the cluster size and the number of strata goes to infinity, using the asymptotic theory given by Serfling in a double array setting. Further, we have proven the consistency of the variance estimation of the PSM estimator using either the jackknife method or the BRR method, using a similar approach. These two variance estimation methods can be re-written to the same form when the estimator is a linear transformation of the outcome and the number of sample clusters is 2 in all strata. The PSM estimator in our setting satisfies such a condition. Fay's method, as an adjusted BRR method, improves the BRR estimate in certain circumstances and it is commonly-used in practice. In addition, the consistency of the BRR estimate can also be generalized to the variance estimate using the Fay's method. In 4.4, we ran a simulation study to compare the performance of the BRR method and Fay's method to the performance of bootstrap methods, including the full bootstrap and the conditional bootstrap. The results have shown that in general when the number of PSU in each stratum in the survey data is 2, the BRR estimate and the Fay's estimate work well and they outperform the full bootstrap and the conditional bootstrap. Fay's method slightly outperforms the BRR method also. The full bootstrap and the conditional bootstrap perform similarly to each other but they underestimate the true variance in this case. However, when the number of the PSU in all strata is large, the performance of the two bootstrap methods is improved and might finally outperform the BRR and the Fay's method, as the bootstrap estimates usually result in less variability. Finally in 4.5, we have used the PATH data to study the performance of these variance estimation methods for the PSM estimator. The results showed that in practice, the BRR method, the Fay's method and the bootstrap methods might not really differ much from each other.

As we have mentioned, we are actually estimating the average causal effect among those treated subjects who can find successful matches in the control group. For most situations in our simulation study, the proportion getting successful matching across the whole sample is high. In those situations, the Monte Carlo mean estimate of the PSM estimator is close enough to the true ACET. For the other scenarios for example where the standard deviation of the cluster specific covariate mean is too big, the matching proportion substantially reduces. In these scenarios, we estimate the effect among those treated subject that can be matched. For this estimand of interest in practice, the PSM estimator is still consistent. Also it is of interest to note that the variability of the cluster specific treatment effect doesn't effect the final variance estimation. This term cancels out between the potential outcomes of the treated and the potential outcomes of the control.

In addition, the sampling design that has been discussed in this chapter is with replacement. Sampling with replacement is used as it doesn't affect the variance estimation. [57] In contract, when we use sampling without replacement but keep assuming a with replacement design, the true variance would be overestimated [57] and need to be adjusted. However, our results can be generalized to the sampling without replacement design. Remember we assume that a small number of clusters is drawn in all strata, that is, we assume m_h is fixed. Also, we assume the number of strata H goes to infinity, thus we have $\sum_{h=1}^{H} m_h / \sum_{h=1}^{H} M_h \to 0$ as $\sum_{h=1}^{H} m_h \to \infty$. Rao and Shao [70] have shown that under this condition, the practice of sampling without replacement still leads to the approximately unbiased variance estimation of the matching estimator using the standard variance estimator, which is $\hat{V}ar_{rep}(\bar{d})$ we discussed in this chapter.

Survey designs may draw different numbers of PSU in each stratum. A commonly used design is to draw 2 PSU in all strata, in which case the BRR method or Fay's method may be advantageous for variance estimation. They substantially reduce the number of replicates compared to the jackknife method, which saves a great amount of time in practice. In addition, we have shown that in this scenario, neither the full bootstrap nor the conditional bootstrap is appropriate. The two bootstrap methods perform similarly and both underestimate the true variance. When more clusters or PSU are drawn in the strata, the performance of the bootstrap methods are improved as the bootstrap population (the number of sample clusters in the strata) increases. When the number of PSU is big enough, the bootstrap methods may work better than the BRR and the Fay's method as the bootstrap base estimators end up with smaller variances. In the same scenario where more than 2 PSU are drawn in all strata, the BRR and the Fay's method can still perform well. A suggested adjustment approach is to split each stratum into several pseudo sub-strata, where 2 PSU remain in each created pseudo sub-stratum, and conduct the same analysis with the constructed pseudo sub-strata. In practice, another potential adjustment is to combine different PSU into 2 big PSU, as what the PATH study does. [38] However, in our simulation we have shown this approach doesn't work well. It turns out to overestimate the variance significantly.

The full bootstrap and the conditional bootstrap for estimating the variance of the matching estimator turn out to be similar across our simulation studies. This is consistent with previous findings [12] where matching is without replacement. When matching is conducted with replacement, the conditional bootstrap method will underestimate the true variance, whereas the full bootstrap still works well or is conservative. Future work will focus on how to derive and evaluate variance estimation for the matching estimator when matching is with replacement.

4.7 Afterthoughts before next Chapter

So far, we have used the PSM technique to investigate the effectiveness of e-cigarette use on the long-term cigarette cessation using PATH survey data. To improve the existing PSM method, we have further explored the DM estimator to make DR causal inference. In addition, we have proved the large sample properties of the variance estimator for the matching estimator of the ACET in complex surveys.

In the last chapter, we will summarize and discuss the findings from Chapters 2 to 4 and give our final conclusions. We will also consider future work that continues our research in the matching estimator causal inference area.

This chapter, in full, has been prepared for submission for publication as ”Chen, Ruifeng; Messer, Karen S. *Large sample properties of the variance estimation of matching estimators in complex survey*”. The dissertation author was the primary author on this paper.

Chapter 5

Conclusions and future work

This dissertation has presented the explicit form of a DM estimator, an improved method of PSM which can make DR causal inference. Furthermore, we have shown how to conduct the variance estimation of the matching estimator in complex surveys with survey weights. The performance of these proposed methodologies has been illustrated in simulation studies and with applications to the PATH tobacco survey data. The motivation of exploring the main statistical methodologies came from a tobacco control study, presented in chapter two, where we used PSM to investigate whether e-cigarette use in a quit attempt has helped US smokers who use them to achieve smoking cessation in the long-term. Results of the use of the DM estimator in such questions were consistent with results of the use of the PSM estimator that we used in the original question. Moreover, we have presented and demonstrated a consistent estimator of the variance of such matching-based estimators variance, for use with complex survey data.

In particular, in chapter two we used the PATH data to address whether the use of e-cigarettes to aid quitting contributed to increased successful smoking/ nicotine cessation in the US population (self-reported 12+ months continuous abstinence [28]). We focused on any e-cigarette use for quitting compared to no use and we used caliper nearest neighbor PSM to match each e-cigarette user with up to two closely matched control respondents on 24 potential confounders identified a priori. We compared population-weighted abstinence rates in the matched samples. This approach estimated the effect of e-cigarette use explicitly among those who choose to use them as a cessation aid, and was less dependent on modelling assumptions than regression-based approaches which estimate the average causal effect among the entire population (both users and non users). [9] A follow-up regression model was conducted to achieve a DR estimator. Bootstrap quantile confidence intervals were used to assess statistical significance. Finally, we verified the appropriateness of the PSM by comparing the kernel density estimates of the PS. The PSM analysis found no evidence for a difference in the proportion who achieved long-term abstinence from cigarettes between those who used e-cigarettes to help quit smoking and the matched sample of those who did not use e-cigarettes as a cessation aid. Instead, more importantly, we found that e-cigarettes users were less likely to be long-term nicotine abstinent at follow up.

In chapter three, we studied the performance of the DM estimator which matches on both the PS score and the PGS score simultaneously in estimating the ACET. In a simulation study, the performance of the proposed DM estimator was compared with other commonly-used estimators in estimating the ACET including the OLS estimator, the PSM estimator and a usual regression-based DR estimator PSM-OLS (PSM-regression more generally). As expected, the DM estimator performs well when at least one of the PS and the PGS is correct. Meanwhile, the DM estimator only pays a very small to modest penalty in efficiency for its double-robustness. In our simulated data, one of our covariates is omitted in predicting the PS, compared to the situation where all covariates are simulated to be related to the PGS, thus the PSM is less efficient than the OLS estimator when both of the models are correct. Importantly, an interaction between one of our covariates and treatment indicator has been added to make the ACET differ from the ACE. We further explored interval estimators for the DM estimator in this chapter, under a SRS design. Here, we considered the full bootstrap which bootstraps sample individuals followed-by re-estimating the PS and re-matching bootstrap pairs, the conditional bootstrap which bootstraps the pairs in the original sample directly without re-estimating the PS or re-conducting the matching, and a parametric method which estimates the variance from the difference of the matched pairs. Results have shown that the full bootstrap consistently works well although it is sometimes conservative. The conditional bootstrap works similar to the parametric approach, and they are more efficient when the DM estimator uses matching without replacement. However, they underestimate the variance when matching with replacement. For both the cigarette cessation rate and the cigarette consumption reduction, the DM estimator ended up with negative results.

In chapter four, we investigated methods for variance estimation of the matching estimator in complex surveys with survey weights. We studied the PSM estimator as an example, which can be generalized to other matching estimators. We showed that the PSM estimator is asymptotically consistent in estimating the ACET, or the causal effect among those matched treated subjects, with given assumptions and under mild regularity conditions, using the asymptotic theory given by Serfling in a double array setting. Furthermore, we proved the asymptotic consistency of the jackknife and the BRR variance estimators for the PSM estimator. These two estimators can be re-written in the same form when the PSM estimator is a linear transformation of the outcome and the number of sample clusters is 2 in all strata. In addition, consistency can be generalized to the Fay's estimate which is an improved BRR method and is commonly-used in practice in survey studies. We used a simulation study to compare the performance of the BRR estimate and Fay's estimate to the performance of the full bootstrap and the conditional bootstrap variance estimators. In these two bootstrap methods, each bootstrap-sample cluster is followed by bootstrap-sampling individuals within clusters to get the final bootstrap sample. The results showed that in general when we have 2 PSU in each stratum in the survey design, the BRR estimate and the Fay's estimate work well and outperform the full bootstrap and the conditional bootstrap. In our simulation, Fay's estimate consistently slightly outperforms the BRR estimate, which ends up with a larger standard deviation and percent bias. The full bootstrap and the conditional bootstrap perform similarly to each other, but they both underestimate the true variance. When the number of PSU's in all strata is increased, the performance of the two bootstrap methods improves. In this scenario, the BRR estimate and the Fay's estimate still work, but we need to create pseudo sub-strata to ensure each pseudo sub-stratum contains 2 PSU. We compare the performance of these variance estimators using the PATH data and we found that using counseling or self-help materials didn't help smokers who want to quit cigarette smoking reduce the cigarette consumption in the long-term. We use it as a case study example in this chapter without controlling for all major variables. The case study also shows that in real data, these variance estimators might not differ much to each other and end up with similar conclusions.

In conclusion, we have used PSM estimator of the ACET to demonstrate that the use of e-cigarettes is not an effective method of quitting cigarettes among US smokers. In fact, it may contribute to continuing use of nicotine. Following this study, we proposed and explicitly studied the DM estimator which gives the PSM estimator more room to be correct in estimating the ACET. Finally, we have assessed methods for variance estimation of the matching estimator in complex surveys and would suggest Fay's method should be used in such circumstances. In the following paragraphs, we discuss some of key findings in these chapters as well as the future work.

First of all, in the second chapter, besides the fact we found that e-cigarettes are not helpful for cigarette cessation, it's worth noticing the low rates of nicotine abstinence found in our study. Nicotine abstinence was measured by e-cigarettes, other tobacco products, and NRT products. Long-term nicotine abstinence was well under 5% for US smokers who used e-cigarettes to quit, and nearly doubled for those who did not. Among those who successfully used e-cigarettes to attain long-term abstinence from cigarettes, two-thirds were still using e-cigarettes during the follow-up year. One thing of particular concern to public health professionals is the high rate of continued smoking of other forms of tobacco among those who successfully quit cigarettes, ranging from 17% of those who successfully used e-cigarettes to quit to 7% among successful pharmaceutical aid users.

An interesting finding in the third chapter is the extra robustness to model mis-specification which is possible for matching estimators. In certain circumstances, exact matching is still possible even though the balancing measure in the PGS or the PS model is incorrect. This allows for extra robustness in the incorrect model specification case for matching estimators as compared to their regression-based counterparts. Methods which are even more independent of the functional form of these models have been proposed [36] and are worth further study. However, there is also the trade-off for the matching estimators when a non-negligible proportion of treated subjects has been discarded after matching, which is not the case in our study but also needs further assessment.

Considering matching with replacement, theory suggests and our simulations in chapter 3 confirm that it has higher variance than matching without replacement. In case of a large covariate imbalance between treatment groups, matching with replacement may reduce bias, however. For estimators incorporating matching with replacement under the SRS design, conditional methods of confidence interval estimation are not recommended. Both the conditional bootstrap and conditional parametric approaches resulted in under-coverage in our simulation scenarios. As the replication rate increases, the under coverage will become more severe. The full bootstrap had conservative coverage for some of the matching estimators, and it became more conservative when matching with replacement, consistent with previous findings [34, 12]. The full bootstrap can be considered for use, as Abadie [4] showed that matching on the estimated PS adds a non-positive adjustment factor to the asymptotic variance of the estimator which matches on the true PS. This lends theoretical support to the idea that the full bootstrap will always be conservative for matching with replacement. As a correction, Bodory [15] proposes the wild bootstrap in which the sample covariates are fixed, and in each bootstrap sample, treatment indicators are resampled and the propensity score is re-estimated. They showed that the wild bootstrap comes closer to the nominal size than the full bootstrap for matching with replacement. Future work in this area will focus on finding a similar adjusted approach for the DM estimator.

We aim to estimate the ACET throughout, however, as we have mentioned, many times we are actually estimating the average causal effect among those treated subjects who can find a successful match in the control group. For example for scenarios in chapter four where the standard deviation of the cluster specific covariate mean is too big, the matching proportion substantially reduces. In these scenarios, we are actually estimating the effect among those treated subject that can be matched. For this estimand of interest in practice, the PSM estimator is still consistent. Another cluster -related variation component, the variability of the cluster specific treatment effect, doesn't effect the final estimate of variance. This term cancels out between the potential outcomes of the treated and the potential outcomes of the control.

Finally, in chapter 4, our simulation results consistently show that the Fay's estimate slightly outperforms the BRR estimate. We randomly choose the adjustment parameter $\rho = 0.3$ in our setting. In the future, it's worth exploring the underlying mechanism and the choice of the adjustment parameter ρ in different scenarios. In addition, the variance estimators based on the full bootstrap method and on the conditional bootstrap turn out to be similar across our simulation studies. This is consistent with previous findings [12] where matching is without replacement. When matching is conducted with replacement, the conditional bootstrap method is going to underestimate the true variance, where the full bootstrap still works well or conservative. Future work will focus on how to derive and evaluate the variance estimation for the matching estimator when matching is with replacement.

Appendix A

Additional materials for Chapter 2

A.1 Measurement detail for pre-identified study covariates in PATH study

Socio-demographics: Use standard derived variables for age, sex, ethnicity, race, education, and income. Note the variable of education comes from PATH Wave 1 database, it's not available in Wave 2 database based on our knowledge.

Nicotine dependence scale: Variables are combined to derive nicotine dependence scale by calculating the mean of the non-missing scores. Nicotine dependence items take the form of a series of statements on emotional and physical responses to nicotine products (e.g. ”I frequently crave product”, ”I usually want to use product right after I wake up”, ”I [would] feel alone without my product”). Respondents are asked to rate their level of agreement with each statement on a 5-point scale, where 1=”Not true of me at all” and 5=”Extremely true of me”. Respondents can also answer ”don't know” or refuse to answer the question; these are treated as missing responses. Responses are rescaled to a 3-point scale, where 1 (not at all) = 0, 2 or 3 = 50 and 4 or 5 =100, summed and divided by the number of non-missing values.

Cigarette consumption: Average number of cigarettes smoked each day. Responses could

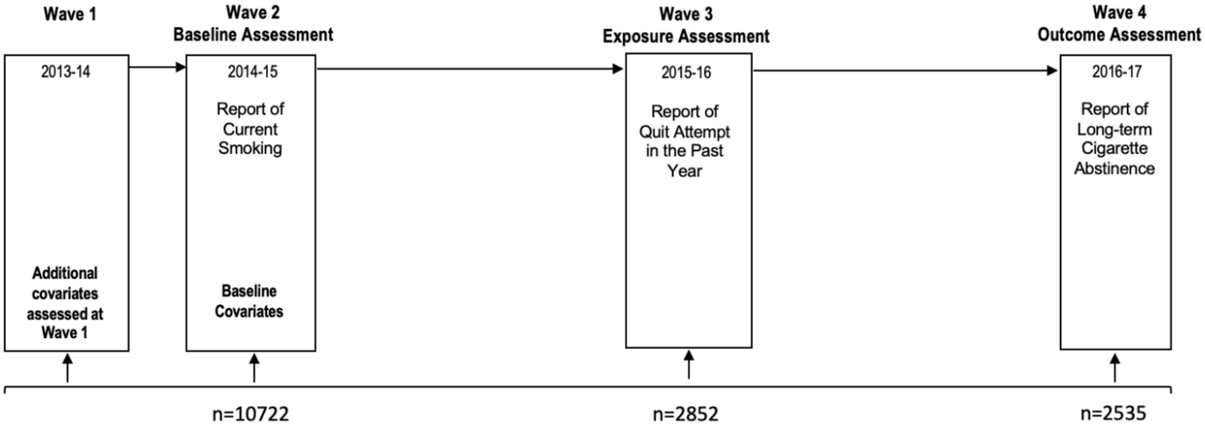


Figure A.1: Data collection schema for e-cigarette use to aid long-term smoking cessation in the US analysis.

Note:
Sample sizes (n) are unweighted.
Current smoking includes both daily and non-daily smoking.
Additional covariates assessed at Wave 1 include the assessments of smoking related diseases and daily e-cigarette use at Wave 1.

be reported as cigarettes or packs. For respondents with missingness, we replaced their cigarette consumption by multiplying average number of cigarettes smoked per day among non-current 30-day smokers with the number of days smoked in the past 30 days, and divided by 30 days. Responses larger than 100 were considered errors and were replaced with 100 (the maximum value is 100).

Length of the QA Prior to Wave 2: Length of last quit attempt in the past 12 months.

Timing of the QA: This was calculated as the date of W3 survey completed minus the end date of the most recent quit attempt reported in W3.

Smoke-free home: Statement that best describes rules about smoking a combustible tobacco product inside home. It's a 3-point scale from 1 (not allowed anywhere or anytime at all) to 3 (allowed anywhere or anytime at all).

Perceived harm of cigarettes: Respondents were asked ”How harmful do you think cigarettes are to health?” and could reply on a 5-point scale from 1 (not at all harmful) to 5 (extremely harmful).

Relative perceived harm of e-cigarettes: Respondents were asked ”Is using e-cigarettes less harmful, about the same, or more harmful than smoking cigarettes?” and could reply on a 3-point scale, where 1=Less harmful, 2=About the same and 3=More harmful.

Exposure to other smokers: ”In the past 7 days, number of hours that you were in close contact with others when they were smoking.”

Pack years of smoking: Calculated by multiplying the number of pack smoked per day by the number of years the respondent smoked regularly, when respondents answered “Yes” to either question ”Smoked same number of cigarettes per day since started smoking fairly regularly” or ”Smoked same number of cigarettes per day since started smoking fairly regularly (Pack measure)”.

Age started smoking fairly regularly Binary variable with answers yes or no.

Interest in quitting cigarettes: On a scale of 1-10 where 1=Not at all interested and 10=Extremely interested.

Self-efficacy about quitting: ”If you did try to quit product altogether in the next 6 months, how likely do you think you would be to succeed?” on a 4-point scale from 1=Not at all likely and 4=Very likely.

Smoking related health diagnoses: Respondents were asked if they had ever been told by a doctor or health professional that they had any of the listed diseases. Group A: Heart Disease: High blood pressure; High cholesterol; Congestive heart failure; A stroke; A heart attack; Some other heart condition. Group B: Respiratory Disease: COPD; chronic bronchitis; emphysema; asthma; some other lung or respiratory condition. Group C: Cancer.

Disorder symptoms for externalizing mental health problems: Respondents were asked the last time they had experienced any of 7 externalizing (e.g., had a hard time paying attention or listening to instructions at school, work or home, bullied or started physical fights). The number of reports of experiencing such symptoms in the past month or the past 2-12 months was summed and coded into a 3-level severity indicator, with those reporting 0 or 1 symptom scored as Low, 2-3 symptoms scored as Moderate and 4 or more scored as High.

Disorder symptoms for internalizing mental health problems: Respondents were asked the last time they experienced any of 4 internalizing disorder symptoms: feeling very trapped, lonely, sad, blue, depressed, or hopeless about the future, feeling very anxious, nervous, tense, scared, panicked, or like something bad was going to happen, had sleep problems. The number of reports of experiencing such symptoms in the past month or the past 2-12 months was summed and coded into a 3-level severity indicator, with those reporting 0 or 1 symptom scored as Low, 2-3 symptoms scored as Moderate and 4 or more scored as High.

Insurance coverage at Wave 2: Respondents who reported currently being covered by at least one type of health insurance, including insurance purchased directly or through an employer or union, Medicare, Medicaid, VA, TRICARE or other military health care and Indian Health Insurance, were scored as having insurance coverage. Missing data on all of these variables were coded to ”did not have insurance”.

Daily cigarettes use at Wave 2: Respondents' cigarettes used at Wave 2, either smoked every day or smoked some days.

Daily e-cigarettes use at Wave 1 or Wave 2: Either daily e-cigarette use at Wave 1 or daily e-cigarette use at Wave 2 were considered prior daily e-cigarette use.

Table A.1: Univariate distribution of study covariates by cessation aid category.

Parameter	Used e-Cigarettes on Quit Attempt (n = 427)			Did Not Use e-Cigarettes on Quit Attempt (n = 2,108)			Used pharmaceutical aid only on Quit Attempt (n = 465)			Used no product on Quit Attempt (n = 1,643)		
	No.	Weighted %	95% CI	No.	Weighted %	95% CI	No.	Weighted %	95% CI	No.	Weighted %	95% CI
Age												
18-34	218	20.5	17.5, 23.4	922	79.5	76.6, 82.5	115	10.3	8.4, 12.3	807	69.2	66.0, 72.4
35-50	127	19.2	15.9, 22.4	546	80.8	77.6, 84.1	144	20.7	17.2, 24.3	402	60.1	56.1, 64.2
≥ 50	82	11.8	9.5, 14.2	640	88.2	85.8, 90.5	206	29.0	24.9, 33.0	434	59.2	54.8, 63.5
Sex												
0 = Male	202	16.8	14.4, 19.2	1012	83.2	80.2, 86.2	191	17.6	15.1, 20.1	821	65.7	62.7, 68.6
1 = Female	225	18.1	15.3, 20.9	1095	81.9	79.1, 84.7	274	20.9	18.3, 23.6	821	61.0	57.7, 64.2
Education												
Less than high school	89	13.2	10.1, 16.4	593	86.8	83.6, 89.9	113	17.1	13.9, 20.3	480	69.7	65.1, 74.2
High school graduate	90	15.1	10.9, 19.2	502	84.9	80.8, 89.1	100	19.7	15.8, 23.5	402	65.3	60.0, 70.6
Some college or higher	230	21.1	18.3, 23.8	944	79.0	76.2, 81.7	242	20.0	17.7, 22.3	702	58.9	55.6, 62.2
Ethnicity												
Hispanic	37	8.7	5.9, 11.6	334	91.3	88.4, 94.1	40	9.6	6.7, 12.6	294	81.6	78.1, 85.2
Non-Hispanic	390	19.1	17.2, 21.1	1732	80.9	78.9, 82.8	413	20.5	18.6, 22.4	1319	60.3	57.8, 62.9
Race												
White	354	20.8	18.5, 23.0	1400	79.2	77.0, 81.5	326	20.0	17.9, 22.0	1074	59.3	56.5, 62.1
Black	26	5.6	3.3, 7.9	433	94.4	92.1, 96.7	90	19.3	15.2, 23.4	343	75.1	70.3, 79.9
Asian	4	*	*	33	88.7	80.6, 96.8	5	*	*	28	76.2	63.2, 89.3
Other	39	18.1	12.0, 24.2	190	81.9	75.8, 88.0	36	14.0	9.7, 18.4	154	67.8	61.0, 74.6
Income												
$< 35,000$	220	14.3	12.2, 16.5	1341	85.7	83.5, 87.8	268	17.5	15.1, 19.9	1073	68.2	65.3, 71.1
$350,00-100,000$	155	22.9	19.6, 26.1	523	77.1	73.9, 80.4	141	21.8	18.5, 25.2	382	55.3	50.9, 59.7
$> 100,000$	35	23.2	15.9, 30.4	110	76.8	69.6, 84.1	30	20.9	14.3, 27.6	80	55.9	47.4, 64.4
Health Insurance Status												
No	70	15.6	10.6, 20.5	414	84.4	79.5, 89.4	50	10.0	7.1, 13.0	364	74.4	69.2, 79.6
Yes	357	17.9	16.2, 19.7	1683	82.1	80.3, 83.8	411	21.2	19.1, 23.3	1272	60.9	58.4, 63.4
Disorder symptoms for externalizing mental health problems												
Low	184	13.2	11.3, 15.2	1241	86.8	84.8, 88.7	262	18.9	16.5, 21.3	979	67.9	64.9, 70.9
Moderate	134	23.8	19.4, 28.3	492	76.2	71.7, 80.6	127	21.3	17.9, 24.6	365	54.9	50.1, 59.7
High	109	22.5	18.5, 26.5	375	77.5	73.5, 81.5	76	17.2	12.8, 21.6	299	60.3	55.2, 65.4
Disorder symptoms for internalizing mental health problems												
Low	153	13.8	11.7, 16.0	1042	86.2	84.0, 88.3	222	19.1	16.6, 21.6	820	67.1	64.0, 70.1
Moderate	127	19.7	15.9, 23.6	530	80.3	76.4, 84.1	124	19.8	16.6, 23.1	406	60.4	56.6, 64.2
High	147	22.4	18.9, 25.9	536	77.6	74.1, 81.1	119	18.5	14.7, 22.3	417	59.1	54.1, 64.2
Smoking-related diseases												
1 = Marked	201	16.2	13.9, 18.5	1069	83.8	81.5, 86.1	308	24.8	21.8, 27.9	761	58.9	55.7, 62.2
Nicotine dependence												
0-33.3	89	14.2	10.8, 17.6	571	85.8	82.4, 89.2	51	8.2	5.5, 10.9	520	77.6	73.3, 81.9
33.4-66.7	172	17.1	14.5, 19.7	839	82.9	80.3, 85.5	203	21.2	18.1, 24.3	636	61.7	58.3, 65.1
66.8-100	165	21.6	18.4, 24.8	648	78.4	75.2, 81.6	207	27.0	23.7, 30.3	441	51.4	47.4, 55.4
Smoke-free home												
1 = Smoking is not allowed anywhere	246	17.5	15.3, 19.7	1174	82.5	80.3, 84.7	233	17.4	15.2, 19.6	941	65.2	62.0, 68.3
Perceived harm of cigarettes												
Not to somewhat harmful	65	12.4	9.1, 15.7	463	87.6	84.3, 90.9	83	15.8	12.2, 19.3	380	71.8	67.3, 76.3
Very/extremely harmful	360	18.6	16.7, 20.5	1641	81.4	79.5, 83.3	380	20.0	18.0, 22.0	1261	61.4	58.8, 64.1
Relative perceived harm of e-cigarettes												
1 = Less harmful	262	27.2	24.3, 30.1	726	72.8	69.9, 75.7	171	18.0	15.2, 20.8	555	54.8	51.5, 58.1
2 = About the same	142	12.5	10.3, 14.7	1064	87.5	85.3, 89.7	236	21.1	18.3, 23.8	828	66.5	63.1, 69.8
3 = More harmful	16	*	*	242	94.3	90.8, 97.8	36	13.3	8.7, 17.8	206	81.0	75.0, 87.0
Second-hand smoking hours in past 7 days												
≤ 10 hours	262	15.6	13.8, 17.3	1479	84.4	82.7, 86.2	331	19.8	17.5, 22.1	1148	64.7	61.8, 67.5
> 10 hours	161	22.1	18.6, 25.6	597	77.9	74.4, 81.4	130	18.0	14.9, 21.2	467	59.9	55.7, 64.0
Age began regular smoking												
18+	160	17.2	14.4, 20.0	874	82.8	80.0, 85.6	204	19.3	16.6, 22.1	670	63.5	59.8, 67.2
< 18	238	20.9	18.4, 23.5	907	79.1	76.5, 81.6	223	21.7	18.8, 24.6	684	57.4	54.4, 60.3
Cigarette consumption												
1-9 CPD	194	15.3	13.2, 17.5	1089	84.7	82.5, 86.8	161	12.6	10.4, 14.9	928	72.0	69.0, 75.0
10-19 CPD	131	20.3	16.4, 24.3	533	79.7	75.7, 83.6	150	23.6	20.0, 27.2	383	56.1	51.4, 60.8
20+ CPD	93	19.4	15.6, 23.2	436	80.6	76.8, 84.4	147	29.8	25.1, 34.5	289	50.8	46.2, 55.4
Pack-years												
< 20	137	19.8	16.5, 23.2	591	80.2	76.8, 83.5	121	17.1	14.0, 20.3	470	63.0	59.1, 66.9
$21-35$	29	19.7	12.3, 27.1	117	80.3	72.9, 87.7	41	31.7	23.0, 40.4	76	48.6	40.0, 57.2
> 35	16	18.0	8.0, 27.9	85	82.0	72.1, 92.0	36	34.5	23.6, 45.4	49	47.5	38.3, 56.8
Interest in quitting cigarettes												
1-7	139	15.0	12.3,17.7	802	85.0	82.3, 87.7	127	15.1	12.0, 18.3	675	69.8	66.3, 73.4
8-9	86	18.5	13.9, 23.2	379	81.5	76.8, 86.1	93	19.6	15.9, 23.2	286	61.9	56.0, 67.8
10 (extremely Interested)	164	19.0	16.3, 21.7	736	81.0	78.3, 83.7	205	23.6	20.5, 26.7	531	57.4	54.0, 60.8
Self-efficacy about quitting												
No intent to quit in next 6 mos	192	20.0	17.3, 22.8	826	80.0	77.2, 82.7	165	17.3	14.6, 20.1	661	62.7	59.5, 65.8
Not at all or a little likely	28	19.3	12.4, 26.3	117	80.7	73.7, 87.6	34	22.7	15.2, 30.3	83	57.9	49.1, 66.8
Somewhat likely	83	18.5	14.4, 22.6	366	81.5	77.4, 85.6	108	25.3	20.4, 30.2	258	56.2	50.5, 61.9
Very likely	60	13.8	9.9, 17.6	360	86.2	82.4, 90.1	81	18.9	14.6, 23.3	279	67.3	61.8, 72.7
Length of the QA reported at W1												
≤ 30 d	171	16.9	14.0, 19.8	845	83.1	80.2, 86.0	201	21.5	18.3, 24.6	644	61.7	57.9, 65.5
30+ d	68	25.0	19.1, 30.7	226	75.0	69.3, 80.9	40	15.4	10.7, 20.0	186	59.7	54.1, 65.4
No quit/no data	188	16.1	13.9, 18.2	1037	83.9	81.38, 86.1	224	18.1	15.7, 20.6	813	65.8	62.5, 69.1
Timing of the QA												
≤ 6 mo	205	15.7	13.6, 17.8	1136	84.3	82.2, 86.4	261	20.2	17.8, 22.6	875	64.1	61.3, 66.9
6+ mo	98	19.3	15.6, 23.0	461	80.7	77.0, 84.4	99	17.9	14.5, 21.4	362	62.8	58.1, 67.5
No quit/no data	124	19.3	16.2, 22.5	511	80.7	77.5, 83.8	105	18.0	14.6, 21.5	406	62.6	58.5, 66.7
Daily cigarettes use at W2												
1=Marked	290	17.4	15.3, 19.5	1455	82.6	80.5, 84.7	383	23.4	21.0, 25.8	1072	59.2	56.4, 61.9
Daily e-cigarette use at W1 or W2												
1=Marked	160	54.3	47.3, 61.3	96	45.7	38.7, 52.7	25	12.3	7.6, 16.9	71	33.4	26.3, 40.6

* Estimate was suppressed because it has low statistical precision. It is based on a denominator sample size of less than 20.

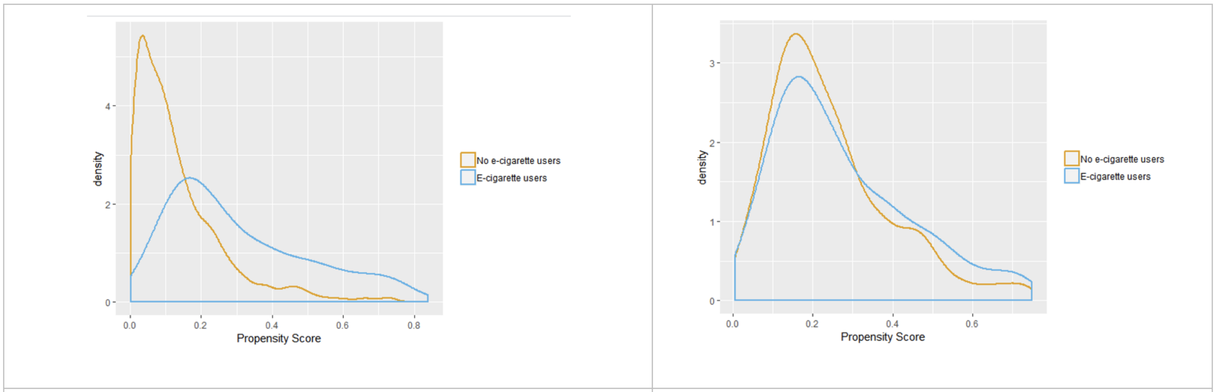
A.2 Improvement in covariate balance with propensity score matching: comparability of groups (kernel density plots and average covariate balance grid across all bootstrap runs)

Boxplots show, for each covariate in the propensity score model, the bootstrap distribution of the mean difference between exposed and non-exposed samples, before matching (left hand plot) and after matching (right hand plot), across all bootstrap samples (1500 bootstrap for both the primary comparison and the secondary comparison). For each bootstrap sample, the covariate is standardized using the entire sample prior to dividing into exposed and non-exposed subjects and taking the mean. Missing observations are imputed for each bootstrap sample as an initial step.

A.2.1 Primary comparison: e-cigarettes on the QA versus no e-cigarettes on the QA

A.2.2 Previous daily use of e-cigarettes across study groups

We presented details of the covariate with the largest residual between-group difference after matching: previous daily use of e-cigarettes. That there was still a residual difference after 1:2 matching reflects the considerable difference in previous daily e-cigarette use between the two study groups. Among those who used e-cigarettes to help them on the QA, 20.8% were previous daily users and a further 21.5% had been fairly regular users. For those who did not use e-cigarettes on the quit attempt, only 3.2% had been previous daily users of e-cigarettes and a further 7.4% had used e-cigarettes fairly regularly.



Before Matching Selected Bootstrap Sample Size
(E-cigarettes use: 406, no E-cigarettes use: 2129)

After Matching (matching is without replacement)
(E-cigarettes use: 375, no E-cigarettes use: 676)

Figure A.2: E-cigarettes on the QA vs no e-cigarettes on the QA: randomly selected one example from 1500 bootstrap runs (PS of e-cigarettes use).

Table A.2: Prior e-cigarette use among US smokers who made a quit attempt in 2015-2016[a], by use or no use of e-cigarettes as a cessation aid.

	E-Cigarettes used to quit (W3) (n=294)		E-cigarettes not used to quit (W3) (n=1881)	
Prior E-cigarette use	%	95% C.L.	%	95% C.L.
Never	15.4	10.6, 20.2	53.7	51.5, 55.9
Ever but not fairly regularly	42.3	37.7, 46.9	35.7	33.5, 38.0
Fairly regularly but not daily W1 or W2	21.5	16.7, 26.4	7.4	6.0, 8.7
Daily W1 or W2	20.8	15.6, 25.9	3.2	2.4, 4.1

Abbreviations: C.L., Wilson Confidence Limit; the QA, last quit attempt; W2, PATH Study Wave 2; W3, PATH Study Wave 3; W4, PATH Study Wave 4.

[a]. Weighted U.S. population estimates.

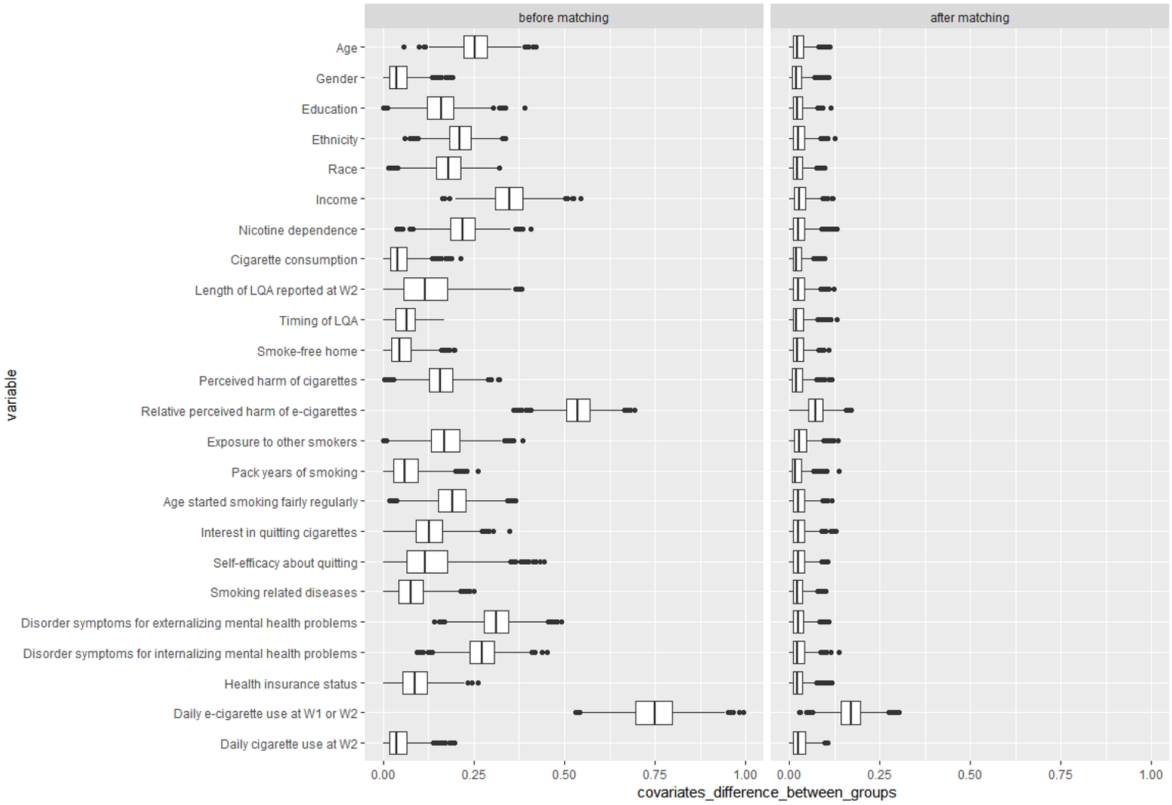
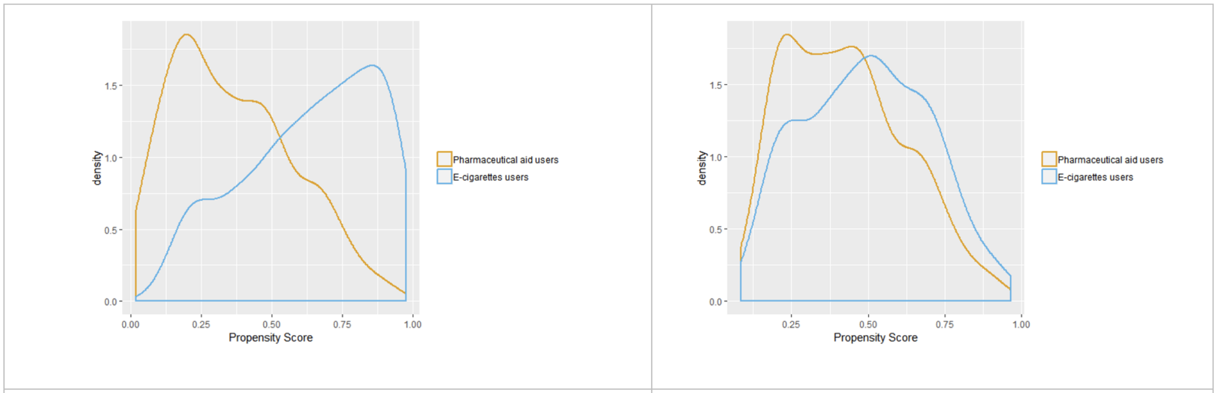


Figure A.3: Standardized differences in 24 important covariates between smokers who used e-cigarettes to quit in 2015-2016 and those who did not, before and after matching.

Note: for a given covariate, we define ”a marked improvement in covariate balance from matching” as a decrease of at least 0.1 units in the median absolute difference of the standardized covariate between exposed and non-exposed subjects, comparing the bootstrap distribution before and after matching. These comparisons do not use the survey weights. For comparison 1, the following 16 covariates below achieved a marked improvement in covariate balance from the matching procedure (ordered by size of the difference in medians): Daily e-cigarette use at W1 or W2, Relative perceived of harm of e-cigarettes, Income, Disorder symptoms for externalizing mental health problems, Disorder symptoms for internalizing mental health problems, Age, Nicotine dependence, Ethnicity, Age started smoking fairly regularly, Race, Exposure to other smokers, Perceived harm of cigarettes, Education, Interest in quitting cigarettes, Self-efficacy about quitting, Length of the QA reported at W2.



Before Matching Selected Bootstrap Sample Size
(E-cigarettes use: 444, pharmaceutical aid use: 448)

After Matching (matching is without replacement)
(E-cigarettes use: 251, pharmaceutical aid use: 353)

Figure A.4: E-cigarettes on the QA vs pharmaceutical aid on the QA: randomly selected one example from 1500 bootstrap runs (PS of e-cigarettes use).

A.2.3 Secondary comparison: e-cigarettes on the QA versus pharmaceutical aid on the QA

A.3 Sensitivity analyses of the main PSM analyses

We present sensitivity analyses to add 1:2 propensity score matching matched pairs or triples as random effects instead of fixed effects in the logistic regression after matching (figure 4) as well as 1:1 propensity score matching with adjusting indicators of matched pairs or triples as fixed effects, to check the robustness of our findings.

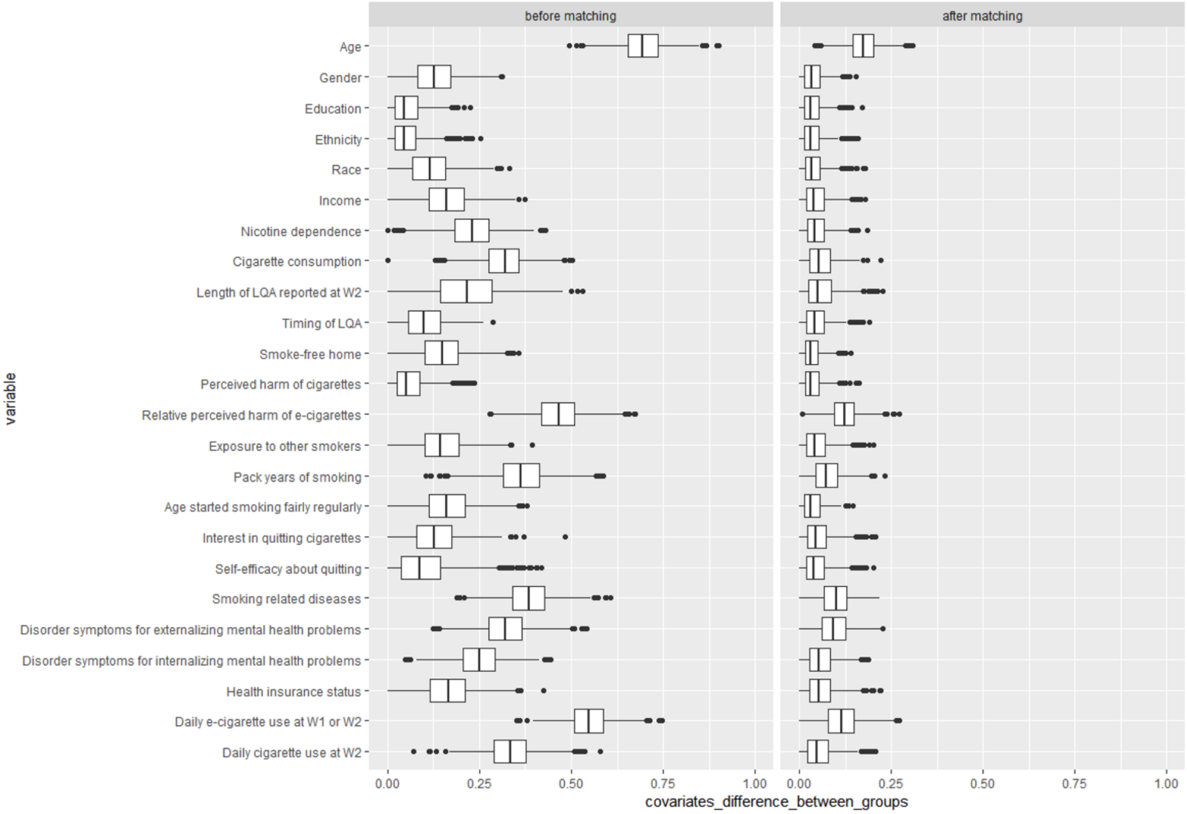


Figure A.5: Standardized differences in 24 important covariates between smokers who used e-cigarettes to quit in 2015-2016 and those who used pharmaceutical aid to quit, before and after matching.

Note: for a given covariate, we define ”a marked improvement in covariate balance from matching” as a decrease of at least 0.1 units in the median difference of the standardized covariate between exposed and non-exposed subjects, comparing the bootstrap distribution before and after matching. These comparisons do not use the survey weights. For comparison 3, the following 16 covariates below achieved a marked improvement in covariate balance from the matching procedure (ordered by size of the difference in medians): Age, Daily e-cigarette use at W1 or W2, Relative perceived of harm of e-cigarettes, Smoking related diseases, Daily cigarette use at W2, Pack years of smoking, Cigarette consumption, Disorder symptoms for externalizing mental health problems, Disorder symptoms for internalizing mental health problems, Nicotine dependence, Length of the QA reported at W2, Age started smoking fairly regularly, Health insurance status, Income, Exposure to other smokers, Gender.

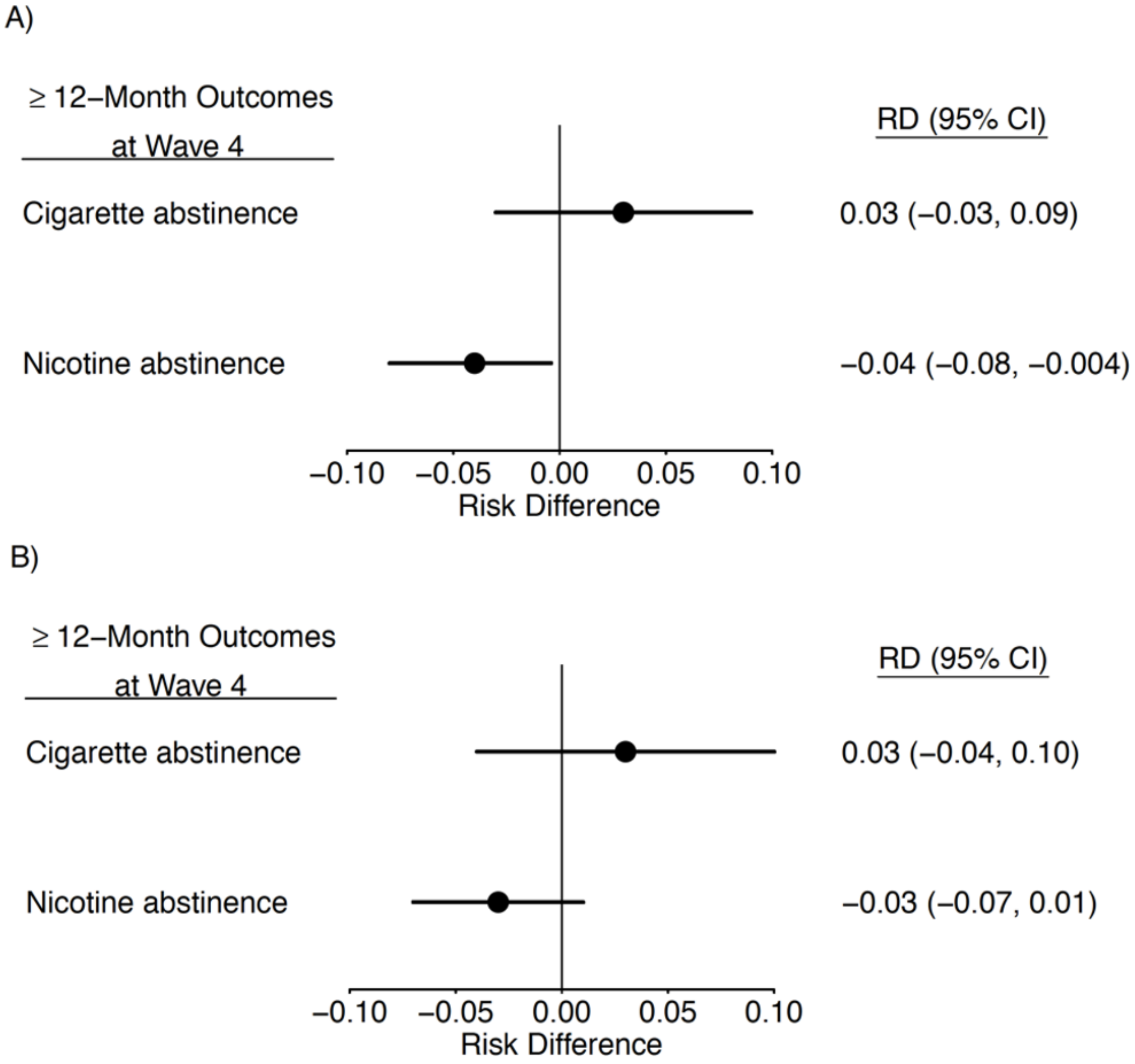


Figure A.6: The 1:2 propensity score matching with matched pairs or triples adjusted as random effects.

Note: adjusted estimate of differences [a] in long-term cigarette abstinence[b] and long-term nicotine abstinence[b], according to cessation aid to quit among US smokers, population assessment of tobacco and health study, United States, wave2 to wave4. A) e-cigarettes vs no e-cigarettes on last quit attempt prior to W3; B) e-cigarettes vs pharmacotherapy on last quit attempt prior to W3.

[a]. The mean difference between fitted values in the exposed group and the unexposed group from an adjusted logistic regression model, with longitudinal survey weights.

[b]. Abstinence from cigarette smoking for 12+ months and from any nicotine product for 12+ months, assessed at W4.

Confidence intervals are Bonferroni adjusted for each comparison.

A.3.1 The 1:2 propensity score matching with matched pairs or triples adjusted as random effects

A.3.2 The 1:1 propensity score matching with matched pairs or triples adjusted as fixed effects

Additionally, we conducted the sensitivity analysis to test the robustness of 1:2 propensity score matching by comparing estimated mean effects of each outcome of each comparison with those effects derived using 1:1 propensity score matching, in which we did not adjust any remaining covariate or overall propensity score due to the excellent covariates balance after matching. (data not shown) All estimates for each comparison and for each outcome were very similar to those estimates in our main propensity score matching analyses, which indicates 1:2 matching didn't introduce significant bias in our study compared to 1:1 matching.

A.4 Sensitivity analyses of logistic regression models for PSM models

We present sensitivity analyses for our primary comparison, use of e-cigarettes on the QA vs no use of e-cigarettes on the QA, related to both long-term (12+ months) cigarettes abstinence and nicotine abstinence; and our secondary comparison, use e-cigarettes on the QA vs use pharmaceutical aid on the QA with same endpoints. Overall, we used logistic regression analyses with longitudinal survey weights in parallel to PSM analyses in this section. Missing covariates were imputed with one simple imputation before logistic regression (seed was set to 1000 in R). Bonferroni correction was adjusted to primary comparison related 2 outcomes, and the secondary comparison related 2 outcomes respectively.

E-cigarette on the QA vs no e-cigarette on the QA	Risk Difference	Risk Ratio	Odds Ratio	Adj 95% CI for OR
12+ months cigarette abstinence at W4	0.02	1.14	1.20	0.78, 1.85

E-cigarette on the QA vs no e-cigarette on the QA	Risk Difference	Risk Ratio	Odds Ratio	Adj 95% CI for OR
12+ months nicotine abstinence at W4	-0.05	0.35	0.38	0.15, 0.99

A.4.1 a. primary comparison: use e-cigarettes on the QA vs did not use e-cigarettes on the QA

12+ months cigarette abstinence: Logistic regression of e-cigarette as a cessation aid compared to no e-cigarette on the QA, controlled for relevant propensity score covariates (age, gender, ethnicity, race, education level, income, nicotine dependence (ND), relative perceived harm of e-cigarettes and daily e-cigarettes use at Wave 1 or Wave 2), corrected for multiple (2) comparisons.

12+ months nicotine abstinence: Logistic regression of e-cigarette as a cessation aid compared to no e-cigarette on the the QA, controlled for relevant propensity score covariates (age, gender, ethnicity, race, education level, income, nicotine dependence (ND), relative perceived harm of e-cigarettes and daily e-cigarettes use at Wave 1 or Wave 2), corrected for multiple (2) comparisons.

Findings from the 4.a of sensitivity analyses of the primary comparison (use e-cigarettes on the QA vs did not use e-cigarettes on the QA)

Outcome 1. 12+ months cigarette abstinence: There was no difference in 12+ month cigarettes abstinence at W4 between those who used an e-cigarette to quit and those who did not. The confidence limits on odds ratio crossed 1.0.

Outcome 2. 12+ months nicotine abstinence: There was statistically significant difference in 12+ month nicotine abstinence at W4 between those who used an e-cigarette to quit and those who did not. The confidence limits on odds ratio didn't cross 1.0. Using an e-cigarette to

E-cigarette on the QA vs pharmaceutical aid on the QA	Risk Difference	Risk Ratio	Odds Ratio	Adj 95% CI for OR
12+ months cigarette abstinence at W4	0.04	1.42	1.36	0.64, 2.88

E-cigarette on the QA vs pharmaceutical aid on the QA	Risk Difference	Risk Ratio	Odds Ratio	Adj 95% CI for OR
12+ months nicotine abstinence at W4	-0.03	0.50	0.42	0.11, 1.57

quit prior to W3 caused 5 percent decrease in 12+ month nicotine abstinence at W4 compared to not using an e-cigarette to quit prior to W3.

A.4.2 b. secondary comparison: use e-cigarettes on the QA vs use pharmaceutical aid on the QA

12+ months cigarette abstinence: Logistic regression of e-cigarette as a cessation aid compared to pharmaceutical aid on the the QA, controlled for relevant propensity score covariates (age, gender, ethnicity, race, education level, income, nicotine dependence (ND), relative perceived harm of e-cigarettes and daily e-cigarettes use at Wave 1 or Wave 2), corrected for multiple (2) comparisons.

12+ months nicotine abstinence: Logistic regression of e-cigarette as a cessation aid compared to pharmaceutical aid on the the QA, controlled for relevant propensity score covariates (age, gender, ethnicity, race, education level, income, nicotine dependence (ND), relative perceived harm of e-cigarettes and daily e-cigarettes use at Wave 1 or Wave 2), corrected for multiple (2) comparisons.

Findings from the 4.b of sensitivity analyses of the second secondary comparison (use e-cigarettes on the QA vs use pharmaceutical aid on the QA)

Outcome 1. 12+ months cigarette abstinence: There was no difference in 12+ month cigarettes abstinence at W4 between those who used an e-cigarette to quit and those who used pharmaceutical aid only to quit. The confidence limits on odds ratio crossed 1.0.

Outcome 2. 12+ months nicotine abstinence: There was no difference in 12+ month

The interaction between e-cigarette use and baseline smoking status	Odds Ratio	Adj 95% CI for OR
12+ months cigarette abstinence at W4	2.15	0.99, 4.68
12+ months nicotine abstinence at W4	3.05	0.37, 25.25

nicotine abstinence at W4 between those who used an e-cigarette to quit and those who used pharmaceutical aid only to quit. The confidence limits on odds ratio crossed 1.0.

A.5 Sensitivity analyses of logistic regression models for interactions between e-cigarette use and key covariates on cigarette abstinence/ nicotine abstinence

We present sensitivity analyses for testing whether e-cigarette use on cigarette abstinence/ nicotine abstinence was different by study key covariates: baseline smoking status (daily vs non-daily cigarette smoking at Wave 2), nicotine dependence (nicotine dependence scale < 50 vs ≥ 50), age (< 35 vs ≥ 35), sex (male vs female), education level (at least some college or higher vs others) and race & ethnicity (non-Hispanic white vs others). Logistic regressions were conducted, and one of the key covariates listed above, the e-cigarette use and their interaction were included as predictors in each of the logistic regression models. Those with missing covariates were removed in the logistic regression models.

A.5.1 a. interaction between e-cigarette use and baseline smoking status on cigarette abstinence/ nicotine abstinence

Logistic regression of e-cigarette use, baseline smoking status and their interaction on cigarette abstinence/ nicotine abstinence

The interaction between e-cigarette use and nicotine dependence	Odds Ratio	Adj 95% CI for OR
12+ months cigarette abstinence at W4	1.52	0.71, 3.25
12+ months nicotine abstinence at W4	6.17	0.78, 48.54

The interaction between e-cigarette use and age	Odds Ratio	Adj 95% CI for OR
12+ months cigarette abstinence at W4	0.92	0.42, 2.05
12+ months nicotine abstinence at W4	0.91	0.17, 4.94

A.5.2 b. interaction between e-cigarette use and nicotine dependence on cigarette abstinence/ nicotine abstinence

Logistic regression of e-cigarette use, nicotine dependence and their interaction on cigarette abstinence/ nicotine abstinence

A.5.3 c. interaction between e-cigarette use and age on cigarette abstinence/ nicotine abstinence

Logistic regression of e-cigarette use, age and their interaction on cigarette abstinence/ nicotine abstinence

A.5.4 d. interaction between e-cigarette use and sex on cigarette abstinence/ nicotine abstinence

Logistic regression of e-cigarette use, sex and their interaction on cigarette abstinence/ nicotine abstinence

The interaction between e-cigarette use and sex	Odds Ratio	Adj 95% CI for OR
12+ months cigarette abstinence at W4	0.50	0.22, 1.12
12+ months nicotine abstinence at W4	0.89	0.19, 4.20

The interaction between e-cigarette use and education level	Odds Ratio	Adj 95% CI for OR
12+ months cigarette abstinence at W4	1.07	0.49, 2.34
12+ months nicotine abstinence at W4	0.21	0.04, 1.04

A.5.5 e. interaction between e-cigarette use and education level on cigarette abstinence/ nicotine abstinence

Logistic regression of e-cigarette use, education level and their interaction on cigarette abstinence/ nicotine abstinence

A.5.6 f. interaction between e-cigarette use and race ethnicity on cigarette abstinence/ nicotine abstinence

Logistic regression of e-cigarette use, race & ethnicity and their interaction on cigarette abstinence/ nicotine abstinence

Findings from outcome 1. 12+ months cigarette abstinence: None of the interactions between the key covariates and e-cigarette use were statistically significant, which indicates that there were no difference of e-cigarette use on cigarette abstinence among those with different baseline smoking status, nicotine dependence, age, sex, education level and race & ethnicity The confidence limits on odds ratio crossed 1.0.

The interaction between e-cigarette use and race & ethnicity	Odds Ratio	Adj 95% CI for OR
12+ months cigarette abstinence at W4	1.54	0.57, 4.17
12+ months nicotine abstinence at W4	0.89	0.17, 4.69

The interaction between e-cigarette use and baseline smoking status on cigarette abstinence was nearly significant, and daily baseline cigarette users were more likely to be cigarette abstinence at Wave 4 if they used e-cigarette at Wave 3, compared to non-daily baseline cigarette users (OR for the interaction: 2.15, 95% CI: 0.99, 4.68). However, e-cigarette use was not significant to cigarette abstinence among either daily baseline cigarette users or non-daily users. In detail, among daily baseline cigarette users, the odds ratio for cigarette abstinence comparing those who used e-cigarette to quit and those who didn't use e-cigarette to quit was 1.52 (95% CI: 0.95, 2.42). Among non-daily baseline cigarette users, the odds ratio for cigarette abstinence comparing those who used e-cigarette to quit and those who didn't use e-cigarette to quit was 0.71 (95% CI: 0.38, 1.30).

Findings from outcome 2. 12+ months nicotine abstinence: None of the interactions between the key covariates and e-cigarette use were statistically significant, which indicates that there were no difference of e-cigarette use on nicotine abstinence among those with different baseline smoking status, nicotine dependence, age, sex, education level and race & ethnicity The confidence limits on odds ratio crossed 1.0.

The interaction between e-cigarette use and education level on nicotine abstinence was nearly significant, and those who were more educated were less likely to be nicotine abstinence at Wave 4 if they used e-cigarette at Wave 3, compared to non-daily baseline cigarette users (OR for the interaction: 0.21 95% CI: 0.04, 1.04). In detail, e-cigarette use was significant to nicotine abstinence among those who were more educated, the odds ratio for nicotine abstinence comparing those who used e-cigarette to quit and those who didn't use e-cigarette to quit was 0.13 (95% CI: 0.03, 0.48). E-cigarette use was not significant to nicotine abstinence among those who were less educated, and the odds ratio for nicotine abstinence comparing those who used e-cigarette to quit and those who didn't use e-cigarette to quit was 0.59 (95% CI: 0.24, 1.46).

Bibliography

[1] Alberto Abadie and Guido W. Imbens. Large Sample Properties of Matching Estimators for Average Treatment Effects. *Econometrica*, 74(1):235–267, JAN 2006.

[2] Alberto Abadie and Guido W. Imbens. Estimation of the conditional variance in paired experiments. *Annales d'Économie et de Statistique*, (91/92):175–187, JUL 2008.

[3] Alberto Abadie and Guido W. Imbens. Bias-corrected matching estimators for average treatment effects. *Journal of Business & Economic Statistics*, 29(1):1–11, JAN 2011.

[4] Alberto Abadie and Guido W. Imbens. Matching on the estimated propensity score. *Econometrica*, 84(2):781–807, MAR 2016.

[5] Alberto Abadie and Jann Spiess. Robust Post-Matching Inference. *Journal of the American Statistical Association*, JAN 2021.

[6] Younathan Abdia, K. B. Kulasekera, Somnath Datta, Maxwell Boakye, and Maiying Kong. Propensity scores based methods for estimating average treatment effect and average treatment effect among treated: a comparative study. *Biometrical Journal*, 59(5):967–985, SEP 2017.

[7] David B. Abrams, Allison M. Glasser, Jennifer L. Pearson, Andrea C. Villanti, Lauren K. Collins, and Raymond S. Niaura. Harm Minimization and Tobacco Control: Reframing Societal Views of Nicotine Use to Rapidly Save Lives. *Annual Review of Public Health*, 39(1):193–213, APR 2018.

[8] Joseph Antonelli, Matthew Cefalu, Nathan Palmer, and Denis Agniel. Doubly robust matching estimators for high dimensional confounding adjustment. *Biometrics*, 74(4):1171–1179, DEC 2018.

[9] Peter C. Austin. An introduction to propensity score methods for reducing the effects of confounding in observational studies. *Multivariate behavioral research*, 46(3):399–424, MAY 2011.

[10] Peter C. Austin. A comparison of 12 algorithms for matching on the propensity score. *Statistics in medicine*, 33(6):1057–1069, MAR 2014.

[11] Peter C. Austin, Nathaniel Jembere, and Maria Chiu. Propensity score matching and complex surveys. *Statistical Methods in Medical Research*, 27(4):1240–1257, JUL 2018.

[12] Peter C. Austin and Dylan S. Small. The use of bootstrapping when using propensity-score matching without replacement: a simulation study. *Statistics in Medicine*, 33(24):4306–4319, OCT 2014.

[13] Tarik Benmarhnia, John P. Pierce, Eric Leas, Martha M. White, David R. Strong, Madison L. Noble, and Dennis R. Trinidad. Can e-cigarettes and pharmaceutical aids increase smoking cessation and reduce cigarette consumption? Findings from a nationally representative cohort of American smokers. *American journal of epidemiology*, 187(11):2397–2404, NOV 2018.

[14] Kaitlyn M. Berry, Lindsay M. Reynolds, Jason M. Collins, Michael B. Siegel, Jessica L. Fetterman, Naomi M. Hamburg, Aruni Bhatnagar, and Emelia J. Benjamin. E-cigarette initiation and associated changes in smoking cessation and reduction: the Population Assessment of Tobacco and Health Study, 2013–2015. *Tobacco control*, 28(1):42–49, JAN 2019.

[15] Hugo Bodory, Lorenzo Camponovo, Martin Huber, and Michael Lechner. The finite sample performance of inference methods for propensity score matching and weighting estimators. *Journal of Business Economic Statistics*, 38(1):183–200, JAN 2020.

[16] Ron Borland, Timea R. Partos, and K. M. Cummings. Systematic biases in cross-sectional community studies may underestimate the effectiveness of stop-smoking medications. *Nicotine & Tobacco Research*, 14(12):1483–1487, DEC 2012.

[17] M A. Brookhart, Sebastian Schneeweiss, Kenneth J. Rothman, Robert J. Glynn, Jerry Avorn, and Til Stürmer. Variable selection for propensity score models. *American Journal of Epidemiology*, 163(12):1149–1156, APR 2006.

[18] Weihua Cao, Anastasios A. Tsiatis, and Marie Davidian. Improving efficiency and robustness of the doubly robust estimator for a population mean with incomplete data. *Biometrika*, 96(3):723–734, SEP 2009.

[19] Ruifeng Chen, John P. Pierce, Eric C. Leas, Tarik Benmarhnia, David R. Strong, Martha M. White, Matthew D. Stone, Dennis R. Trinidad, Sara B. McMenamin, and Karen Messer. Effectiveness of e-cigarettes as aids for smoking cessation: evidence from the PATH Study cohort, 2017-2019. *Tobacco Control*, FEB 2022. Epub ahead of print.

[20] Ruifeng Chen, John P. Pierce, Eric C. Leas, Martha M. White, Sheila Kealey, David R. Strong, Dennis R. Trinidad, Tarik Benmarhnia, and Karen Messer. Use of Electronic Cigarettes to Aid Long-Term Smoking Cessation in the United States: Prospective Evidence From the PATH Cohort Study. *American Journal of Epidemiology*, 189(12):1529–1537, DEC 2020.

[21] Hongying Dai and Adam M. Leventhal. Association of electronic cigarette vaping and subsequent smoking relapse among former smokers. *Drug and alcohol dependence*, 199:10–17, JUN 2019.

[22] Hongying Dai and Adam M. Leventhal. Prevalence of e-Cigarette Use Among Adults in the United States, 2014-2018. *JAMA*, 322(18):1824–1827, NOV 2019.

[23] Marco daCosta DiBonaventura, Jan-Samuel Wagner, Yong Yuan, Gilbert L'Italien, Paul Langley, and W. R. Kim. Humanistic and economic impacts of hepatitis C infection in the United States. *Journal of Medical Economics*, 13(4):709–718, DEC 2010.

[24] Olive J. Dunn. Multiple comparisons among means. *Journal of the American statistical association*, 56(293):52–64, MAR 1961.

[25] Bradley Efron. *The jackknife, the bootstrap and other resampling plans*. CBMS-NSF Regional Conference Series in Applied Mathematics, JAN 1982.

[26] Bradley Efron and Charles Stein. The Jackknife Estimate of Variance. *The Annals of Statistics)*, 9(3):586–596, MAY 1981.

[27] Amy L. Fairchild, Ronald Bayer, and Ju S. Lee. The E-Cigarette Debate: What Counts as Evidence? *American Journal of Public Health*, 109(7):1000–1006, JUL 2019.

[28] Elizabeth A. Gilpin, John P. Pierce, Arthur J. Farkas, and Arthur J. Farkas. Duration of smoking abstinence and success in quitting. *Journal of the National Cancer Institute*, 89(8):572, APR 1997.

[29] Tilmann Gneiting and Adrian E. Raftery. Strictly proper scoring rules, prediction, and estimation. *Journal of the American statistical Association*, 102(477):359–378, MAR 2007.

[30] Lawrence W. Green, Jonathan E. Fielding, and Ross C. Brownson. The Debate About Electronic Cigarettes: Harm Minimization or the Precautionary Principle. *Annual Review of Public Health*, 39(1):189–191, APR 2018.

[31] Peter Hajek, Anna Phillips-Waller, Dunja Przulj, Francesca Pesola, Katie M. Smith, Natalie Bisal, Jinshuo Li, Steve Parrott, Peter Sasieni, Lynne Dawkins, Louise Ross, Maciej Goniewicz, Qi Wu, and Hayden J. McRobbie. A randomized trial of e-cigarettes versus nicotine-replacement therapy. *The New England Journal of Medicine*, 380(7):629–637, FEB 2019.

[32] Scott D. Halpern, Michael O. Harhay, Kathryn Saulsgiver, Christine Brophy, Andrea B. Troxel, and Kevin G. Volpp. A pragmatic trial of e-cigarettes, incentives, and drugs for smoking cessation. *The New England Journal of Medicine*, 378(24):2302–2310, JUN 2018.

[33] Ben B. Hansen. The prognostic analogue of the propensity score. *Biometrika*, 95(2):481–488, JUN 2008.

[34] Jennifer Hill and Jerome P. Reiter. Interval estimation for treatment effects using propensity score matching. *Statistics in Medicine*, 25(13):2230–2256, JUL 2006.

[35] Daniel E. Ho, Kosuke Imai, Gary King, and Elizabeth A. Stuart. Matching as Nonparametric Preprocessing for Reducing Model Dependence in Parametric Causal Inference. *Political Analysis*, 15(3):199–236, JAN 2007.

[36] Zonghui Hu, Dean A. Follmann, and Naisyin Wang. Estimation of mean response via the effective balancing score. *Biometrika*, 101(3):613–624, SEP 2014.

[37] Jidong Huang, Zongshuan Duan, Julian Kwok, Steven Binns, Lisa E. Vera, Yoonsang Kim, Glen Szczypka, and Sherry L. Emery. Vaping versus JUULing: how the extraordinary growth and marketing of JUUL transformed the US retail e-cigarette market. *Tobacco Control*, 28(2):146–151, FEB 2019.

[38] Andrew Hyland, Bridget K. Ambrose, Kevin P. Conway, Nicolette Borek, Elizabeth Lambert, Charles Carusi, Kristie Taylor, Scott Crosse, Geoffrey T. Fong, K. M. Cummings, David Abrams, John P. Pierce, James Sargent, Karen Messer, Maansi Bansal-Travers, Ray Niaura, Donna Vallone, David Hammond, Nahla Hilmi, Jonathan Kwan, Andrea Piesse, Graham Kalton, Sharon Lohr, Nick Pharris-Ciurej, Victoria Castleman, Victoria R. Green, Greta Tessman, Annette Kaufman, Charles Lawrence, Dana M van Bemmel, Heather L. Kimmel, Ben Blount, Ling Yang, Barbara O'Brien, Cindy Tworek, Derek Alberding, Lynn C. Hull, Yu-Ching Cheng, David Maklan, Cathy L. Backinger, and Wilson M. Compton. Design and methods of the Population Assessment of Tobacco and Health (PATH) Study. *Tobacco Control*, 26(4):371–378, JUL 2017.

[39] Guido W Imbens. Nonparametric estimation of average treatment effects under exogeneity: A review. *Review of Economics and statistics*, 86(1):4–29, FEB 2004.

[40] David R. Judkins. Fay's method for variance estimation. *Journal of Official Statistics*, 6(3):223, SEP 1990.

[41] Sara Kalkhoran, Yuchiao Chang, and Nancy A. Rigotti. Electronic Cigarette Use and Cigarette Abstinence Over 2 Years Among U.S. Smokers in the Population Assessment of Tobacco and Health Study. *Nicotine & Tobacco Research*, 22(5):728–733, APR 2020.

[42] Joseph D. Y. Kang and Joseph L. Schafer. Demystifying double robustness: A comparison of alternative strategies for estimating a population mean from incomplete data. *Statistical science*, 22(4):523–539, NOV 2007.

[43] Karin A. Kasza, K. M. Cummings, Matthew J. Carpenter, Monica E. Cornelius, Andrew J. Hyland, and Geoffrey T. Fong. Use of stop-smoking medications in the United States before and after the introduction of varenicline. *Addiction*, 110(2):346–355, FEB 2015.

[44] Edward L. Korn and Barry I. Graubard. Examples of Differing Weighted and Unweighted Estimates from a Sample Survey. *The American Statistician*, 49(3):291–295, AUG 1995.

[45] D. Krewski and J. N. K. Rao. Inference From Stratified Samples: Properties of the Linearization, Jackknife and Balanced Repeated Replication Methods. *The Annals of Statistics*, 9(5):1010–1019, SEP 1981.

[46] Daniel Krewski. On the stability of some replication variance estimators in the linear case. *Journal of Statistical Planning and Inference*, 2(1):45–51, JAN 1978.

[47] Dennis Z. Kuo, T. M. Bird, and J. M. Tilford. Associations of family-centered care with health care outcomes for children with special health care needs. *Maternal and child health journal*, 15(6):794–805, AUG 2011.

[48] Finbarr P. Leacy and Elizabeth A. Stuart. On the joint use of propensity and prognostic scores in estimation of the average treatment effect on the treated: a simulation study. *Statistics in medicine*, 33(20):3488–3508, SEP 2014.

[49] Eric C. Leas, John P. Pierce, Tarik Benmarhnia, Martha M. White, Madison L. Noble, Dennis R. Trinidad, and David R. Strong. Effectiveness of pharmaceutical smoking cessation aids in a nationally representative cohort of American smokers. *Journal of the National Cancer Institute*, 110(6):581–587, JUN 2018.

[50] Michael Lechner. Some practical issues in the evaluation of heterogeneous labour market programmes by matching methods. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 165(1):59–82, FEB 2002.

[51] Myoung-jae Lee and Sanghyeok Lee. Double robustness without weighting. *Statistics & Probability Letters*, 146:175–180, MAR 2019.

[52] David Lenis, Trang Q. Nguyen, Nianbo Dong, and Elizabeth A. Stuart. It's all about balance: propensity score matching in the context of complex survey data. *Biostatistics*, 20(1):147–163, JAN 2019.

[53] David T. Levy, Ron Borland, Eric N. Lindblom, Maciej L. Goniewicz, Rafael Meza, Theodore R. Holford, Zhe Yuan, Yuying Luo, Richard J. O'Connor, Raymond Niaura, and David B. Abrams. Potential deaths averted in USA by replacing cigarettes with e-cigarettes. *Tobacco Control*, 27(1):18–25, JAN 2018.

[54] Roderick J. Little and Donald B. Rubin. Causal effects in clinical and epidemiological studies via potential outcomes: concepts and analytical approaches. *Annual Review of Public Health*, 21(1):121–145, MAY 2000.

[55] Roderick J. Little and Sonya Vartivarian. On weighting the rates in non-response weights. *Statistics in medicine*, 22(9):1589–1599, MAY 2003.

[56] Roderick JA. Little. Inference with survey weights. *Journal of Official Statistics*, 7(4):405–424, DEC 1991.

[57] Sharon L. Lohr. *Sampling: Design and Analysis*. Cengage Learning, DEC 1999.

[58] Thomas Lumley. Analysis of complex survey samples. *Journal of statistical software*, 9(8):1–9, APR 2004.

[59] Ann McNeill1, Leonie S. Brose1, Robert Calder, Linda Bauld, and Debbie Robson. Evidence review of e-cigarettes and heated tobacco products 2018: *A report commissioned by Public Health England. London: Public Health England* 6, 2018.

[60] Richard Miech, Lloyd Johnston, Patrick M. O'Malley, Jerald G. Bachman, and Megan E. Patrick. Trends in adolescent vaping, 2017–2019. *The New England Journal of Medicine*, 381(15):1490–1491, OCT 2019.

[61] Paul L. Morgan, Michelle L. Frisco, George Farkas, and Jacob Hibel. A propensity score matching analysis of the effects of special education services. *The Journal of special education*, 43(4):236–254, FEB 2010.

[62] Stephen L. Morgan and David J. Harding. Matching estimators of causal effects: Prospects and pitfalls in theory and practice. *Sociological methods & research*, 35(1):3–60, AUG 2006.

[63] Engineering National Academies of Sciences and Medicine. *Public Health Consequences of E-Cigarettes*. Washington, DC: The National Academies Press, MAY 2018.

[64] US Department of Commerce C.B. National Cancer Institute-sponsored. Tobacco Use Supplement to the Current Population Survey: What is the TUS-CPS? MAR 2014.

[65] US Department of Health and Human Services. Smoking cessation: a report of the Surgeon General. *Atlanta: US Department of Health and Human Services*, 2020.

[66] John P. Pierce, Tarik Benmarhnia, Ruifeng Chen, Martha M. White, David B. Abrams, Bridget K. Ambrose, Carlos Blanco, Nicolette Borek, Kelvin Choi, Blair Coleman, Wilson M. Compton, Kenneth M. Cummings, Cristine D. Delnevo, Tara Elton-Marshall, Maciej L. Goniewicz, Shannon Gravely, Geoffrey T. Fong, Dorothy Hatsukami, James Henrie, Karin A. Kasza, Sheila Kealey, Heather L. Kimmel, Jean Limpert, Raymond S. Niaura, Carolina Ramôa, Eva Sharma, Marushka L. Silveira, Cassandra A. Stanton, Michael B. Steinberg, Ethel Taylor, Maansi Bansal-Travers, Dennis R. Trinidad, Lisa D. Gardner, Andrew Hyland, Samir Soneji, and Karen Messer. Role of e-cigarettes and pharmacotherapy during attempts to quit cigarette smoking: The PATH Study 2013-16. *PLoS One*, 15(9):e0237938, SEP 2020.

[67] John P. Pierce, Karen Messer, Eric C. Leas, Sheila Kealey, Martha M. White, and Tarik Benmarhnia. A source of bias in studies of e-cigarettes and smoking cessation. *Nicotine & Tobacco Research*, 22(5):861–862, MAY 2020.

[68] Samuel D. Pimentel, Rachel R. Kelz, Jeffrey H. Silber, and Paul R. Rosenbaum. Large, Sparse Optimal Matching With Refined Covariate Balance in an Observational Study of the Health Outcomes Produced by New Surgeons. *Journal of the American Statistical Association*, 110(510):515–527, APR 2015.

[69] National Addiction HIV Data Archive Program. Population Assessment of Tobacco and Health (PATH) Study Series. https://www.icpsr.umich.edu/icpsrweb/NAHDAP/ series/606. Accessed June 22, 2020.

[70] J. N. K. Rao and Jun Shao. Modified balanced repeated replication for complex survey data. *Biometrika*, 86(2):403–415, JUN 1999.

[71] James M. Robins, Steven D. Mark, and Whitney K. Newey. Estimating exposure effects by modelling the expectation of exposure conditional on confounders. *Biometrics*, 48(2):479–495, JUN 1992.

[72] James M. Robins, Andrea Rotnitzky, and Lue P. Zhao. Estimation of regression coefficients when some regressors are not always observed. *Journal of the American statistical Association*, 89(427):846–866, SEP 1994.

[73] Paul R. Rosenbaum. *Observation and experiment: An introduction to causal inference*. Harvard University Press, 2017.

[74] Paul R. Rosenbaum and Donald B. Rubin. The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70(1):41–55, APR 1983.

[75] Paul R. Rosenbaum and Donald B. Rubin. Reducing bias in observational studies using subclassification on the propensity score. *Journal of the American statistical Association*, 79(387):516–524, SEP 1984.

[76] Paul R. Rosenbaum and Donald B. Rubin. Constructing a control group using multivariate matched sampling methods that incorporate the propensity score. *The American Statistician*, 39(1):33–38, FEB 1985.

[77] Brian L. Rostron, Catherine G. Corey, Joanne T. Chang, Dana M. van Bemmel, Mollie E. Miller, and Cindy M. Chang. Associations of cigarettes smoked per day with biomarkers of exposure among US adult cigarette smokers in the population assessment of tobacco and health (PATH) study wave 1 (2013–2014). *Cancer Epidemiology and Prevention Biomarkers*, 28(9):1443–1453, SEP 2019.

[78] Donald B. Rubin. Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology*, 66(5):688–701, OCT 1974.

[79] Donald B. Rubin. *Matched sampling for causal effects*. Cambridge University Press, SEP 2006.

[80] Donald B. Rubin and Neal Thomas. Characterizing the Effect of Matching Using Linear Propensity Score Methods with Normal Distributions. *Biometrika*, 79(4):797–809, DEC 1992.

[81] Donald B. Rubin and Neal Thomas. Matching using estimated propensity scores: relating theory to practice. *Biometrics*, 52(1):249–264, MAR 1996.

[82] Shaun R. Seaman and Stijn Vansteelandt. Introduction to double robust methods for incomplete data. *Statistical science*, 33(2):184–197, 2018.

[83] Robert J. Serfling. *Approximation Theorems of Mathematical Statistics*. New York, NY, 2009.

[84] Jun Shao and Dongsheng Tu. *The Jackknife and Bootstrap*. Springer-Verlag, Inc., New York, 1995.

[85] Heng Shu and Zhiqiang Tan. Improved estimation of average treatment effects on the treated: Local efficiency, double robustness, and beyond. AUG 2018. arXiv.1808.01408.

[86] Samir S. Soneji, Hai-Yen Sung, Brian A. Primack, John P. Pierce, and James D. Sargent. Quantifying population-level health benefits and harms of e-cigarette use in the United States. *PLoS One*, 13(3):e0193328, MAR 2018.

[87] David R. Strong, Jennifer Pearson, Sarah Ehlke, Thomas Kirchner, David Abrams, Kristie Taylor, Wilson M. Compton, Kevin P. Conway, Elizabeth Lambert, Victoria R. Green, Lynn C. Hull, Sarah E. Evans, Michael Cummings, Maciej Goniewicz, Andrew Hyland, and Raymond Niaura. Indicators of dependence for different types of tobacco product users: Descriptive findings from Wave 1 (2013–2014) of the Population Assessment of Tobacco and Health (PATH) study. *Drug and alcohol dependence*, 178:257–266, SEP 2017.

[88] Elizabeth A. Stuart. Matching methods for causal inference: A review and a look forward. *Statistical Science*, 25(1):1–21, FEB 2010.

[89] Elizabeth A. Stuart, Gary King, Kosuke Imai, and Daniel Ho. MatchIt: nonparametric preprocessing for parametric causal inference. *Journal of statistical software*, 42(8):1–28, 2011.

[90] Laura Sweet, Theodore M. Brasky, Sarah Cooper, Nathan Doogan, Alice Hinton, Elizabeth G. Klein, Haikady Nagaraja, Amanda Quisenberry, Wenna Xi, and Mary E. Wewers. Quitting behaviors among dual cigarette and e-cigarette users and cigarette smokers enrolled in the tobacco user adult cohort. *Nicotine & Tobacco Research*, 21(3):278–284, MAR 2019.

[91] Carl-Erik Särndal, Bengt Swensson, and Jan Wretman. *Model Assisted Survey Sampling*. Springer Series in Statistics, OCT 2003.

[92] Zhiqiang Tan. Bounded, efficient and doubly robust estimation with inverse weighting. *Biometrika*, 97(3):661–682, SEP 2010.

[93] Yebin Tao and Haoda Fu. Doubly robust estimation of the weighted average treatment effect for a target population. *Statistics in medicine*, 38(3):315–325, OCT 2018.

[94] Nguyen T. T. Thuong. Impact of health insurance on healthcare utilisation patterns in Vietnam: a survey-based analysis with propensity score matching method. *BMJ open*, 10(10):e040062, OCT 2020.

[95] Robert Tibshirani. Regression Shrinkage and Selection via the Lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, 58(1):267–288, JAN 1996.

[96] Yves Tillé. *Sampling Algorithms*. Springer New York, MAR 2006.

[97] Human Services United States Department of Health, National Institutes of Health, National Institute on Drug Abuse, United States Department of Health, Human Services, Food & Drug Administration, and Center for Tobacco Products. Population Assessment of Tobacco and Health (PATH) Study [United States] Restricted-Use Files. Inter-university Consortium for Political and Social Research [distributor], 2021-12-16.

[98] Karel Vermeulen and Stijn Vansteelandt. Bias-reduced doubly robust estimation. *Journal of the American Statistical Association*, 110(511):1024–1036, JUL 2015.

[99] Yongji Wang, Hongwei Cai, Chanjuan Li, Zhiwei Jiang, Ling Wang, Jiugang Song, and Jielai Xia. Optimal Caliper Width for Propensity Score Matching of Three Treatment Groups: A Monte Carlo Study. *PloS one)*, 8(12):e81045, DEC 2013.

[100] Kenneth E. Warner and David Mendez. E-cigarettes: Comparing the Possible Risks of Increasing Smoking Initiation with the Potential Benefits of Increasing Smoking Cessation. *Nicotine & Tobacco Research*, 21(1):41–47, JAN 2019.

[101] Shannon L. Watkins, Johannes Thrul, Wendy Max, and Pamela M. Ling. Real-world effectiveness of smoking cessation strategies for young and older adults: findings from a nationally representative cohort. *Nicotine & Tobacco Research*, 22(9):1560–1568, SEP 2020.