Improved Strategies Using the Propensity Score to Estimate the Effect of a Healthcare Policy

BY
LAUREN AMELIA ABDERHALDEN
B.A. Mathematics, University of California at San Diego, 2007
M.S. Statistics, San Diego State University, 2009

DISSERTATION
Submitted as partial fulfillment of the requirements for the degree of Doctor of Philosophy in Public Health Sciences (Biostatistics) in the Graduate College of the University of Illinois at Chicago, 2018

Chicago, Illinois

Defense Committee:
Hakan Demirtas, Chair and Advisor
Saria Awadalla, Biostatistics
Hua Yun Chen, Biostatistics
Sally Freels, Biostatistics
Kevin Stroupe, Edward Hines, Jr. VA Hospital
ACKNOWLEDGEMENTS

To my husband and our growing family
ACKNOWLEDGMENTS

Firstly, I would like to express great appreciation to my advisor Dr. Hakan Demirtas, for the continuous support of my Ph.D. study, high expectations of produced work, patience, humor, life insights, smart guidance, motivation, and immense knowledge. His continued guidance helped me not only see this dissertation to completion but made me into a better researcher all around.

In addition to my advisor, I would like to thank the other members of my dissertation committee: Dr. Kevin Stroupe, Dr. Sally Freels, Dr. Hua Yun Chen, and Dr. Saria Awadalla for their support, participation, and helpful suggestions. It was the exciting research with Dr. Stroupe that inspired the topic of this dissertation.

Benn Williams writing sessions, Write-On, as well as his book recommendation, How to Write a Lot by Paul Silvia instilled the motivation in me to define scheduled writing times. I will continue a writing schedule in the future.

I would like to thank my parents: my dad who helped me greatly with additional computing infrastructure to carry out the vast amount of simulations, and both parents for supporting me through this journey and life in general.

Finally, it is impossible to put into words the immense thank you to dearest husband Felician. You supported me through the tough times and celebrated with me the great, provided insight, motivation, your laptop which will never be the same after months of simulations, warm hugs, and understanding at times when I needed it most.
# TABLE OF CONTENTS

1. **INTRODUCTION** ............................................................................................................................................. 1

2. **BACKGROUND** ............................................................................................................................................... 6
   2.1 Natural experiments in healthcare policy estimation .............................................................................. 6
   2.2 Difference-in-differences .......................................................................................................................... 6
      2.2.1 Structure of the data considered ......................................................................................... 11
   2.3 Counterfactual framework ................................................................................................................... 12
   2.4 Propensity score ...................................................................................................................................... 16
   2.5 Inverse probability weighting .......................................................................................................... 17
   2.6 Augmented inverse probability weighted estimator ........................................................................... 19
   2.7 Previous approaches ........................................................................................................................ 20
      2.7.1 Propensity score matching ............................................................................................... 20
      2.7.2 Propensity score weighting and sensitivity analysis .......................................................... 21
      2.7.3 Four-group weighted design .......................................................................................... 22
   2.8 Forward thoughts ............................................................................................................................. 25

3. **METHODS** .................................................................................................................................................... 27
   3.1 Two-way fixed effects regression estimator ...................................................................................... 27
   3.2 Separate time weighted regression estimator ................................................................................... 28
   3.3 Four-group weighted design .............................................................................................................. 29
   3.4 Empirical standard error estimates ..................................................................................................... 29
      3.4.1 Basic bootstrap method ...................................................................................................... 29
      3.4.2 Percentile bootstrap method .......................................................................................... 30
   3.5 Augmented inverse probability weighted difference-in-differences estimator .............................. 31
   3.6 Estimators for comparison ................................................................................................................. 32
   3.7 Estimate of interest ............................................................................................................................ 33
   3.8 Discussion ............................................................................................................................................ 34

4. **EXAMPLE FROM THE MEDICAL EXPENDITURE PANEL SURVEY** ............................................................... 40
   4.1 Introduction ........................................................................................................................................... 40
   4.2 Methods ................................................................................................................................................. 40
      4.2.1 Data source and study population ....................................................................................... 40
      4.2.2 Statistical analysis ............................................................................................................. 43
   4.3 Results ................................................................................................................................................ 45
      4.3.1 Study population ............................................................................................................... 45
      4.3.2 Propensity score weight diagnostics .............................................................................. 53
      4.3.3 Prescription utilization and cost .......................................................................................... 64
   4.4 Discussion ............................................................................................................................................ 86

5. **SIMULATION STUDIES** ............................................................................................................................. 89
   5.1 Simulation using artificial data .......................................................................................................... 89
      5.1.1 Algorithm ........................................................................................................................... 90
      5.1.2 Evaluation criteria ........................................................................................................... 91
      5.1.3 Results ................................................................................................................................... 93
<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>5.2 Simulation devised around real data</td>
<td>99</td>
</tr>
<tr>
<td>5.2.1 Data generation</td>
<td>99</td>
</tr>
<tr>
<td>5.2.2 Design elements</td>
<td>101</td>
</tr>
<tr>
<td>5.2.3 Algorithm</td>
<td>106</td>
</tr>
<tr>
<td>5.2.4 Examination of simulated datasets</td>
<td>108</td>
</tr>
<tr>
<td>5.2.5 Results</td>
<td>123</td>
</tr>
<tr>
<td>6. DISCUSSION</td>
<td>154</td>
</tr>
<tr>
<td>7. SUMMARY OF CONTRIBUTIONS</td>
<td>165</td>
</tr>
<tr>
<td>CITED LITERATURE</td>
<td>167</td>
</tr>
<tr>
<td>APPENDICES</td>
<td>174</td>
</tr>
<tr>
<td>APPENDIX A</td>
<td>175</td>
</tr>
<tr>
<td>APPENDIX B</td>
<td>185</td>
</tr>
<tr>
<td>APPENDIX C</td>
<td>196</td>
</tr>
<tr>
<td>APPENDIX D</td>
<td>205</td>
</tr>
<tr>
<td>APPENDIX E</td>
<td>210</td>
</tr>
<tr>
<td>APPENDIX F</td>
<td>217</td>
</tr>
<tr>
<td>APPENDIX G</td>
<td>220</td>
</tr>
<tr>
<td>VITA</td>
<td>238</td>
</tr>
</tbody>
</table>
# LIST OF TABLES

<table>
<thead>
<tr>
<th>TABLE</th>
<th>PAGE</th>
</tr>
</thead>
<tbody>
<tr>
<td>I. THE DIFFERENCE-IN-DIFFERENCES EFFECT ESTIMATE</td>
<td>7</td>
</tr>
<tr>
<td>II. DIFFERENCE-IN-DIFFERENCES AS REPRESENTED BY A TWO-WAY FIXED EFFECTS REGRESSION MODEL</td>
<td>10</td>
</tr>
<tr>
<td>III. GROUP IDENTIFICATION IN THE FOUR-GROUP WEIGHTED DESIGN</td>
<td>23</td>
</tr>
<tr>
<td>IV. STUDY SAMPLE DEMOGRAPHICS</td>
<td>47</td>
</tr>
<tr>
<td>V. CORRELATIONS OF STUDY COVARIATES</td>
<td>50</td>
</tr>
<tr>
<td>VI. STUDY SAMPLE DEMOGRAPHICS IN THE SUBGROUP</td>
<td>51</td>
</tr>
<tr>
<td>VII. CORRELATIONS OF STUDY COVARIATES IN THE SUBGROUP</td>
<td>52</td>
</tr>
<tr>
<td>VIII. DISTRIBUTION OF PROPENSITY SCORES AND DERIVED WEIGHTS IN SEPARATE TIME WEIGHTED AND FOUR-GROUP WEIGHTED MODELS</td>
<td>54</td>
</tr>
<tr>
<td>IX. SEPARATE TIME WEIGHTED STUDY SAMPLE DEMOGRAPHICS</td>
<td>61</td>
</tr>
<tr>
<td>X. FOUR-GROUP WEIGHTED STUDY SAMPLE DEMOGRAPHICS</td>
<td>62</td>
</tr>
<tr>
<td>XI. SUMMARY STATISTICS OF PERCENT STANDARDIZED DIFFERENCES</td>
<td>63</td>
</tr>
<tr>
<td>XII. DISTRIBUTION OF PROPENSITY SCORES AND DERIVED WEIGHTS IN SEPARATE TIME WEIGHTED AND FOUR-GROUP WEIGHTED MODELS IN THE SUBGROUP</td>
<td>65</td>
</tr>
<tr>
<td>XIII. SUMMARY STATISTICS OF PERCENT STANDARDIZED DIFFERENCES IN THE SUBGROUP</td>
<td>70</td>
</tr>
<tr>
<td>XIV. UNWEIGHTED MEANS FOR NUMBER OF PRESCRIPTIONS, COST OF MEDICAID PRESCRIPTIONS, AND TOTAL CHARGES</td>
<td>71</td>
</tr>
<tr>
<td>XV. REGRESSION COEFFICIENT ESTIMATES FROM COVARIATE ADJUSTED TWO-WAY FIXED EFFECTS MODELS</td>
<td>74</td>
</tr>
<tr>
<td>XVI. DIFFERENCE-IN-DIFFERENCES ESTIMATES OF NUMBER AND COST OF PRESCRIPTIONS USING 13 ESTIMATORS</td>
<td>78</td>
</tr>
<tr>
<td>XVII. REGRESSION COEFFICIENT ESTIMATES FROM COVARIATE ADJUSTED TWO-WAY FIXED EFFECTS MODELS OF TOTAL CHARGES IN THE SUBGROUP</td>
<td>83</td>
</tr>
<tr>
<td>Table</td>
<td>Description</td>
</tr>
<tr>
<td>-------</td>
<td>-------------</td>
</tr>
<tr>
<td>XVIII.</td>
<td>DID ESTIMATES OF TOTAL CHARGES USING 13 ESTIMATORS IN THE SUBGROUP</td>
</tr>
<tr>
<td>XIX.</td>
<td>SIMULATION SETTINGS OF SELECTION BIAS</td>
</tr>
<tr>
<td>XX.</td>
<td>SIMULATION RESULTS IN A STUDY USING ARTIFICIAL DATA</td>
</tr>
<tr>
<td>XXI.</td>
<td>SIMULATION SETTINGS OF DESIRED EFFECT SIZES</td>
</tr>
<tr>
<td>XXII.</td>
<td>SIMULATION SETTINGS OF ACTUAL EFFECT SIZE MULTIPIERS</td>
</tr>
<tr>
<td>XXIII.</td>
<td>COMPARISON OF THE ACTUAL AND BIG SIMULATED DATASETS</td>
</tr>
<tr>
<td>XXIV.</td>
<td>REGRESSION COEFFICIENT ESTIMATES FROM ACTUAL AND SIMULATED DATASETS FOR NUMBER OF PRESCRIPTIONS</td>
</tr>
<tr>
<td>XXV.</td>
<td>COMPARISON OF THE ACTUAL AND BIG SIMULATED DATASETS IN THE SUBGROUP</td>
</tr>
<tr>
<td>XXVI.</td>
<td>REGRESSION COEFFICIENT ESTIMATES FROM ACTUAL AND SIMULATED DATASETS FOR TOTAL CHARGES</td>
</tr>
<tr>
<td>XXVII.</td>
<td>SIMULATION RESULTS IN 4 TYPES OF SMALL POVERTY COVARIATE CONFOUNDING FOR NUMBER OF PRESCRIPTIONS</td>
</tr>
<tr>
<td>XXVIII.</td>
<td>SIMULATION RESULTS IN 3 TYPES OF MODERATE POVERTY COVARIATE CONFOUNDING FOR NUMBER OF PRESCRIPTIONS</td>
</tr>
<tr>
<td>XXIX.</td>
<td>SIMULATION RESULTS IN 3 TYPES OF LARGE POVERTY COVARIATE CONFOUNDING FOR NUMBER OF PRESCRIPTIONS</td>
</tr>
<tr>
<td>XXX.</td>
<td>SIMULATION RESULTS IN 4 TYPES OF SMALL AGE COVARIATE CONFOUNDING FOR NUMBER OF PRESCRIPTIONS</td>
</tr>
<tr>
<td>XXXI.</td>
<td>SIMULATION RESULTS IN 3 TYPES OF MODERATE AGE COVARIATE CONFOUNDING FOR NUMBER OF PRESCRIPTIONS</td>
</tr>
<tr>
<td>XXXII.</td>
<td>SIMULATION RESULTS IN 3 TYPES OF LARGE AGE COVARIATE CONFOUNDING FOR NUMBER OF PRESCRIPTIONS</td>
</tr>
<tr>
<td>XXXIII.</td>
<td>SIMULATION RESULTS IN 4 TYPES OF SMALL POVERTY COVARIATE CONFOUNDING FOR TOTAL CHARGES</td>
</tr>
<tr>
<td>XXXIV.</td>
<td>SIMULATION RESULTS IN 3 TYPES OF MODERATE POVERTY COVARIATE CONFOUNDING FOR TOTAL CHARGES</td>
</tr>
</tbody>
</table>
LIST OF TABLES (continued)

XXXV. SIMULATION RESULTS IN 3 TYPES OF LARGE POVERTY COVARIATE CONFOUNDING FOR TOTAL CHARGES........................................................... 140

XXXVI. SIMULATION RESULTS IN 4 TYPES OF SMALL AGE COVARIATE CONFOUNDING FOR TOTAL CHARGES........................................................... 141

XXXVII. SIMULATION RESULTS IN 3 TYPES OF MODERATE AGE COVARIATE CONFOUNDING FOR TOTAL CHARGES......................................................... 142

XXXVIII. SIMULATION RESULTS IN 3 TYPES OF LARGE AGE COVARIATE CONFOUNDING FOR TOTAL CHARGES...................................................... 143
LIST OF FIGURES

<table>
<thead>
<tr>
<th>FIGURE</th>
<th>PAGE</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. The difference-in-differences treatment effect</td>
<td>8</td>
</tr>
<tr>
<td>2. Difference-in-differences treatment effect and counterfactual outcome</td>
<td>15</td>
</tr>
<tr>
<td>3. Flowchart of study sample inclusion</td>
<td>46</td>
</tr>
<tr>
<td>4. Distribution of propensity scores using a separate time weighted estimator</td>
<td>56</td>
</tr>
<tr>
<td>5. Distribution of propensity scores using a four-group weighted estimator</td>
<td>57</td>
</tr>
<tr>
<td>6. Distribution of propensity score weights using a separate time weighted estimator</td>
<td>58</td>
</tr>
<tr>
<td>7. Distribution of propensity score weights using a four-group weighted estimator</td>
<td>59</td>
</tr>
<tr>
<td>8. Distribution of propensity scores using a separate time weighted estimator in the subgroup</td>
<td>66</td>
</tr>
<tr>
<td>9. Distribution of propensity scores using a four-group weighted estimator in the subgroup</td>
<td>67</td>
</tr>
<tr>
<td>10. Distribution of propensity score weights using a separate time weighted estimator in the subgroup</td>
<td>68</td>
</tr>
<tr>
<td>11. Distribution of propensity score weights using a four-group weighted estimator in the subgroup</td>
<td>69</td>
</tr>
<tr>
<td>12. Forest plots comparing difference-in-differences estimators with linear model specification</td>
<td>79</td>
</tr>
<tr>
<td>13. Forest plots comparing difference-in-differences estimators with generalized linear model specification</td>
<td>80</td>
</tr>
<tr>
<td>14. Forest plots comparing estimators for total charges</td>
<td>85</td>
</tr>
<tr>
<td>15. Coverage rates in a study using artificial data</td>
<td>95</td>
</tr>
<tr>
<td>16. Correlational relationships among variables in actual and big simulated datasets: all data</td>
<td>110</td>
</tr>
<tr>
<td>17. Correlational relationships among variables in actual and big simulated datasets: group 1</td>
<td>111</td>
</tr>
<tr>
<td>18. Correlational relationships among variables in actual and big simulated datasets: group 2</td>
<td>112</td>
</tr>
</tbody>
</table>
LIST OF FIGURES (continued)

19. Correlational relationships among variables in actual and big simulated datasets: group 3 ................................................................. 113

20. Correlational relationships among variables in actual and big simulated datasets: group 4 ......................................................................................................................................... 114

21. Correlational relationships among variables in actual and big simulated datasets: all subgroup........................................................................................................................................... 118

22. Correlational relationships among variables in actual and big simulated datasets: subgroup, group 1 ......................................................................................................................................... 119

23. Correlational relationships among variables in actual and big simulated datasets: subgroup, group 2 ......................................................................................................................................... 120

24. Correlational relationships among variables in actual and big simulated datasets: subgroup, group 3 ......................................................................................................................................... 121

25. Correlational relationships among variables in actual and big simulated datasets: subgroup, group 4 ......................................................................................................................................... 122

26. Coverage results under small poverty covariate confounding for number of Medicaid prescriptions ................................................................................................................................................ 131

27. Coverage results under medium poverty covariate confounding for number of Medicaid prescriptions ................................................................................................................................................ 132

28. Coverage results under large poverty covariate confounding for number of Medicaid prescriptions ................................................................................................................................................ 133

29. Coverage results under small age covariate confounding for number of Medicaid prescriptions ................................................................................................................................................ 134

30. Coverage results under medium age covariate confounding for number of Medicaid prescriptions ................................................................................................................................................ 135

31. Coverage results under large age covariate confounding for number of Medicaid prescriptions ................................................................................................................................................ 136

32. Coverage results under small poverty covariate confounding for total charges ........ 144

33. Coverage results under medium poverty covariate confounding for total charges ...... 145

34. Coverage results under large poverty covariate confounding for total charges ........ 146

35. Coverage results under small age covariate confounding for total charges ........ 147

36. Coverage results under medium age covariate confounding for total charges ........ 148
LIST OF FIGURES (continued)

37. Coverage results under large age covariate confounding for total charges .................. 149
## LIST OF ABBREVIATIONS

<table>
<thead>
<tr>
<th>Abbreviation</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>ACA</td>
<td>Affordable Care Act</td>
</tr>
<tr>
<td>AIPW</td>
<td>augmented inverse probability weighted</td>
</tr>
<tr>
<td>ATE</td>
<td>average treatment effect</td>
</tr>
<tr>
<td>ATET</td>
<td>average treatment effect on the treated</td>
</tr>
<tr>
<td>CI</td>
<td>confidence interval</td>
</tr>
<tr>
<td>CT</td>
<td>common trend assumption</td>
</tr>
<tr>
<td>ESS</td>
<td>effective sample size</td>
</tr>
<tr>
<td>DID</td>
<td>difference-in-differences</td>
</tr>
<tr>
<td>IPW</td>
<td>inverse probability weighting</td>
</tr>
<tr>
<td>MEPS</td>
<td>Medical Expenditure Panel Survey</td>
</tr>
<tr>
<td>NEPT</td>
<td>no effect on the pre treatment population assumption</td>
</tr>
<tr>
<td>RMSE</td>
<td>root-mean-square error</td>
</tr>
<tr>
<td>RXMCD</td>
<td>cost of Medicaid prescription medications</td>
</tr>
<tr>
<td>RXTOT</td>
<td>number of prescription medications</td>
</tr>
<tr>
<td>SD</td>
<td>standard deviation</td>
</tr>
<tr>
<td>SITA</td>
<td>strongly ignorable treatment assignment</td>
</tr>
<tr>
<td>SUTVA</td>
<td>stable unit treatment value assumption</td>
</tr>
<tr>
<td>TOTTCH</td>
<td>total healthcare charges</td>
</tr>
</tbody>
</table>
SUMMARY

Healthcare policy data are usually observational by nature, where assignment to a policy group is seldom random. Difference-in-differences is a common approach for estimating the effect of a healthcare policy. Propensity score methods are helpful in mitigating selection bias between treatment groups. The conjunct use of difference-in-differences and propensity score methods are beginning to gain traction in healthcare policy effect estimation. Contemporary methods fail to take into account the sampling variability of propensity scores, and result in an erroneous precision.

This dissertation compares current estimators, offers a refinement to the standard error of current methods, and proposes a new estimator. Proposed methods use a resampling estimation approach to obtain the variance which is found to greatly improve precision by overcoming the erraticism introduced from unknown sources of propensity score sampling variability. Further, a proposed augmented inverse probability weighted estimator holds potential.

Estimation methods are evaluated to measure the effect of 2014 Medicaid expansion policy on number of prescriptions and cost of Medicaid prescriptions. It is found that Medicaid expansion contributed to an additional four medications per year for low-income individuals without significantly increasing the cost. Thirteen candidate estimators were applied to a real data example, evaluated in a simulation study using artificial data, and rigorously examined in a comprehensive simulation experiment devised around the real data example. It is found that the proposed estimation methods offer the much needed fix to the precision.
1. INTRODUCTION

Policymakers are tasked with providing recommendations to programs which best serve the public with the most efficient use of resources. Accurate and precise effect estimation of healthcare utilization, spending, and general health attributable to a policy or program is necessary for making informed decisions. In the majority of healthcare service settings, random allocation to a particular policy or program group is unfeasible. Natural experiments are increasingly being conducted to evaluate the effect of a healthcare policy or program. A fundamental problem still remains: selection bias. Often those who purchase or opt into a health insurance policy or program do so for specific, individualistic reasons such as access to care, medication formulary, or provider preference. Distributional differences between individuals enrolled or affected by the policy, and individuals not enrolled or unaffected by the policy, are likely to exist in this situation.

Studies in healthcare policy aim to evaluate the effect of a treatment (i.e. policy/program) in a setting where treatment assignment is almost never randomly allocated. Under random treatment allocation, there is no systematic or selective process influencing treatment assignment. In this setting, treatment groups are likely to look similar in distributional makeup and free from selection bias. However, randomized controlled trials in the effect estimation of a healthcare policy are nearly always infeasible. Reasons include ethical consideration and cost. In the absence of experimental design, selection bias of whom receives treatment and whom does not is of fundamental concern. Recent technological advances and information storage capabilities have facilitated the vast collection of observational healthcare data. Policies governing access to care, prescription formulary, beneficiary out-of-pocket expenses, and more are overseen by current healthcare policy. Policies are quickly rolled out, without proper
knowledge of how the policy might influence healthcare utilization and cost. In the setting of a natural experiment, it is imperative to incorporate techniques emulating the setting of random treatment allocation comparable to a randomized controlled trial.

Difference-in-differences (DID) approach is a popular method in applied economic research to estimate the effect of a policy or program in observational data (Heckman et al., 1999). The motivation behind DID estimation is simple: estimate the treatment effect from observational data by negating out any background, secular, environmental, or miscellaneous system changes to the outcome. Data are gathered pre and post intervention for both a treated and control group. Differences between pre and post outcomes are calculated for both a treated and control group. The difference between the differences is then calculated, as naturally, the difference-in-differences.

Propensity score methods were derived in statistical and biostatistical fields to mitigate bias arising from the imbalance of characteristics between treatment groups. Propensity score techniques are increasingly being applied in observational settings, when treatment assignment is not randomly allocated. The aim of these methods are to balance treatment groups on observed covariates. The propensity score is “the conditional probability of receiving the treatment given the observed covariates” (Rosenbaum, 2002). Adjustments using propensity scores generally fall into one of four categories: matching, covariate adjustment using the propensity score, stratification, or weighting.

Difference-in-differences approach is popular in applied economics/econometrics and propensity score techniques are popular in statistics/biostatistics. However, DID and propensity score techniques are seldom used in conjunction (Stuart et al., 2014; Fu et al., 2007). Recently, researchers have begun combining DID estimation with propensity score techniques to estimate
the effect of a healthcare policy while addressing the selection bias inherit to the data. A few
combined estimators have been proposed without guidance on how these conjunct
methodologies might alter point and coverage estimates. The use of propensity score methods to
address treatment group selection bias while failing to adjust for the prior estimation of the
propensity scores may result in an erroneous variance estimate. The alteration to the DID
variance estimand from simultaneously using DID and inverse probability weighting (IPW) by
propensity score methods is unknown. There is current debate involving estimation of the
sampling variability of treatment effects when using propensity scores (Austin and Small, 2014).
There has been a call for future research considering the performance of DID estimators under a
“broader range of specification features, outcomes, [and] heterogeneous program effects” (Ryan
et al., 2015). Furthermore, there has been a call for research to investigate the effect of
propensity score use on inferences (Williamson et al., 2012; Austin and Small, 2014). Existing
research offers limited guidance and empirical evidence on how precision in DID effect
estimation might be affected and refined by incorporating propensity score weighting methods to
account for selection bias.

We will focus on propensity score weighting approaches incorporated with DID
estimation. A parsimonious estimation strategy is to estimate the DID in a regression model with
inverse probability weighting of observations using the propensity score. There are several ways
a propensity score weighting strategy could be derived. In these approaches, the standard error
obtained from the regression model might not be precise if there exists sampling variability in
the propensity score. An alternate approach could be an augmented estimator, in which the basic
IPW estimator is augmented with the weighted average of regression estimators involving the
treatment and control groups. Such an estimator would be doubly robust, or insensitive to either
misspecification of the propensity score model or the outcome regression model. Guidance in the standard error estimation of various DID estimation methods have not been established for different types and magnitudes of covariate selection bias.

A simulation study was conducted to evaluate the statistical properties of several DID estimation methods under various scenarios of covariate selection bias. Accuracy and precision of the estimates were evaluated under seven alternate model specifications: (1) naïve two-way fixed effects regression estimator, (2) naïve two-way fixed effects regression estimator including covariates, (3) separate time weighted regression estimator, (4) separate time weighted regression estimator including covariates, (5) four-group weighted regression estimator, (6) four-group weighted regression estimator including covariates, and (7) augmented IPW estimator. Empirical bootstrap coverage estimates are proposed for models (4), (6), and (7). DID estimator specification and variance estimation procedures were evaluated using well-accepted measures of accuracy and precision. A conclusion is drawn, and recommendations are made for an accurate, precise IPW DID estimator.

Existing and proposed DID estimation strategies were applied to a real data example. Ideas and methods were illustrated to ascertain the causal effect of Medicaid expansion on healthcare utilization as cost. More specifically, the estimators were applied to determine the number of prescription medications and cost of Medicaid prescription medications attributed to Medicare expansion.

Estimator performance might vary under different types of selection bias (time, group and time, group by time), propensity score sampling variability, and propensity score or outcome model misspecification. Failing to adjust for the prior estimation of propensity scores might generate estimates with too conservative (i.e. too wide) coverage rates. It is hypothesized that the
variance is improved when it is estimated through an empirical approach by successfully accounting for the prior estimation of propensity scores. A doubly robust estimator is expected to be more robust to propensity score or outcome model misspecification. The augmented estimator with an empirical standard error is hypothesized to yield improved coverage rates under a broad range of selection bias, propensity score sampling variability, and model misspecification scenarios. Careful attention must be paid in estimator selection when combining propensity score methods and DID estimation to avoid improper inference. Incorporating methods from econometrics and biostatistics fields to obtain a DID estimate while addressing selection bias is a valuable and promising approach in healthcare policy effect estimation.

Current healthcare reform is a timely topic. A natural experiment could be conducted to evaluate the impact of a new policy such as Medicaid expansion. Accurate and precise effect estimation of utilization and cost attributable to a healthcare policy is critically important. Policymakers require the most correct information available to make the most informed decisions regarding the United States’ healthcare system. There has never been a more crucial time for accurate and precise effect estimation of a healthcare policy available to applied researchers.
2. **BACKGROUND**

2.1 **Natural experiments in healthcare policy estimation**

Policymakers and program administrators are interested in the effect of healthcare policies and public health programs in order to guide future implementations. Observational studies are often used to evaluate the impact of a healthcare policy implementation (Dimick and Ryan, 2014). However, observational studies can be plagued with many sources of potential bias due to the nonexperimental essence of the data. Due to this reason, there exists “almost a complete lack of an evidence base on the cost effectiveness of public health interventions” (Wanless, 2004). Natural experiments could offer the best available evidence in the evaluation of a public health intervention (Petticrew et al., 2005).

2.2 **Difference-in-differences**

Difference-in-differences estimation strategy is frequently used in applied economics to evaluate the effect of a public policy change or intervention from observational data. Oftentimes the difference in post and pre outcomes of a policy change is attributed to the effect of that policy. However, this method does not take into account any time-dependent changes that might have occurred unrelated to the policy change. The DID study design accounts for such background changes by using a comparison group experiencing the same time-dependent changes but not experiencing the policy (Angrist and Pischke, 2008). Consider two groups, treatment and control, across two time periods, pre and post policy implementation. The intuition is straightforward: compute the difference in the treated group’s post and pre outcomes, and separately compute the difference in the control group’s post and pre outcomes. Then take the difference of the two calculated differences. Simply, the DID estimate is,

\[
\text{DID} = (\bar{y}_{treatment, post} - \bar{y}_{treatment, pre}) - (\bar{y}_{control, post} - \bar{y}_{control, pre})
\]
(Wooldridge, 2010). Alternatively, the first difference could be taken within post and pre time periods between treatment and control groups

\[
\text{DID} = (\bar{y}_{\text{treatment,post}} - \bar{y}_{\text{treatment,pre}}) - (\bar{y}_{\text{control,post}} - \bar{y}_{\text{control,pre}}) \\
= (\bar{y}_{\text{treatment,post}} - \bar{y}_{\text{control,post}}) - (\bar{y}_{\text{treatment,pre}} - \bar{y}_{\text{control,pre}}).
\]

The DID estimate can be thought of as completing a two by two table with cells containing the average outcomes of treated and control, pre and post data. The marginal differences are then taken. The difference in these marginal differences is the DID. A representation of this can be seen in Table I. The DID estimate can also be realized visually. The difference in the observed outcome of the treated to what the outcome would have been had the trend remained parallel to the control group is the DID, or simply the policy effect. Figure 1 displays this visual representation of the DID effect.

**TABLE I**

THE DIFFERENCE-IN-DIFFERENCES EFFECT ESTIMATE

<table>
<thead>
<tr>
<th></th>
<th>Pre</th>
<th>Post</th>
<th>Difference</th>
</tr>
</thead>
<tbody>
<tr>
<td>Treated</td>
<td>$\bar{y}_{\text{treatment,pre}}$</td>
<td>$\bar{y}_{\text{treatment,post}}$</td>
<td>$\bar{y}<em>{\text{treatment,post}} - \bar{y}</em>{\text{treatment,pre}}$</td>
</tr>
<tr>
<td>Control</td>
<td>$\bar{y}_{\text{control,pre}}$</td>
<td>$\bar{y}_{\text{control,post}}$</td>
<td>$\bar{y}<em>{\text{control,post}} - \bar{y}</em>{\text{control,pre}}$</td>
</tr>
</tbody>
</table>

\[
\text{DID} = (\bar{y}_{\text{treatment,post}} - \bar{y}_{\text{treatment,pre}}) \\
- \bar{y}_{\text{control,pre}} - \bar{y}_{\text{control,post}} \\
= (\bar{y}_{\text{treatment,post}} - \bar{y}_{\text{control,post}}) \\
- (\bar{y}_{\text{treatment,pre}} - \bar{y}_{\text{control,pre}})
\]
Figure 1. The difference-in-differences treatment effect
Landmark papers in DID estimation involve natural experiments in employment, including the effect of employment on a minimum wage increase, immigration on employment, and wages and disability benefits on time out of work after injury (Card and Krueger, 1993; Card, 1990; Meyer et al., 1995). Health services’ applications of DID include change in hospital volume, cost, and quality of care among different agent payment levels (Farrar et al., 2009), racial disparities in health services’ use following mandatory enrollment in managed care programs among Medicaid beneficiaries (Tai-Seale et al., 2001), the impact of limiting Medicare’s bariatric surgery coverage to only Centers of Excellence (Dimick et al., 2013), adherence to diabetes medication following a new health benefit design (Zeng et al., 2010), introduction of a new healthcare policy and delivery system, the Alternate Quality Contract onto medical spending, quality, and performance (Song et al., 2011; Song et al., 2012), out-of-pocket mental healthcare expenditures (Stuart et al., 2014), and the effect of VA medication acquisition of Veterans enrolled in Medicare Part D (Stroupe et al., 2017). To more clearly convey the objectives of this study, we assume that all assumptions of the DID estimator have been fully satisfied. For details and proofs of these assumptions the reader is referred to excellent sources by Abadie (2005) and Lechner (2011).

Let $Y(i, p)$ be the outcome for subject $i$ at time $p$, where $p$ equal to 0 indicates the pre treatment period and equal to 1 indicates the post treatment period. Henceforth, we will use small $p$ when referring to actual values of the time indicator and big $P$ when referring to the random variable, such as in regression equations. Define $Z$ as the binary treatment indicator of what is actually observed, equal to 1 for those receiving treatment and equal to 0 for those receiving control. Assuming the selection on treatment assignment does not depend on any transitory shock, and equality of variances in average outcomes of both the treated and control groups had
the treated group not been treated (Abadie, 2005), the DID can be estimated by the least squares estimate of,

\[
\text{DID} = \{E[Y(i, 1) | Z = 1] - E[Y(i, 0) | Z = 1]\} - \{E[Y(i, 1) | Z = 0] - E[Y(i, 0) | Z = 0]\}.
\]

The DID is estimable via a two-way fixed effects model interacting period and treatment group indicators,

\[
Y = \beta_0 + \beta_P P + \beta_Z Z + \beta_{PZ} PZ + \epsilon.
\]

This is the overwhelmingly most popular estimation strategy of DID effect estimation. For a continuous outcome \(Y\) where the link function is equal to the mean, the DID effect can be estimated by a two-way fixed effects model and represented exactly by the parameter estimate of the interaction term, \(\hat{\beta}_{PZ}\). Table II shows the DID treatment effect as obtained from a regression relationship.

**TABLE II**

DIFFERENCE-IN-DIFFERENCES AS REPRESENTED BY A TWO-WAY FIXED EFFECTS REGRESSION MODEL

<table>
<thead>
<tr>
<th></th>
<th>Pre</th>
<th>Post</th>
<th>Difference</th>
</tr>
</thead>
<tbody>
<tr>
<td>Treatment</td>
<td>(\beta_0 + \beta_Z)</td>
<td>(\beta_0 + \beta_P + \beta_Z + \beta_{PZ})</td>
<td>((\beta_0 + \beta_P + \beta_Z + \beta_{PZ}) - (\beta_0 + \beta_Z) = \beta_P + \beta_{PZ})</td>
</tr>
<tr>
<td>Control</td>
<td>(\beta_0)</td>
<td>(\beta_0 + \beta_P)</td>
<td>((\beta_0 + \beta_P) - (\beta_0) = \beta_P)</td>
</tr>
<tr>
<td>Difference</td>
<td>DID = ((\beta_P + \beta_{PZ}) - (\beta_P) = \beta_{PZ})</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
2.2.1 **Structure of the data considered**

We examine the particular case where individuals are not necessarily recorded both pre and post policy implementation. This DID scenarios is contrary to the case one would typically consider of longitudinal recordings on the same individuals. In the case considered here, pre and post groups within the same treatment regime are made up of possibly different subjects. Subjects are allowed to be in the treatment group in the pre period and the control group in the post period, and vice versa or exit and enter the study altogether. All subjects have at least one, and at most two recorded outcomes. If two outcomes are recorded on the same subject, one outcome is recorded pre policy implementation and the other is recorded post policy implementation. One benefit in examining this scenario is that it is less restrictive on the data. Methods derived in this context can be applied to more general types of data structures. Such methods have a broader range of applicability in comparison to the scenario of repeated measures on the same individual. For example, an individual could switch from control to treated groups when, after realizing the current state they reside does not offer Medicaid coverage at their particular income level, decides to relocate to a state that does. This scenario might particularly be the case for individuals living close to state lines or managing a complex, chronic health condition. The structure of data examined here of non-same-person-recorded outcomes represents a more general structure of data to that typically examined in DID effect estimation, as follow-up of the same subjects can be costly and unfeasible at times. Non-same-person-recorded data in DID represents a likely scenario of available observational data in which accurate and precise estimation methods are required.

In the case investigated here of non-same-subject-recorded outcomes, pre and post outcomes are unlikely to be correlated. Had subjects been recorded both pre and post
intervention, the correlation of outcomes within subjects would need to be taken into account. Failure to account for correlated responses within the same unit would result in downwardly biased standard error estimates, resulting in a misleadingly narrow confidence interval (CI) (Cameron and Miller, 2015; Bertrand et al., 2004; Moulton, 1986; Moulton, 1990).

The structure of the data examined in this dissertation is of cross sectional recordings of pre and post time periods as opposed to longitudinal recordings. Our motivation is to identify an improved estimation method of the standard error in DID effect estimation.

2.3 **Counterfactual framework**

We continue to consider the situation of two groups measured across two time periods. Denote $Y_Z(i,p)$ as the potential outcome response of subject $i$ had subject $i$ received treatment $Z$ (=1 if treated, =0 if control) in period $p$ (=0 in pre period, =1 in post period). Each subject observes the outcome $Y$ and vector of covariates $X$ measured prior to the receipt of treatment, or if measured after, unaffected by the group assignment. For ease of notation, exclude the subject argument $i$, and let $Y_Z(p)$ denote the potential response under treatment $Z$ in period $p$. Each individual has an associated vector $(Y_0(p), Y_1(p))$ of potential, or counterfactual responses had the individual received control or treatment. The stable unit treatment value assumption (SUTVA) (Angrist et al., 1996) states that potential outcomes $(Y_0(p), Y_1(p))$ are unrelated to the treatment status of other individuals and thus completely determined. When SUTVA holds, the observed outcome will equate with the potential outcome under the treatment received

$$Y(p) = zY_1(p) + (1 - z)Y_0(p).$$

(SUTVA)

Positivity states that each person has a probability greater than zero of being in both the treatment and control groups. Exogeneity assumes covariates in $X$ are not changed by the treatment received, $X_1 = X_0 = X$. Strongly ignorable treatment assignment (SITA), or
exchangeability, asserts there is no unmeasured confounding in expectation, $(Y_0(p), Y_1(p)) \perp X|\nu$ where $\nu$ is a subset of $x$ and $\perp$ denotes statistical independence. We assume SUTVA, positivity, exogeneity, and SITA are fully satisfied.

Of course $Y_0(p)$ and $Y_1(p)$ in the same period $p$ are never concurrently observed for any subject but rather represent the hypothetical responses had the subject received both treatment and control during the same period, $p$. The means of the hypothetical distributions $Y_0(p)$ and $Y_1(p)$ represent the responses had every individual received control or treatment, respectively. The means of these distributions, $E[Y_0(p)]$ and $E[Y_1(p)]$ represent the mean responses had all individuals received both treatments at time $p$. The central necessity in causal effect estimation relies on the identification of expected potential outcomes from the observed data $(Y, Z, X)$. The average treatment effect on the treated (ATET) is the treatment effect on the group of individuals who received treatment. Estimation of the ATET is the main interest in this dissertation and can be formally represented as,

$$ATET(x) = E[Y_1(1) - Y_0(1) | X = x, Z = 1].$$

Important assumptions specific to DID are common trend (CT) and no effect on the pre-treatment population (NEPT). The CT assumption states that changes over time in expected potential control outcomes given $X$ are independent of treatment assignment in the post period. More specifically, trends of treated and control groups given $X$ would have been the same regardless of if treatment was ever applied. The CT assumption equates these trends in treatment and control groups conditional on $X$,

$$E[Y_0(1) | X = x, Z = 1] - E[Y_0(0) | X = x, Z = 1]$$

$$= E[Y_0(1) | X = x, Z = 0] - E[Y_0(0) | X = x, Z = 0]$$

$$= E[Y_0(1) | X = x] - E[Y_0(0) | X = x].$$
The NEPT assumption states that there was no effect of treatment in the pre treatment period,

\[ \text{ATET}(0, x) = 0 \quad \forall x \in X \]  

(NEPT).

The counterfactual outcome of the treated group post policy implementation is what the trend would have been had it followed a parallel path to the observed trend of the control group (CT assumption). The difference in the counterfactual outcome and observed outcome of the treated group post policy implementation is the DID. An illustration of the counterfactual outcome in the context of DID estimation is displayed in Figure 2.

The ATET can be estimated using observed data provided the aforementioned assumptions have been satisfied (Lechner, 2011):

\[ \text{ATET}(x) = E[Y_1(1) - Y_0(1) | X = x, Z = 1] \]

\[ = E[Y(1) | X = x, Z = 1] - E[Y_0(1) | X = x, Z = 1] \]  

(SUTVA)

Noting that,

\[ E[Y_0(1) | X = x, Z = 1] = E[Y_0(1) | X = x, Z = 0] - E[Y_0(0) | X = x, Z = 0] + E[Y_0(0) | X = x, Z = 1] \]  

(CT)

\[ = E[Y(1) | X = x, Z = 0] - E[Y(0) | X = x, Z = 0] + E[Y_0(0) | X = x, Z = 1] \]  

(SUTVA)

The last term reduces to,

\[ E[Y_0(0) | X = x, Z = 1] = E[Y_1(0) | X = x, Z = 1] \]  

(NEPT)

\[ = E[Y(0) | X = x, Z = 1] \]  

(SUTVA)

And finally,

\[ \text{ATET}(x) = \{E[Y(1) | X = x, Z = 1] - E[Y(0) | X = x, Z = 1]\} - \{E[Y(1) | X = x, Z = 0] - E[Y(0) | X = x, Z = 0]\} \]

All components of ATET(x) are observable, and thus ATET(x) is identifiable (Lechner, 2011).
Figure 2. Difference-in-differences treatment effect and counterfactual outcome
2.4 Propensity score

A popular method to minimize selection bias in observational studies are propensity score estimation techniques (Rosenbaum and Rubin, 1983). Failure to account for treatment selection bias can lead to biased estimates. The propensity score, is the conditional probability of treatment assignment on observed pre treatment characteristics $X$, or if recorded post treatment, unaffected by treatment, $e(X) = P(Z = 1|X)$. It is bounded by 0 and 1, $0 < e(X) < 1$. Treatment assignment is independent of potential outcomes given the propensity score.

The most common way to estimate propensity scores, and the method employed throughout this paper is via logistic regression. Alternate estimation methods of the propensity score do exist and include probit regression, neural networks, recursive partitioning, and boosting (Westreich and Cole, 2010; Lee et al., 2011; McCaffrey et al., 2004; Setoguchi et al., 2008). Throughout this dissertation, we estimate propensity scores via the logistic regression model parameterized by $\alpha_m$, $e(X, \alpha) = \{1 + \exp(-X'\alpha)\}^{-1}$, where $m$ is the number of recorded characteristics. The maximum likelihood estimator of $\alpha$, $\hat{\alpha}$ can be obtained by solving the estimating equations,

$$\sum_{i=1}^{n} \Psi_j(Z_i, X_i; \alpha) = \sum_{i=1}^{n} \frac{Z_i - e(X_i, \alpha)}{e(X_i, \alpha)\{1 - e(X_i, \alpha)\}} \frac{\partial}{\partial \alpha} \{e(X_i, \alpha)\}$$

$$= \left[ Z_i - \frac{\exp(X_i\alpha)}{1 + \exp(X_i\alpha)} \right] (X_i)_j$$

$$= [Z_i - e(X_i, \alpha)](X_i)_j$$

by setting each $j= 1, \ldots, m + 1$ equations equal to zero and solving for $\alpha_j$ (McCullagh and Nelder, 1989). To simplify notation, we write the propensity score $e(X_i, \alpha)$ as $e(X)$, and $e(X_i)$ when evaluated at $i$. 
2.5 Inverse probability weighting

Inverse probability weighting methods were introduced in survey sampling methodology (Horvitz and Thompson, 1952). A new pseudo-population can be imagined as derived from \( w \) copies of each subjects’ data. Copies of the subjects’ data, \( w \), can be thought of as weights which are greater for treated observations who look dissimilar to other treated observations, and less for treated observations who look similar to other treated observations. Similarly, more weight is given to control observations who look dissimilar to other control observations, and less weight is given to control observations who look similar to other control observations. The purpose of this weighting strategy is to balance treated and control groups by putting all observations on a level playing field. Given (1) and \( Z(1 - Z) = 0 \), it follows that \( E[ZY/e(X)] = E[ZY_1/e(X)] \). By assuming SUTVA and SITA,

\[
E \left[ \frac{Z Y}{e(X)} \right] = E \left\{ E \left[ \frac{Z Y_1}{e(X)} | Y_1, X \right] \right\} = E \left\{ \frac{Y_1}{e(X)} E[Z | Y_1, X] \right\} = E(Y_1)
\]

(Lunceford and Davidian, 2004). Similarly, \( E \left[ \frac{(1-Z) Y}{1-e(X)} \right] = E(Y_0) \). Intuitively, each unit is weighted by the inverse propensity of it being in the particular treatment regime observed, then averaged across all units in the same treatment group. Substituting the estimated propensity scores from the logistic regression model into IPW is equivalent to solving the estimating equations

\[
\sum_{i=1}^{n} u(\theta; Y_i, Z_i, x_i) = 0 \quad \text{for the parameter } \theta = (\mu, \mu_0, \alpha)’, \text{ where }
\]

\[
u(\theta; Y, Z, x) = \begin{pmatrix} (Y - \mu_1) Ze^{-1} \\ (Y - \mu_0)(1 - Z)(1 - e)^{-1} \\ x(Z - e) \end{pmatrix}.
\]

The resulting estimator, \( \hat{\theta} \), is asymptotically normally distributed.

Lunceford and Davidian (2004) prove an interesting property of IPW estimators. For \( \alpha \) unknown and therefore estimated, the large-sample variance of the IPW estimator is actually
reduced in comparison to the situation where the true propensity scores had been known. The authors recommend even in situations where the propensity scores are known, estimation of the propensity scores is still beneficial from an efficiency standpoint. When using propensity scores to determine IPWs, the derived weights come with inexact precision. Propensity scores are rarely known using real data. In practice, it might be unknown how close the estimated weights are in comparison to their true value. Previous studies have shown that failing to take into account prior estimation of the propensity scores in matched propensity score estimators (Abadie and Imbens, 2016), stratified (Williamson et al., 2012), and IPW propensity score estimators (Lunceford and Davidian, 2004; Williamson et al., 2014) will result coverage intervals that are too conservative (i.e. too wide). Using robust standard errors might not be enough to account for the sampling variability of the propensity scores. Wald 95% confidence intervals from a weighted regression with robust standard errors might still be too conservative because they do not take into account the prior estimation of the propensity scores (Robins et al., 2000). Asymptotic variance estimation formulas have been derived for the IPW estimator in a single-difference treatment effect but not for a double DID treatment effect estimate.

Diagnostic testing and graphical approaches can be used to compare distributions across treatment groups. Balance of treatment groups can be evaluated via standardized differences. Further, side-by-side boxplots, empirical cumulative distribution functions, as well as Kolmogorov-Smirnov testing can be used to compare distributions between the two groups (Austin and Stuart, 2015). The distribution of weights should be checked as IPWs can be unstable when propensity scores are close to one (Kang and Schafer, 2007). To account for the estimation of weights, empirical methods such as robust/sandwich variance estimators (Joffé et al., 2004) or bootstrap methods (Austin and Stuart, 2015) can be applied.
2.6 **Augmented inverse probability weighted estimator**

The augmented inverse probability weighted (AIPW) estimator extends the IPW estimator by including additional predictive information from the covariate(s). Covariate information $X$ “contains information about the probability of treatment, but it also contains predictive information about the outcome variable” (Glynn and Quinn, 2009). In a single difference, AIPW estimators exhibit the “doubly robust” property, such that the estimator will perform well if either the propensity score or outcome regression model is correctly specified. This offers the researcher two chances at making the correct inference, a property not shared by IPW or likelihood-based estimators (Bang and Robins, 2005). Doubly robust methods could increase efficiency and reduce bias (Lunceford and Davidian, 2004). For a continuous outcome $Y$, an AIPW estimator tended to have a lower root-mean-square error (RMSE) compared to IPW estimators in a study by Glynn and Quinn (2009). Furthermore, the AIPW estimator is dramatically superior under partial misspecification of the propensity score model (Glynn and Quinn, 2009). The AIPW estimator of the average treatment effect (ATE) in a single difference is,

$$
\text{ATE}_{\text{AIPW}} = \frac{1}{n} \sum_{i=1}^{n} \left[ \frac{Z_{i}Y_{i}}{\hat{e}(X_{i})} - \frac{(1-Z_{i})Y_{i}}{1-\hat{e}(X_{i})} \right] - \frac{(Z_{i}-\hat{e}(X_{i}))}{\hat{e}(X_{i})(1-\hat{e}(X_{i}))} \left[ \left( 1 - \hat{e}(X_{i}) \right) \hat{E}(Y_{i}|Z_{i} = 1, X_{i}) + \hat{e}(X_{i}) \hat{E}(Y_{i}|Z_{i} = 0, X_{i}) \right].
$$

Lunceford and Davidian (2004) derive an empirical sandwich estimator of the sampling variability of $\text{ATE}_{\text{AIPW}},$

$$
\text{Var}(\text{ATE}_{\text{AIPW}}) = \frac{1}{n^2} \sum_{i=1}^{n} \hat{f}_{i}^2
$$

where
\[ \hat{I}_i = \left[ \frac{Z_i Y_i}{\hat{e}(X_i)} - \frac{(1 - Z_i)Y_i}{1 - \hat{e}(X_i)} \right] \]

\[ - \frac{(Z_i - \hat{e}(X_i))}{\hat{e}(X_i)(1 - \hat{e}(X_i))} \left( (1 - \hat{e}(X_i))\hat{E}(Y_i|Z_i = 1, X_i) + \hat{e}(X_i)\hat{E}(Y_i|Z_i = 0, X_i) \right) \]

\[ - \text{ATE}_{AIPW}. \]

2.7 Previous approaches

Current approaches to address selection bias in DID are limited (Stuart et al., 2014). Covariate regression adjustment in DID typically does not sufficiently resolve selection bias (Abadie, 2005). Propensity score methods could offer potential in addressing the selection bias issue in DID. These methods generally fall into four categories: matching, stratification, covariate adjustment by the propensity score, and weighting. There have been some developments in propensity score adjustment using matching and stratification in the context of DID estimation. The weighted two-way fixed effects regression estimator was shown to be algebraically equivalent to the causal DID estimator. Statistical properties and estimator sensitivity have not been established for different weighting schemes and under various types and levels of selection bias in DID estimation. The following are some studies which incorporate propensity score approaches in DID estimation.

2.7.1 Propensity score matching

Werner et al. (2009) uses a propensity score matched design in DID estimation to examine quality of nursing home post-acute care following the rollout of Nursing Home Compare, a website from the Centers for Medicare and Medicaid Services which publishes key statistics on nursing home care, quality, and safety. The aim of this investigation was to determine the effect of public reporting of nursing home quality measures. All nursing homes were subject to the new reporting standards, except small nursing homes. The authors chose to
compare large nursing homes (treated group) to small nursing homes (control group). They use a
two-pronged approach to isolate the desired effect estimate. First, within-subject changes were
examined. Then a DID model was applied to a matched sample which matched similar patients
pre and post implementation. It was a unique application of propensity score matching in that
persons were matched within the same treatment group, pre and post time periods. More
specifically, persons were matched within the same unit (i.e. nursing home) with replacement
and this within-unit matching was accounted for in Huber-White robust standard error
calculations (Huber, 1967; White, 1980). This study was unique in a few ways. Typically, treated
individuals would have been matched to similar control individuals on attributes in the pre
implementation period, or if measured post implementation, unaffected treatment assignment.
Another aspect of the propensity score matching which made this study unique was the
composition of the control group. Clearly, small nursing homes are much different on a
multitude of different factors not limited to patient case mix and for-profit status. The Werner et
al. (2009) study offered an interesting approach in propensity score adjustment in DID
estimation. The criticism we have is the failure to address the imbalance between vastly different
treatment and control groups.

2.7.2 Propensity score weighting and sensitivity analysis

Song et al. (2012) use a DID approach to estimate the effect of the Alternative Quality
Contract of Blue Cross Blue Shield of Massachusetts on spending and quality. Treatment was
not randomly assigned. The treated group consisted of prior-risk organizations who had past
involvement with risk-based contracts. The control group consisted of organizations who had
previously operated fee-for-service. A multivariate linear regression model was used to estimate
the DID effect of spending changes. Propensity score weights were calculated using variables
age, sex, risk, and cost sharing. There are no details on how the propensity score weights were constructed, though we are assuming IPW. This calls into question whether prediction of prior-risk versus no-prior-risk organization is a function of patient’s mix, risk, and cost sharing. They do several sensitivity analyses to support their results, including a test for differences in pre intervention spending trends between prior-risk and no-prior-risk groups, and indicate that they found no self-selection bias. This is more or less a test of the constant slope assumption in DID model estimation in the pre period and then inferring it would hold on towards the post period. Additional details around the propensity score weights and accompanying diagnostics would have strengthened confidence in the claims. An important lesson gleaned from this investigation and that of Werner et al. (2009) is the importance of examining outcomes much after program implementation. Both studies found a large transitory shock in the year following program implementation which eventually stabilized. Results examining only two time points of pre and post implementation in DID estimation should be taken with caution as a substantial portion of the estimated effect could be attributed to a transitory shock of the program introduction. Other examples which use propensity score weighting in DID estimation are Zeng et al. (2010) and Song et al. (2011).

2.7.3 Four-group weighted design

Stuart et al. (2014) proposed an interesting approach using IPW in DID estimation. The authors simultaneously generate IPWs for each of the four possible scenarios (treatment pre, treatment post, control pre, control post) and use them in weighted regression models. This method does not assume longitudinal recordings of the same individual and is thus appropriate for DID estimation in the absence of within-subject correlation. An illustrative example of the approach is applied to estimate the effect of enrollee out-of-pocket expenses for mental
healthcare attributed to a new healthcare program, the Accountable Care Organization model implemented by Blue Cross Blue Shield of Massachusetts. They compare an unweighted estimator, an estimator weighted separately at each time point, and a new proposed four-group weighted estimator.

The author’s proposed four-group weighted approach begins with defining four groups from the two by two DID design: group 1 = treatment pre, group 2 = treatment post, group 3 = control pre, group 4 = control post. Group delineation of the four groups can be seen in Table III.

**TABLE III**

GROUP IDENTIFICATION IN THE FOUR-GROUP WEIGHTED DESIGN

<table>
<thead>
<tr>
<th></th>
<th>Pre</th>
<th>Post</th>
</tr>
</thead>
<tbody>
<tr>
<td>Treated</td>
<td>group 1</td>
<td>group 2</td>
</tr>
<tr>
<td>Control</td>
<td>group 3</td>
<td>group 4</td>
</tr>
</tbody>
</table>

Propensity scores are estimated via a multinomial logistic regression model. Group 1 was arbitrarily chosen as the reference group in order to derive weights similar to a common group, though any of the four groups can serve as the reference group. Four separate propensity scores are calculated for each subject as the probability of being in each one of the four groups. The four estimated propensity scores sum to one for each subject. Propensity scores are then used to construct weights proportional to the probability of being in group 1 relative to the group they
were actually in. Weights are calculated as the probability of being in group 1 versus the remaining groups (2, 3, and 4), separately. The weight for an individual \( i \) is,

\[
w_i = \frac{e_1(X_i)}{e_g(X_i)}
\]

where \( g \) represents the actual group membership of individual \( i \). Individuals in group 1 have a weight equal to one. The DID estimate of interest is decomposed as,

\[
\text{DID}_{G=1} = E_{X|G=1}E[Y_1(1) - Y_0(1) | X = x] \\
= E_{X|G=1}[Y_1|X = x,Z = 1] - E_{X|G=1}[Y_1(0) | X = x,Z = 1] \\
- (E_{X|G=1}[Y_0|X = x,Z = 0] - E_{X|G=1}[Y_0(0) | X = x,Z = 0]).
\]

Consistent estimates for each of the four groups are generated as,

\[
\hat{\mu}_g = \frac{\sum_{i=1}^{n} I(G_i = g)Y_iw_i}{\sum_{i=1}^{n} I(G_i = g)w_i},
\]

These estimates can be easily calculated from a fitted regression model such as,

\[
Y = \beta_0 + \beta_P + \beta_Z + \beta_{PZ}PZ
\]

using a long dataset with observations of subjects at time period \( p \).

The authors apply the proposed estimator to four DID scenarios: no difference, group difference, group and time difference, and group by time difference. The four-group weighted estimator is evaluated and compared to the naïve two-way fixed effects regression estimator and an estimator that weights groups separately at each time period. Metrics in bias and coverage are used to compare estimators. As expected, the naïve DID model performs poorly, with higher bias and lower confidence interval coverage under scenarios of selection bias. Stuart et al.’s (2014) proposed four-group weighted approach yielded the lowest bias estimates and good coverage. While authors promote the four-group weighted design, they caution the researcher that along with the negligible bias comes the possibility of increased standard errors.
2.8 Forward thoughts

The central interest of this dissertation is to determine an improved estimation strategy of the standard error in DID estimation which addresses group and time selection bias by IPW using the propensity score. Prior studies have shown that incorporating propensity score weighting methods into DID estimation in a natural experiment improves the statistical properties of the estimate by balancing covariates in the nonrandomly allocated treatment groups. To date, the recommended estimator is Stuart et al.’s (2014) four-group weighted estimator using a model-based standard error. However, this estimator has some shortcomings. While the estimator exhibits excellent bias properties, it comes at the cost of decreased precision. We hypothesize that failure to take into account the prior estimation of the propensity score in IPW models will result in confidence intervals that are too conservative. To account for this, we propose the use of an empirical standard error to improve precision while retaining the progress made in bias. This is the main aim of this dissertation. Bootstrapping procedures are seldom applied together with propensity score methods (Austin and Small, 2014). Secondary aims include examining the change in precision by including covariates to the outcome model used in the propensity score model. We hypothesize that a DID estimator which includes covariates in the outcome model will be more precise to a similar estimator which does not. Finally, we propose an AIPW estimator. We anticipate such an estimator will be the most robust under either propensity score or outcome model misspecification. However, such an augmented estimator might not perform as well as four-group weighted estimators in certain types of selection bias. Robustness is not tested for in this investigation. A further study could be conducted to confirm the aforementioned robustness hypotheses.

The degree to which naïve, unweighted DID estimators, the common practice in DID
estimation, do not reflect the true estimate and associated coverage could lead the researcher and ultimately policymaker into making ill-informed policy decisions. There is a critical need for accurate and precise effect estimation of current healthcare policy implementations.
3. METHODS

The goal of this dissertation is to identify an improved estimation method of the standard error resulting in increased precision of the DID treatment effect using IPW by the propensity score. The methods proposed in this dissertation are, (1) a variance estimation strategy in the separate time weighted estimator and Stuart et al.’s (2014) four-group weighted estimator, and (2) an AIPW DID estimator.

3.1 Two-way fixed effects regression estimator

The overwhelmingly most common method in DID estimation is the two-way fixed effects regression estimator. The DID effect via a two-way fixed effects model can be specified as,

\[ Y = \beta_0 + \beta_P P + \beta_Z Z + \beta_{PZ} PZ. \]

The DID effect can be represented exactly by the parameter estimate of the interaction term, \( \hat{\beta}_{PZ} \).

The model-based standard error of the interaction term is taken as the standard error estimate of the DID.

Two versions of each regression model specification outlined in this dissertation will be examined by, (1) the method outlined and (2) the method outlined including the covariate(s) already used in the propensity score model, to the outcome model. Controlling for covariates when estimating the treatment effect is a way to improve precision. Subjects’ covariates are likely correlated to their outcomes and potential outcomes. Including these covariates to the outcome model can aid in setting aside possible variation in the outcome attributed from the covariates. The treatment effect on outcomes can then be estimated with greater power and precision. This gain in precision is especially highlighted when covariates are highly predictive of the outcome. For example, the two-way fixed effects model including covariate(s) \( X \) is
specified as,

\[ Y = \beta_0 + \beta_P P + \beta_Z Z + \beta_{PZ} PZ + \beta_X X. \]

### 3.2 Separate time weighted regression estimator

Consider a model that weights each subject depending on their treatment group, separately at each time period. More specifically, split the data into pre intervention and post intervention groups. Estimate propensity scores via a logistic regression model parameterized by \( \alpha_{m \times 1} \),

\[ P(Z = 1|X) = e(X, \alpha) = \{1 + \exp(-X'\alpha)\}^{-1} \]

in both the pre period and post period, separately.

- **Pre data:** \( P(Z = 1|X, \text{groups 1 and 3}) = e_{pre}(X, \alpha) = \{1 + \exp(-X'\alpha)\}^{-1} \)
- **Post data:** \( P(Z = 1|X, \text{groups 2 and 4}) = e_{post}(X, \alpha) = \{1 + \exp(-X'\alpha)\}^{-1} \)

Generate individual predicted probabilities of treatment assignment, \( \hat{e} \),

- Subjects in group 1: \( \hat{e}_{pre} \) where group = 1
- Subjects in group 2: \( \hat{e}_{post} \) where group = 2
- Subjects in group 3: \( \hat{e}_{pre} \) where group = 3
- Subjects in group 4: \( \hat{e}_{post} \) where group = 4

Append together all individual predicted probabilities from the four groups. Subjects who received treatment (groups 1 and 2) are weighted by \( \frac{1}{e} \) and subjects who did not receive treatment (groups 3 and 4) are weighted by \( \frac{1}{1-e} \). Obtain the DID estimate and associated standard error via a weighted two-way fixed effects regression model using the generated weights,

\[ Y = \beta_0 + \beta_P P + \beta_Z Z + \beta_{PZ} PZ. \]
3.3 Four-group weighted design

Stuart et al.’s (2014) four-group weighted design detailed in the previous section will be examined. While this method delivers excellent bias properties, the compromise comes at perhaps an inflated standard error. A goal is to offer an improved estimation strategy of the standard error in the four-group weighted estimator. Bootstrap resampling methods will be used to account for the sampling variability of the propensity scores. The median bootstrap estimate will be explored as a candidate point estimate of the DID, though the main focus of these strategies will be on the precision as opposed to the accuracy.

3.4 Empirical standard error estimates

To account for the sampling variability in the estimated propensity scores, standard errors are computed via two bootstrap methods: (1) basic (a.k.a., non-studentized pivotal) method and (2) percentile method.

3.4.1 Basic bootstrap method

Denote the true population parameter as \( \theta \), and the statistic of interest as \( W \), where \( W = g(\hat{\theta}; \theta) \) for a continuous function \( g \). In our case, that statistic of interest, \( W \), is the DID. If the distribution of \( W \) was known then \( w_\alpha \), where \( \text{Prob}(W \leq w_\alpha) = \alpha \), could be computed such that \( (-\infty, \hat{\theta} - w_\alpha) \) provides a one-sided 100(1 − \( \alpha \))% confidence interval for \( \theta \). Let \( \hat{\theta} \) be an estimator for \( \theta \). The intuition behind the basic bootstrap confidence interval comes from pretending \( \hat{\theta} \) is the true parameter value and considering the relationship of \( \hat{\theta} \) and estimates of \( \hat{\theta} \), \( \hat{\theta}^* \) (Carpenter and Bithell, 2000). The algorithm for computing a two-sided 95% basic bootstrap confidence interval is as follows,

1. Generate 999 bootstrap samples of \( \theta \)
2. Calculate \( w_i^* = \hat{\theta}_i^* - \hat{\theta} \), for \( i \in 1, ..., 999 \)
3. The 0.025\textsuperscript{th} quantile of $W^*$ is estimated by $w_j^*$ where $j = 0.025 \times (999 + 1) = 25$.

Likewise, the 0.0975\textsuperscript{th} quantile of $W^*$ is estimated by $w_j^*$ where $j = 975$.

4. The two-sided 95\% confidence interval is $(2 \times \hat{\theta} - w_{975}^*, 2 \times \hat{\theta} - w_{25}^*)$

3.4.2 **Percentile bootstrap method**

The percentile method is a popular bootstrap method due to its simplicity. The algorithm for the percentile bootstrap method is as follows,

1. Generate 999 bootstrap samples of $\theta$
2. Sort the bootstrap estimates in ascending order, $\hat{\theta}_1^*, ... \hat{\theta}_{999}^*$
3. The two-sided 95\% percentile confidence interval is, $(\hat{\theta}_{25}^*, \hat{\theta}_{975}^*)$

An advantage of the percentile method over the basic method is that it is transformation respecting (Carpenter and Bithell, 2000). Percentile bootstrap confidence intervals are good in general but a disadvantage can be substantial coverage error when the distribution of $\hat{\theta}$ is not symmetric (Efron and Tibshirani, 1994).

Another promising bootstrap procedure is the bias-corrected and accelerated confidence interval (Efron and Tibshirani, 1994). Bias-corrected and accelerated constructed confidence interval have been suggested as a promising technique in the variance adjustment of a treatment effect estimate obtained using propensity scores (Austin and Small, 2014; Williamson et al., 2012). This bootstrap approach was not examined here due to its computational intensity (Carpenter and Bithell, 2000). It uses a jackknife-type, leave-one-out procedure to obtain the acceleration parameter. Future investigations could examine whether this procedure is advisable in DID estimation using propensity scores.

There are possibly many ways one can perform the resampling procedures for the bootstrap in separate time weighted and four-group weighted estimators. We have tried several
such procedures. The method we use for the entirety of this investigation is by first generating all the desired weights, then sampling with replacement subjects’ entire data including its associated weight. More explicitly, each subjects’ data and associated weight are computed using all the data, such that each subject’s information can be represented by \((Y_i, P_i, Z_i, X_i, w_i)\). This set of subjects’ information is the unit of resampling. Subject weights are only calculated once using the full dataset. The median of the bootstrap estimates, \(\hat{\theta}^{*}_{500}\), is taken as the DID effect estimate. Another natural approach might be to sample the data and recalculate the weights each time for the bootstrap sample to use in the regression model. However, this approach did not work well.

3.5 **Augmented inverse probability weighted difference-in-differences estimator**

We extend the AIPW estimator to accommodate a second difference to estimate the DID. First, estimate the ATE using the AIPW estimator for each pre treatment and post treatment difference separately, \(\hat{ATE}_{AIPW,pre}\) and \(\hat{ATE}_{AIPW,post}\). To do this, the data is first stratified into distinct subsets of pre and post data. Propensity scores are generated using the pre and post data separately to compute separate first-difference effects, \(\hat{ATE}_{AIPW,pre}\) and \(\hat{ATE}_{AIPW,post}\). The difference between the two estimated ATEs, \(\hat{ATE}_{AIPW,post} - \hat{ATE}_{AIPW,pre}\) is taken as the final AIPW DID point estimate. Three versions of an empirical standard error are calculated for the proposed AIPW DID estimator: sandwich, basic bootstrap, and percentile bootstrap.

The sandwich standard error of the bootstrap is obtained by first retaining the sandwich standard error of both \(\hat{ATE}_{AIPW,pre}\) and \(\hat{ATE}_{AIPW,post}\). Since pre and post differences are assumed to be independent of one other, the final sandwich standard error of the difference is calculated in a straightforward manner,

\[
\hat{\sigma} = \sqrt{\hat{\sigma}^2_{AIPW,pre} + \hat{\sigma}^2_{AIPW,post}}.
\]
Bootstrap standard errors for the AIPW DID estimator are obtained by first sampling observations with replacement from pre and post sets of the data, the same size as the pre and post samples. Once the data have been sampled, $\hat{ATE}_{AIPW, pre}$ and $\hat{ATE}_{AIPW, post}$ are calculated on the two sets. The difference between the two estimated $\hat{ATE}$s, $\hat{ATE}_{AIPW, post} - \hat{ATE}_{AIPW, pre}$ is retained as the bootstrap parameter estimate. This sampling, calculating, and retaining procedure is repeated for the desired number of bootstrap replications. Basic and percentile bootstrap confidence intervals are obtained using the list of bootstrap estimates.

### 3.6 Estimators for comparison

We compare 13 estimator specifications to estimate the DID effect of an outcome $Y$:

1. Naïve two-way fixed effects model: $Y = \beta_0 + \beta_P P + \beta_Z Z + \beta_{PZ} PZ$, where the estimate of $\beta_{PZ}$ is taken as the DID estimate

2. Naïve two-way fixed effects regression model including the covariate(s) $X$: $Y = \beta_0 + \beta_P P + \beta_Z Z + \beta_{PZ} PZ + \beta_X X$, where the estimate of $\beta_{PZ}$ is taken as the DID estimate

3. Separate time weighted estimator

4a. Separate time weighted estimator including the covariate(s) $X$

4b. Basic bootstrap

4c. Percentile bootstrap

5. Four-group weighted estimator

6a. Four-group weighted estimator including the covariate(s) $X$

6b. Basic bootstrap

6c. Percentile bootstrap

7a. AIPW estimator with sandwich standard error

7b. Basic bootstrap
7c. Percentile bootstrap

3.7 Estimate of interest

The DID in linear models can exactly be represented as the coefficient of the interaction term and its associated standard error. However, in nonlinear models, DID effect estimation and its associated variability is not as straightforward as directly pulling a model coefficient. The interaction term in nonlinear models, such as Poisson and Gamma models, is still the primary coefficient of interest in DID estimation (Puhani, 2012). It is the marginal effect we wish to obtain as the DID, or how the conditional mean of the outcome changes with a change to the \( PZ \) interaction term. Special considerations of the cross difference, or interaction term, in nonlinear models (Ai and Norton, 2003) need not be taken into account in the case of DID estimation because the treatment effect in nonlinear DID models is not the cross difference, but rather the difference in cross differences (Puhani, 2012). Applying the delta method to the nonlinear transformation of the interaction term’s coefficient is an appropriate method for obtaining the DID standard error (Puhani, 2012). The marginal effect can be obtained by either calculus or a finite-difference approach, where calculus methods are typically used in the case of a continuous regressor, and a finite-difference approach used in the case of a dichotomous regressor (Cameron and Trivedi, 2005). For regression models with an exponential conditional mean function, such as Poisson and Gamma, \( E[Y|X] = \exp(X\beta) \), the finite-difference marginal effect of the interaction term \( PZ \) in the treated observations is \( \Delta E[Y|X]/\Delta PZ = \exp(X\beta) - \exp(X\beta - \beta_{PZ}) \), and in the control observations is \( \Delta E[Y|X]/\Delta PZ = \exp(X\beta + \beta_{PZ}) - \exp(X\beta) \). The finite differences are calculated for all observations and averaged across the entire sample for the overall finite-difference marginal effect of the interaction term, or the DID. Functions \textit{margins} in
Stata, and \textit{marg} of package modmarg in R (Gold et al., 2017) compute this exactly along with its associated delta method derived standard error.

3.8 \textbf{Discussion}

Including covariates to the outcome model in estimators are hypothesized to improve the statistical precision of the DID estimate. The improvement to precision is expected to hold when comparing all versions of models with and without covariate adjustment (naïve, separate time weighted, and four-group weighted). Covariate adjustment leads to a gain in statistical power. When covariates are useful in predicting the outcome, the model is better able to identify the effect of the treatment on the outcome. This leads to a smaller standard error, and more precise estimate of the treatment effect. On the other hand, if the covariate is not predictive of the outcome, there are no gains to the precision and even acquires a penalty in a degree of freedom. Even worse is when covariates are predictive of the treatment. This can lead to bias. All estimators except for the AIPW have covariate-adjusted and unadjusted versions so we are able to make inferences regarding covariate adjustment to changes in the statistical precision. The AIPW estimator already incorporates covariate information to the outcome model and is not compared to any version of the AIPW estimator which does not include covariate information.

In a study estimating the impact of a healthcare policy, it is natural to be interested in the effect of the policy on a population who is affected or eligible for the policy, rather than the effect of the policy on an individual. The marginal treatment effect is the effect of the treatment on an entire population. A randomized controlled trial allows for the estimation of the marginal treatment effect. If the goal of an observational study is to answer the same question as a randomized controlled trial, estimation of the marginal treatment effect might be more advantageous (Austin, 2011). Our intent is to estimate the effect of a policy or program on a
population as in a randomized controlled trial where the policy is administered to one group and not the other. Applied to the healthcare setting, a specific aim might be to estimate the effect of a new healthcare policy in terms of utilization or cost on a population, as opposed to an individual. Given the observational nature of most healthcare data and desired population treatment effect, it is the marginal treatment effect we wish to estimate.

In the past, researchers have used covariate regression adjustment to justify differences between treated and control groups. Regression adjustment allows for the estimation of the conditional treatment effect, or the average effect of a treatment on an individual. Propensity score techniques allow for the estimation of the marginal treatment effect, or the average effect of a treatment on a population (Austin, 2011). For example, the interpretation of a model coefficient in a multivariate linear regression model is the estimated change to the outcome when the regressor is increased by one, holding all other variables constant. More concretely, for a dichotomous regressor, the interpretation can be imagined as the difference in outcomes of two individuals, exactly the same on all recorded attributes, except the dichotomous covariate of interest. This is the conditional treatment effect. The conditional treatment effect can be dependent on model specification and covariate information, which is avoided when using propensity score methods (Martens et al., 2007). In an observational study involving a continuous outcome, no unmeasured confounding, and knowledge of the true outcome model, the marginal treatment effect will equal the conditional treatment effect (Austin, 2011). Therefore, in a hypothetical investigation using observational data and correctly specified propensity score and outcome models with continuous outcomes, covariate regression adjustment should deliver estimates similar to those using propensity score methods. Propensity
score techniques are more appropriate compared to covariate adjustment for obtaining the desired, marginal effect.

Naïve covariate regression adjustment and propensity score methods deliver different interpretations of the treatment effect, conditional versus marginal. While it is true we are interested in estimating the marginal treatment effect, basic covariate adjustment represents a widely used method in DID effect estimation of a policy or program. Naïve models which regression adjust for covariates are expected to perform well in the case of a continuous outcome, no group or time bias, and both the propensity score model and outcome model are correctly specified. Conditional estimates produced from naïve models will coincide to the desired marginal estimates under these conditions. Both being a widely used strategy as well as delivering the desired effect estimate under some plausible, represented scenarios is the rationale for including covariate adjustment as a competing estimator.

We expect empirical standard error estimates in separate time weighted and four-group weighted models to deliver coverage rates closer to the nominal rate compared to estimators which use model-derived standard errors. Model-based standard errors are expected to yield conservative (i.e. too wide) confidence intervals due to the unaccounted sampling variability of the propensity score. The resampling procedures in empirical standard errors account for the prior estimation of propensity scores and are expected to improve precision.

Separate time weighted and AIPW estimators are similar in that they both balance treatment groups at each time period. Procedures of the two estimators follow the same structure. First, the sample is stratified into pre and post sets. Separate estimates are computed on the stratified sets. Lastly, the estimates are combined and final inferences are made. While there exists similarities between the two estimators, the AIPW estimator utilizes information in the
conditioning set to improve upon a potential shortcoming in robustness of the separate time weighted estimator. In a Monte Carlo simulation study by Glynn and Quinn (2009), a single-differenced treatment effect was evaluated using AIPW, IPW, and naïve covariate-adjusted estimators. Authors report the AIPW estimator performed similar to an IPW estimator and naïve covariate adjustment under fully correct model specification. However, under partial misspecification of either the outcome or propensity score model, the AIPW estimator performed dramatically better compared to the IPW estimator and naïve covariate adjustment. A potential drawback of the AIPW estimator compared to the separate time weighted estimator is an increase in required computation power and programming for the analyst.

Covariate selection bias in DID can occur across group, across time, or both. It is important to use an estimator which addresses the type of selection bias in the underlying data. We anticipate the four-group weighted estimator will deliver superior overall balance to covariates compared to the separate time weighted estimator. The four-group weighted approach is the only method which simultaneously weights all treatment groups and time periods (treatment pre, treatment post, control pre, control post) in a single, parsimonious weighting procedure resulting in the four groups being similar on a set of covariates. We expect the distributions of propensity scores to have better overlap across treatment groups and time periods. Better covariate balance should translate to reduced bias by more accurately resembling a randomized controlled trial. The four-group weighted approach should yield reduced bias compared to the separate time weighted and AIPW estimators, however at potentially the cost of an inflated variance in a bias-variance tradeoff. Such a tradeoff might be accentuated when there exists extreme weights and therefore weights should be examined for outliers (Stuart et al., 2014).
It was suggested that we explore an additional estimator which first stratifies the sample into treated and control groups, then balances across pre and post time periods within each treatment group. This additional estimator was not evaluated as treatment groups were hypothesized to remain relatively balanced throughout time in the particular context of the study design. While individuals are allowed to move freely between policy-affected and policy-unaffected groups, movement between groups would typically present burden to the individual of either relocating their residence, or switching provider practice, for example. Treatment groups are hypothesized to remain nearly balanced across time due to such burden, however this needs to be carefully checked. If it is found the case mix of treatment groups change substantially over time, a separate group weighted estimator should be considered.

The aim of this investigation was to identify guidance in standard error estimation of the DID marginal treatment effect. Propensity score weighting is advantageous to covariate adjustment because it allows the estimation of the desired marginal treatment effect. Covariate adjustment in naïve two-way fixed effects estimators were still examined due to their popularity in DID estimation and ability to yield the desired marginal effect estimate under certain conditions. Estimation of the standard error is expected to be more efficient in versions of outcome models including covariates. Empirical standard errors are expected to outperform model-based standard errors by accounting for the sampling variability of weighted observations. The separate time weighted estimator weights treatment groups separately at each time period. The method accounts for differences in treatment and control groups at each time period, however it fails to take into account changes that might happen between treatment and control groups over time. The four-group weighted design accounts for such possible temporal changes in case mix within each group and also between groups (Stuart et al., 2014). The four-group
weighted approach is expected to achieve improved balance to covariates across all groups and time. Separate time weighted and AIPW estimators might offer improved precision while addressing all significant sources of covariate imbalance compared to the four-group weighted approach at perhaps the compromise of increased bias. An advantage of the AIPW over the separate time weighted estimator is robustness in the situation of either propensity score or outcome model misspecification. Separate time weighted and four-group weighted propensity score approaches are expected to improve covariate balance, thus reducing bias and delivering a more accurate estimate. However, efforts in bias reduction might come at a cost of worsened precision. Bootstrap resampling of weighted observations could offer a correction to the precision.
4. EXAMPLE FROM THE MEDICAL EXPENDITURE PANEL SURVEY

4.1 Introduction

The purpose of the Patient Protection and Affordable Care Act (ACA) of March 2010 was to have universal healthcare coverage in the United States through insurance market reforms, mandates, subsidies, health insurance exchanges, and Medicaid expansion (Gruber, 2011). The ACA expanded Medicaid coverage to include persons up to 138% of the federal poverty level which is approximately an income of $16,400 for a single adult in 2016 (Ayanian et al., 2017). An interesting piece of the ACA’s Medicaid expansion was that the decision to expand was of the states, as ruled by the Supreme Court in June 2012. The condition from the federal government to the states was as follows: the federal government begins with paying the full cost of the newly covered persons; in 2017, the federal share is reduced to 95%, and further reduced to 90% in 2020. Future cost sharing with states is the reason some states opted out of ACA’s Medicaid expansion program. States that chose to participate in the 2014 Medicaid expansion as of October 31, 2013 were Arkansas, Arizona, California, Colorado, Connecticut, Delaware, District of Columbia, Hawaii, Illinois, Indiana, Kentucky, Maryland, Massachusetts, Michigan, Minnesota, Nevada, New Jersey, New Mexico, New York, North Carolina, North Dakota, Ohio, Oregon, Rhode Island, Vermont, Washington, and West Virginia (McMorrow et al., 2015).

The objective of this applied example was to estimate the impact of the 2014 Medicaid expansion on pharmacy utilization and cost using the methods detailed in this dissertation.

4.2 Methods

4.2.1 Data source and study population

The Medical Expenditure Panel Survey (MEPS) is a nationally representative survey of
the noninstitutionalized United States’ non-military population conducted by the Agency for Healthcare Research and Quality. The survey was designed to assess national and regional estimates of annual healthcare use including healthcare utilization, expenditures, medication use, sources of payment, and insurance coverage (Cohen, 2002). The core of MEPS data, the Household Component is collected for two calendar years on selected households. It is a panel design where sample panels are interviewed five times over the course of 30 months. Every year a new panel is added into the survey. A limitation the MEPS data is that it does not allow for state-level estimates (Cohen, 2002). The data used in all analyses were the MEPS’s Household Component files H163 and H171 as available on January 2nd, 2016.

The study population consisted of working age adults greater than or equal to 18 years of age and less than 65 years of age. Older individuals were excluded as healthcare access issues are different for the population over the age of 65, most of whom are Medicare eligible (Horner-Johnson et al., 2014). Households with income less than 138% of the federal poverty level were considered. Henceforth, the measure of federal poverty level will be referred to as poverty. Lower values of poverty correspond to more poor individuals. Data were compared between years 2013 (pre Medicaid expansion) and 2014 (post Medicaid expansion). As of January 1st, 2014, states in the Northeast census region mostly participated in Medicaid expansion (all states expanded except Maine, New Hampshire, and Pennsylvania). New Hampshire later adopted Medicaid expansion on August 15th, 2014 and Pennsylvania on January 1st, 2015. States in the South mostly did not participate in Medicaid expansion (all states did not expand except Arkansas, Delaware, District of Columbia, Kentucky, Maryland, and West Virginia). Louisiana later adopted Medicaid expansion on July 1st, 2016. The West and Midwest census regions were mixed in terms of proportion of Medicaid-expanded states. The Northeast census region (mostly
expanded) was compared to the South region (mostly did not expand). Henceforth we will refer to the Northeast census region as the Medicaid-expanded or treated group, and the South census region as the non-Medicaid-expanded or control group. We acknowledge the fact that treatment groups are not entirely comprised of Medicaid-expanded and non-Medicaid-expanded states. This limitation is further addressed in the discussion section.

The outcomes were yearly number of prescription medications ($RXTOT$) and yearly total cost of Medicaid prescription medications ($RXMCD$). We controlled for variables poverty level, age, gender, race, ethnicity, education (less than high school, high school, college), insurance (any private, public only, uninsured), and MEPS priority conditions ($0, 1, \geq 2$). Priority conditions were determined by the Agency for Healthcare Research and Quality because of their prevalence, expense, and relevance to policy. Priority conditions consisted of: hypertension, heart disease, stroke, emphysema, chronic bronchitis, high cholesterol, cancer, diabetes, joint pain, arthritis, asthma, and attention deficit hyperactivity disorder. Responses of ‘not ascertained,’ ‘don’t know,’ and ‘refused’ in any MEPS priority conditions were coded as missing and excluded from the analysis.

An additional outcome of yearly total healthcare charges ($TOTCH$) was used in a subgroup analysis. Total charges outcome was defined as the sum of direct payments for care. These payments include payments by Medicaid and private insurance as well as patient out-of-pocket payments. Expenditures included in- and out-patient care, home health care, prescribed medications, dental visits, eye glasses, etc. The subgroup analysis was conducted on the group of individuals whose total charges were greater and $0$ and less than $10,000$, and did not have private insurance. The purpose of the subgroup analysis on an additional outcome was to locate an example likely to observe a significant DID treatment effect on a continuous outcome. A
significant effect of a continuous outcome attributed to the policy will later be important for future simulations gauging estimator performance.

4.2.2 Statistical analysis

Continuous variables were described using the mean, standard deviation (SD), skewness, and kurtosis, while categorical variables were described using the frequency and percentage. Differences were tested between treatment and control groups in the pre intervention period and post intervention period separately. Two-sided $t$ tests were used to test for differences in continuous variables, and Chi-square tests were used to test for differences in categorical variables. Correlations between MEPS variables were calculated. Pearson correlations were calculated for continuous-continuous variable combinations, point-biserial for binary-continuous, point-polyserial for continuous-ordinal, phi coefficient for binary-binary, and polychoric correlations for binary-ordinal and ordinal-ordinal variable combinations.

The distribution of propensity scores were analyzed via summary statistics including the sample size, mean, median, SD, 25th percentile, 75th percentile, interquartile range, minimum, and maximum in separate time weighted and four-group weighted models. The distribution of propensity score weights were analyzed both descriptively and visually in histograms by treatment group and time period. Treated subjects with very low propensity scores and control subjects with very high propensity scores should be examined as they are at risk of having extreme weights. Extreme weights in a group can increase the variability of the estimated average treatment effect (Cole and Hernán, 2008).

Propensity score weights were evaluated for covariate balance. Distributions of weighted covariates and standardized differences were used to assess balance. Differences in variables across groups were tested for via weighted $t$ tests and Chi-square tests. A non-significant
difference between covariates indicated good balance. Standardized differences were calculated for unweighted, separate time weighted, and four-group weighted models across treatment group and time period to quantitatively compare covariate balance between the weighted samples. The standardized difference compares the difference in means relative to the pooled SD (Flury and Riedwyl, 1986). Standardized differences for dichotomous variables were also calculated (Austin, 2009b). Dissimilar to $t$ tests and hypothesis tests, the sample size does not affect the standardized difference, and therefore is appropriate for comparing groups in weighted samples (Austin and Stuart, 2015). A standardized bias greater than 10% may suggest there exists meaningful covariate imbalance (Austin, 2009a; Normand et al., 2001; Mamdani et al., 2005).

Difference-in-differences analysis was applied to ascertain any differences in prescription acquisition and the total cost of Medicaid prescriptions attributable to the 2014 Medicaid expansion. Similar analysis was applied to the subgroup of individuals without private insurance and total charges $1-$9,999 to determine any change in total medical charges for this group. Raw DID were calculated using the unweighted differences in mean outcomes of expanded and nonexpanded states between pre and post implementation periods. The distributional characteristics of outcome variables were examined.

Regression estimates from naïve, covariate-adjusted models including all covariates were examined to reveal any significant relationships between covariates and outcomes as well as provide a naïve regression-adjusted estimate of the DID.

The 13 estimators detailed in the prior section were applied to the MEPS data. Difference-in-differences marginal effect point estimates and associated 95% confidence intervals were evaluated and compared across estimators. Stata version 13.1 was used for data
management and R version 3.4.2 was used to evaluate each of the 13 estimators using the MEPS data.

4.3 Results

4.3.1 Study population

There were 36,940 individuals in the pre and 34,875 individuals in the post Medicaid expansion period. Ninety percent of individuals were excluded because of income above 138% of the federal poverty level, age, and census region. Less than 5% of observations were not included in the study due to missing covariate information in the pre and post periods. Of the 3,552 study-eligible individuals with nonmissing covariate information in the pre period, 26% were treated (i.e., in Medicaid-expanded states), and of the 3,278 study-eligible individuals with nonmissing covariate information in the post period, 27% were treated (i.e., in Medicaid-expanded states). A flow chart of the data construction process is presented in Figure 3.

The subgroup of individuals with total charges greater than $0, less than $10,000, and with private insurance consisted of 38% (n=2,592) of the full sample. Of the 1,383 individuals in the subgroup pre period, 29% were treated. Similarly, of the 1,209 individuals in the subgroup post period, 29% were treated.

Distributional characteristics and differences in the study sample demographics, insurance status, and health status between treatment and control groups in pre and post periods are displayed in Table IV. There was no difference between treatment groups in terms of poverty and education in both the pre and post intervention periods. Individuals in Medicaid-expanded states were older than those in nonexpanded states in the pre period (marginally significant, \( p=0.08 \)), however there was no age difference in the post period. The proportion of males to females in the pre period were similar, however there was a greater proportion of males in the
Figure 3. Flowchart of study sample inclusion

Pre-Medicaid Expansion
2013 MEPS Data (n=36,940)

Excluded (n=33,216)
- Poverty > 138% (n=24,346) or Poverty < 0% (n=6)
- Age < 18 or Age >= 65 (n=14,572)
- Midwest (n=6,596) or West (n=10,472) region

Eligible (n=3,724)

Missing covariates (n=172): Gender (n=0), Black (n=67), Hispanic (n=0), Education (n=56), Insurance (n=0), Chronic Conditions (n=63)

Included in study (n=3,552)

Medicaid Expanded (n=916)
Not Medicaid Expanded (n=2,636)

Post-Medicaid Expansion
2014 MEPS Data (n=34,875)

Excluded (n=31,434)
- Poverty > 138% (n=23,116) or Poverty < 0% (n=10)
- Age < 18 or Age >= 65 (n=13,977)
- Midwest (n=6,299) or West (n=9,940) region

Eligible (n=3,441)

Missing covariates (n=163): Gender (n=0), Black (n=81), Hispanic (n=0), Education (n=38), Insurance (n=0), Chronic Conditions (n=57)

Included in study (n=3,278)

Medicaid Expanded (n=871)
Not Medicaid Expanded (n=2,407)
<table>
<thead>
<tr>
<th>Demographic</th>
<th>Medicaid Expanded</th>
<th>Non-Medicaid Expanded</th>
<th>$p$-value$^a$</th>
<th>Standardized Difference, %</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Pre (n=916)</td>
<td>Post (n=871)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>RXTOT$^b$</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mean</td>
<td>14.3</td>
<td>18.3</td>
<td>0.002</td>
<td>12.6</td>
</tr>
<tr>
<td>SD</td>
<td>31.1</td>
<td>34.4</td>
<td>&lt;0.001</td>
<td>24.3</td>
</tr>
<tr>
<td>Skewness</td>
<td>3.4</td>
<td>3.0</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Kurtosis</td>
<td>13.3</td>
<td>12.8</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Pre (n=2,636)</td>
<td>Post (n=2,407)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mean</td>
<td>10.8</td>
<td>11.0</td>
<td>&lt;0.001</td>
<td>24.3</td>
</tr>
<tr>
<td>SD</td>
<td>23.5</td>
<td>25.1</td>
<td>&lt;0.001</td>
<td>12.6</td>
</tr>
<tr>
<td>Skewness</td>
<td>3.5</td>
<td>4.3</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Kurtosis</td>
<td>15.8</td>
<td>28.0</td>
<td></td>
<td></td>
</tr>
<tr>
<td>RXMCD$^b$</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mean</td>
<td>769.8</td>
<td>936.8</td>
<td>0.005</td>
<td>12.1</td>
</tr>
<tr>
<td>SD</td>
<td>4,301.2</td>
<td>3,827.9</td>
<td>&lt;0.001</td>
<td>19.1</td>
</tr>
<tr>
<td>Skewness</td>
<td>14.5</td>
<td>7.6</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Kurtosis</td>
<td>288.7</td>
<td>70.4</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Poverty</td>
<td></td>
<td></td>
<td>0.33</td>
<td>3.8</td>
</tr>
<tr>
<td>Mean</td>
<td>68.6</td>
<td>69.3</td>
<td>0.14</td>
<td>5.7</td>
</tr>
<tr>
<td>SD</td>
<td>41.1</td>
<td>40.4</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Skewness</td>
<td>-0.26</td>
<td>-0.29</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Kurtosis</td>
<td>-1.03</td>
<td>-1.00</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Age, years</td>
<td></td>
<td></td>
<td>0.08</td>
<td>6.6</td>
</tr>
<tr>
<td>Mean</td>
<td>38.7</td>
<td>38.6</td>
<td>0.51</td>
<td>2.6</td>
</tr>
<tr>
<td>SD</td>
<td>13.4</td>
<td>13.5</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Skewness</td>
<td>0.16</td>
<td>0.18</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Kurtosis</td>
<td>-1.14</td>
<td>-1.19</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Male, %</td>
<td></td>
<td></td>
<td>0.32</td>
<td>3.8</td>
</tr>
<tr>
<td>Black, %</td>
<td></td>
<td></td>
<td>&lt;0.001</td>
<td>8.7</td>
</tr>
<tr>
<td>Hispanic, %</td>
<td></td>
<td></td>
<td>&lt;0.001</td>
<td>15.4</td>
</tr>
<tr>
<td>Education, %</td>
<td></td>
<td></td>
<td>&lt;0.001</td>
<td>17.8</td>
</tr>
<tr>
<td>Less than high school</td>
<td>37.2</td>
<td>35.3</td>
<td>34.8</td>
<td>33.1</td>
</tr>
<tr>
<td>High school</td>
<td>35.7</td>
<td>34.2</td>
<td>35.4</td>
<td>36.2</td>
</tr>
<tr>
<td>College</td>
<td>27.1</td>
<td>30.5</td>
<td>29.9</td>
<td>30.7</td>
</tr>
<tr>
<td>Insurance, %</td>
<td></td>
<td></td>
<td>0.22</td>
<td>6.7</td>
</tr>
<tr>
<td>Any private</td>
<td></td>
<td></td>
<td>&lt;0.001</td>
<td>59.6</td>
</tr>
<tr>
<td>Public only</td>
<td></td>
<td></td>
<td>&lt;0.001</td>
<td>59.0</td>
</tr>
<tr>
<td>Uninsured</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
TABLE IV

STUDY SAMPLE DEMOGRAPHICS (continued)

<table>
<thead>
<tr>
<th>Demographic</th>
<th>Medicaid Expanded</th>
<th>Non-Medicaid Expanded</th>
<th>p-value*</th>
<th>Standardized Difference, %</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Pre (n=916)</td>
<td>Post (n=871)</td>
<td>Pre (n=2,636)</td>
<td>Post (n=2,407)</td>
</tr>
<tr>
<td>Conditions, %</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0</td>
<td>36.6</td>
<td>36.3</td>
<td>42.1</td>
<td>38.7</td>
</tr>
<tr>
<td>1</td>
<td>25.6</td>
<td>23.2</td>
<td>20.1</td>
<td>21.4</td>
</tr>
<tr>
<td>≥2</td>
<td>37.9</td>
<td>40.5</td>
<td>37.9</td>
<td>39.9</td>
</tr>
</tbody>
</table>

*a testing for a difference between Medicaid-expanded and non-Medicaid-expanded groups using t tests with unequal variance assumption

*b abbreviations: RXTOT=total number of prescriptions, RXMCD=total spending on Medicaid prescriptions
nonexpanded group in the post period ($p=0.03$). Individuals in nonexpanded states tended to be Black, not Hispanic, uninsured, and have less MEPS priority conditions.

Percent standardized differences in the unweighted sample are also presented in Table IV. Half of the covariates (Black, Hispanic, and insurance in both periods; conditions in the pre period) exhibit standardized differences in excess of 10%. Insurance status had the greatest degree of covariate imbalance between treatment groups of nearly 60%. This is attributable to the nearly double amount of publicly insured individuals in Medicaid-expanded states (pre 61.1% and post 62.0% versus pre 33.0% and post 34.5%) and half the amount of uninsured individuals (pre 25.0% and post 21.8% versus pre 47.6% and post 44.0%) compared to nonexpanded states.

Correlations between the variables are presented in Table V. There were strong positive correlations between the number and cost of prescriptions with variables age and number of priority conditions. There was a negative correlation with number of prescriptions and Hispanic. The strongest correlation was between age and conditions, with greater age associated with more conditions ($r=0.504$). Other strong positive associations were observed between variables Black-education and Black-conditions. Strongest negative correlations were observed between variables Black-Hispanic, Hispanic-education, and Hispanic-conditions.

Study sample demographics in the subgroup are displayed in Table VI. In the subgroup there were no differences between treatment groups in terms of poverty, age, gender, and education in both the pre and post periods. Contrary to the full sample, nonexpanded states tended to be not Black and Hispanic. Nonexpanded states observed more individuals on public insurance and with more comorbidities in the pre period ($p=0.04$). Correlations between variates in the subgroup are displayed in Table VII. Similar to the full sample, age and higher number of
### TABLE V

**CORRELATIONS OF STUDY COVARIATES**

<table>
<thead>
<tr>
<th></th>
<th>RXTOT&lt;sup&gt;a&lt;/sup&gt;</th>
<th>RXMCD&lt;sup&gt;a&lt;/sup&gt;</th>
<th>Poverty</th>
<th>Age</th>
<th>Male</th>
<th>Black</th>
<th>Hispanic</th>
<th>Education</th>
<th>Conditions</th>
</tr>
</thead>
<tbody>
<tr>
<td>RXTOT&lt;sup&gt;a&lt;/sup&gt;</td>
<td>1</td>
<td>0.406</td>
<td>0.031</td>
<td>0.384</td>
<td>0.077</td>
<td>0.044</td>
<td>-0.149</td>
<td>-0.001</td>
<td>0.504</td>
</tr>
<tr>
<td>RXMCD&lt;sup&gt;a&lt;/sup&gt;</td>
<td>1</td>
<td>0.002</td>
<td>0.104</td>
<td>0.032</td>
<td>0.032</td>
<td>0.032</td>
<td>-0.030</td>
<td>-0.008</td>
<td>0.175</td>
</tr>
<tr>
<td>Poverty</td>
<td>1</td>
<td>0.078</td>
<td>-0.079</td>
<td>-0.035</td>
<td>0.014</td>
<td>0.014</td>
<td>0.089</td>
<td>&lt;0.001</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>Age</td>
<td>1</td>
<td>-0.009</td>
<td>0.012</td>
<td>-0.065</td>
<td>-0.018</td>
<td>-0.018</td>
<td>0.584</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Male</td>
<td>1</td>
<td>-0.041</td>
<td>-0.010</td>
<td>-0.094</td>
<td>-0.094</td>
<td>-0.070</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Black</td>
<td>1</td>
<td>-0.558</td>
<td>0.143</td>
<td>0.123</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Hispanic</td>
<td>1</td>
<td>-0.298</td>
<td>-0.276</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Education</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>&lt;0.001</td>
<td></td>
</tr>
<tr>
<td>Conditions</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<sup>a</sup>abbreviations: RXTOT=number of prescriptions, RXMCD=cost of Medicaid prescriptions
## TABLE VI

### STUDY SAMPLE DEMOGRAPHICS IN THE SUBGROUP

<table>
<thead>
<tr>
<th>Demographic</th>
<th>Medicaid Expanded</th>
<th>Non-Medicaid Expanded</th>
<th>p-value&lt;sup&gt;a&lt;/sup&gt;</th>
<th>Standardized Difference, %</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Pre (n=401)</td>
<td>Post (n=354)</td>
<td>Pre (n=982)</td>
<td>Post (n=855)</td>
</tr>
<tr>
<td>Total charges (TOTTCH)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>0.74</td>
<td>0.004</td>
</tr>
<tr>
<td>Mean</td>
<td>2077.7</td>
<td>2575.5</td>
<td>2125.1</td>
<td>2117.0</td>
</tr>
<tr>
<td>SD</td>
<td>2381.7</td>
<td>2575.5</td>
<td>2480.5</td>
<td>2436.3</td>
</tr>
<tr>
<td>Skewness</td>
<td>1.4</td>
<td>0.9</td>
<td>1.4</td>
<td>1.4</td>
</tr>
<tr>
<td>Kurtosis</td>
<td>1.2</td>
<td>-0.1</td>
<td>1.0</td>
<td>1.2</td>
</tr>
<tr>
<td>Poverty</td>
<td></td>
<td></td>
<td>0.37</td>
<td>&gt;0.99</td>
</tr>
<tr>
<td>Mean</td>
<td>64.9</td>
<td>69.0</td>
<td>67.1</td>
<td>69.0</td>
</tr>
<tr>
<td>SD</td>
<td>39.4</td>
<td>38.2</td>
<td>40.0</td>
<td>41.6</td>
</tr>
<tr>
<td>Skewness</td>
<td>-0.3</td>
<td>-0.3</td>
<td>-0.2</td>
<td>-0.2</td>
</tr>
<tr>
<td>Kurtosis</td>
<td>-1.0</td>
<td>-0.9</td>
<td>-1.0</td>
<td>-1.0</td>
</tr>
<tr>
<td>Age, years</td>
<td></td>
<td></td>
<td>0.88</td>
<td>0.97</td>
</tr>
<tr>
<td>Mean</td>
<td>39.0</td>
<td>39.2</td>
<td>38.9</td>
<td>39.3</td>
</tr>
<tr>
<td>SD</td>
<td>13.3</td>
<td>13.2</td>
<td>13.4</td>
<td>13.3</td>
</tr>
<tr>
<td>Skewness</td>
<td>0.1</td>
<td>0.1</td>
<td>0.2</td>
<td>0.1</td>
</tr>
<tr>
<td>Kurtosis</td>
<td>-1.2</td>
<td>-1.1</td>
<td>-1.2</td>
<td>-1.2</td>
</tr>
<tr>
<td>Male, %</td>
<td>35.2</td>
<td>31.9</td>
<td>31.5</td>
<td>33.7</td>
</tr>
<tr>
<td>Black, %</td>
<td>62.6</td>
<td>67.0</td>
<td>57.3</td>
<td>58.8</td>
</tr>
<tr>
<td>Hispanic, %</td>
<td>59.1</td>
<td>55.1</td>
<td>66.9</td>
<td>65.4</td>
</tr>
<tr>
<td>Education, %</td>
<td></td>
<td></td>
<td>0.78</td>
<td>&gt;0.99</td>
</tr>
<tr>
<td>Less than high school</td>
<td>36.9</td>
<td>38.1</td>
<td>37.5</td>
<td>38.1</td>
</tr>
<tr>
<td>High school</td>
<td>36.4</td>
<td>34.8</td>
<td>34.5</td>
<td>34.7</td>
</tr>
<tr>
<td>College</td>
<td>26.7</td>
<td>27.1</td>
<td>28.0</td>
<td>37.1</td>
</tr>
<tr>
<td>Insurance, %</td>
<td></td>
<td></td>
<td>&lt;0.001</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>Public only</td>
<td>75.1</td>
<td>79.7</td>
<td>47.6</td>
<td>51.0</td>
</tr>
<tr>
<td>Uninsured</td>
<td>24.9</td>
<td>20.3</td>
<td>52.4</td>
<td>49.0</td>
</tr>
<tr>
<td>Conditions, %</td>
<td></td>
<td></td>
<td>0.04</td>
<td>0.96</td>
</tr>
<tr>
<td>0</td>
<td>33.9</td>
<td>29.4</td>
<td>31.9</td>
<td>29.4</td>
</tr>
<tr>
<td>1</td>
<td>26.4</td>
<td>23.7</td>
<td>21.5</td>
<td>23.0</td>
</tr>
<tr>
<td>≥2</td>
<td>39.7</td>
<td>46.9</td>
<td>46.6</td>
<td>47.6</td>
</tr>
</tbody>
</table>

<sup>a</sup>testing for a difference between Medicaid-expanded and non-Medicaid-expanded groups using t tests with unequal variance assumption
<table>
<thead>
<tr>
<th></th>
<th>TOTTCH&lt;sup&gt;a&lt;/sup&gt;</th>
<th>Poverty</th>
<th>Age</th>
<th>Male</th>
<th>Black</th>
<th>Hispanic</th>
<th>Education</th>
<th>Conditions</th>
</tr>
</thead>
<tbody>
<tr>
<td>TOTTCH&lt;sup&gt;a&lt;/sup&gt;</td>
<td>1</td>
<td>-0.004</td>
<td>0.112</td>
<td>-0.043</td>
<td>0.055</td>
<td>-0.173</td>
<td>-0.044</td>
<td>0.472</td>
</tr>
<tr>
<td>Poverty</td>
<td>1</td>
<td>0.071</td>
<td>-0.089</td>
<td>-0.035</td>
<td>0.043</td>
<td>0.094</td>
<td>0.003</td>
<td></td>
</tr>
<tr>
<td>Age</td>
<td>1</td>
<td>-0.065</td>
<td>0.009</td>
<td>-0.026</td>
<td>-0.056</td>
<td>0.572</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Male</td>
<td>1</td>
<td>-0.052</td>
<td>0.007</td>
<td>-0.039</td>
<td>-0.048</td>
<td>0.088</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Black</td>
<td>1</td>
<td>-0.558</td>
<td>0.112</td>
<td>0.103</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Hispanic</td>
<td>1</td>
<td>-0.228</td>
<td>0.235</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Education</td>
<td>1</td>
<td>-0.209</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Conditions</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>1</td>
</tr>
</tbody>
</table>

<sup>a</sup>abbreviations: TOTTCH=total charges
priority conditions were associated with greater medical charges. Likewise the strongest correlation was with age and conditions. Correlational relationships were preserved with Black and Hispanic. New to the subgroup was an observed inverse correlation between education and priority conditions.

4.3.2 Propensity score weight diagnostics

Summary statistics of propensity scores and propensity score weights are shown in Table VIII. The separate time weighted model exhibited greater central tendency and spread of propensity scores compared to the four-group weighted model. In the separate time weighted model, mean propensity scores ranged from 0.24-0.33 (SD=0.12-0.14), while mean propensity scores in the four-group weighted model ranged from 0.12-0.17 (SD=0.06-0.07). The interquartile range of estimated propensity scores in the separate time weighted model was double that in the four-group weighted model. In the separate time weighted model, interquartile ranges of propensity scores in pre and post periods were 0.25 and 0.23 in Medicaid-expanded states and 0.16 and 0.17 in nonexpanded states, while the interquartile range in the four-group weighted model was 0.12 and 0.08 for Medicaid-expanded and nonexpanded states, respectively.

Naturally, the central tendency in propensity score derived weights was also greater in the separate time weighted model compared to the four-group weighted model. The mean weight in the separate time weighted model in the pre and post period was 3.93 and 3.81 in expanded states and 1.35 and 1.36 in nonexpanded states. In the four-group weighted model, the mean weight in the pre and post period was 1.00 and 1.05 in expanded states and 0.35 and 0.38 in nonexpanded states. The spread of propensity score derived weights was considerably greater in the separate time weighted model compared to the four-group weighted model in expanded states (SD: separate time weighted: pre=2.23, post=2.28 / four-group weighted: pre=0.00, post=0.16;
### TABLE VIII

**DISTRIBUTION OF PROPENSITY SCORES AND DERIVED WEIGHTS IN SEPARATE TIME WEIGHTED AND FOUR-GROUP WEIGHTED MODELS**

<table>
<thead>
<tr>
<th>Estimator</th>
<th>Period</th>
<th>n</th>
<th>Mean</th>
<th>Median</th>
<th>SD</th>
<th>25th Percentile</th>
<th>75th Percentile</th>
<th>IQR</th>
<th>Minimum</th>
<th>Maximum</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Propensity Scores</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Separate time weighted</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Expanded</td>
<td>Pre</td>
<td>916</td>
<td>0.32</td>
<td>0.32</td>
<td>0.14</td>
<td>0.19</td>
<td>0.44</td>
<td>0.25</td>
<td>0.07</td>
<td>0.63</td>
</tr>
<tr>
<td></td>
<td>Post</td>
<td>871</td>
<td>0.33</td>
<td>0.33</td>
<td>0.14</td>
<td>0.21</td>
<td>0.44</td>
<td>0.23</td>
<td>0.07</td>
<td>0.61</td>
</tr>
<tr>
<td>Nonexpanded</td>
<td>Pre</td>
<td>2636</td>
<td>0.24</td>
<td>0.19</td>
<td>0.12</td>
<td>0.15</td>
<td>0.31</td>
<td>0.16</td>
<td>0.07</td>
<td>0.65</td>
</tr>
<tr>
<td></td>
<td>Post</td>
<td>2407</td>
<td>0.24</td>
<td>0.21</td>
<td>0.12</td>
<td>0.15</td>
<td>0.32</td>
<td>0.17</td>
<td>0.07</td>
<td>0.62</td>
</tr>
<tr>
<td><strong>Four-group weighted</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Expanded</td>
<td>Pre</td>
<td>916</td>
<td>0.17</td>
<td>0.17</td>
<td>0.07</td>
<td>0.10</td>
<td>0.22</td>
<td>0.12</td>
<td>0.04</td>
<td>0.34</td>
</tr>
<tr>
<td></td>
<td>Post</td>
<td>871</td>
<td>0.17</td>
<td>0.17</td>
<td>0.07</td>
<td>0.10</td>
<td>0.22</td>
<td>0.12</td>
<td>0.04</td>
<td>0.34</td>
</tr>
<tr>
<td>Nonexpanded</td>
<td>Pre</td>
<td>2636</td>
<td>0.12</td>
<td>0.10</td>
<td>0.06</td>
<td>0.08</td>
<td>0.16</td>
<td>0.08</td>
<td>0.03</td>
<td>0.35</td>
</tr>
<tr>
<td></td>
<td>Post</td>
<td>2407</td>
<td>0.12</td>
<td>0.11</td>
<td>0.06</td>
<td>0.08</td>
<td>0.16</td>
<td>0.08</td>
<td>0.03</td>
<td>0.35</td>
</tr>
</tbody>
</table>

| **Propensity Score Weights** |        |     |      |        |     |                 |                 |      |         |         |
| **Separate time weighted** |        |     |      |        |     |                 |                 |      |         |         |
| Expanded          | Pre    | 916 | 3.93 | 3.10   | 2.23| 2.27            | 5.14            | 2.87 | 1.60    | 13.95   |
|                   | Post   | 871 | 3.81 | 3.01   | 2.28| 2.26            | 4.79            | 2.53 | 1.63    | 14.46   |
| Nonexpanded       | Pre    | 2636| 1.35 | 1.24   | 0.25| 1.17            | 1.46            | 0.29 | 1.07    | 2.83    |
|                   | Post   | 2407| 1.36 | 1.26   | 0.27| 1.18            | 1.48            | 0.30 | 1.07    | 2.63    |
| **Four-group weighted** |        |     |      |        |     |                 |                 |      |         |         |
| Expanded          | Pre    | 916 | 1.00 | 1.00   | 0.00| 1.00            | 1.00            | 0.00 | 1.00    | 1.00    |
|                   | Post   | 871 | 1.05 | 1.04   | 0.16| 0.94            | 1.16            | 0.22 | 0.71    | 1.53    |
| Nonexpanded       | Pre    | 2636| 0.35 | 0.24   | 0.25| 0.17            | 0.46            | 0.29 | 0.07    | 1.84    |
|                   | Post   | 2407| 0.38 | 0.29   | 0.27| 0.19            | 0.49            | 0.30 | 0.07    | 1.86    |
interquartile range: separate time weighted: pre=2.87, post=2.53 / four-group weighted: pre=0.00, post=0.22). Interestingly, the spread of weights was exactly the same in separate time weighted as the four-group weighted model in nonexpanded states (SD: pre=0.25, post=0.27; interquartile range: pre=0.29, post=0.30).

The distribution in propensity scores are presented visually in Figure 4 for the separate time weighted model and in Figure 5 for the four-group weighted model. It can be seen visually the variance of the weights is greater for the separate time weights compared to the four-group weights. The propensity score distributions by treatment group look similar when comparing both models. Expanded states exhibited a bimodal distribution, while nonexpanded states exhibited a right-skewed distribution.

Propensity score weight distributions are presented in Figure 6 for the separate time weighted and Figure 7 for the four-group weighted model. All distributions of weights were right skewed and slightly bimodal, except for the pre treatment group in the four-group weighted model, of which all observations were given a weight of one as this was the reference group. The considerable difference in propensity weight’s spread between expanded (min=1.60, max=14.46) and nonexpanded states (min=1.07, max=2.83) in the separate time weighted model is noteworthy from the histograms in Figure 6. The distribution of propensity score weights in the four-group weighted model are displayed in Figure 7. Compared individually, separate time weighted and four-group weighted models’ weights look similar within expanded states and nonexpanded states between pre and post periods, with the exception of the reference group in the four-group weighted model. This offers the impression that distributional attributes of subjects’ covariates did not change much within treatment groups across pre and post periods.

Weighted study sample demographics and percent standardized differences in separate
Figure 4. Distribution of propensity scores using a separate time weighted estimator
Figure 5. Distribution of propensity scores using a four-group weighted estimator
Figure 6. Distribution of propensity score weights using a separate time weighted estimator
Figure 7. Distribution of propensity score weights using a four-group weighted estimator.
time weighted and four-group weighted models are presented in Tables IX and X, respectively. Weighted sample demographics offer notable improvement to the covariate balance in Medicaid-expanded and nonexpanded groups compared to the unweighted sample. Standardized differences in both weighted models are all less than 10% for all groups. This is indicative that both separate time weighted and four-group weighted models were sufficient in balancing covariates across treatment groups and time periods.

Table XI presents summary statistics of percent standardized differences in the unweighted and weighted samples. The distributional composition of the treatment groups remain relatively constant between pre and post periods. This can be seen from percent standardized differences no greater than 10% in the unweighted sample of Medicaid-expanded and non-Medicaid-expanded groups. These results indicate that a weighting scheme to balance treatment groups across time periods, within group, may not be necessary. The true value of the weighting techniques in this sample is in the balancing of covariates across treatment groups, not necessarily across time periods.

Propensity score weighting methods in separate time weighted and four-group weighted models offered considerable improvement to the balance of treatment groups. The four-group weighted model did a better job compared to the separate time weighted model in balancing covariates both across treatment groups and time periods. The mean standardized difference was approximately 2% in the separate time weighted model and 1% in the four-group weighted model, compared to 15% in the unweighted model. All percent standardized differences in the separate time weighted and four-group weighted samples were less than 10% and considered acceptable.
### TABLE IX

**SEPARATE TIME WEIGHTED STUDY SAMPLE DEMOGRAPHICS**

<table>
<thead>
<tr>
<th>Demographic</th>
<th>Medicaid Expanded</th>
<th>Non-Medicaid Expanded</th>
<th>Standardized Difference, %</th>
<th>Standardized Difference, %</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Pre (n=3598.8)</td>
<td>Post (n=3322.8)</td>
<td>Medicaid Expanded</td>
<td>Non-Medicaid Expanded</td>
</tr>
<tr>
<td>Poverty</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mean</td>
<td>69.6</td>
<td>70.4</td>
<td>69.7</td>
<td>71.0</td>
</tr>
<tr>
<td>SD</td>
<td>41.7</td>
<td>41.8</td>
<td>41.4</td>
<td>42.1</td>
</tr>
<tr>
<td>Age, years</td>
<td></td>
<td></td>
<td>1.6</td>
<td></td>
</tr>
<tr>
<td>Mean</td>
<td>37.4</td>
<td>37.7</td>
<td>38.0</td>
<td>38.3</td>
</tr>
<tr>
<td>SD</td>
<td>13.0</td>
<td>13.1</td>
<td>13.5</td>
<td>13.3</td>
</tr>
<tr>
<td>Male, %</td>
<td>41.2</td>
<td>40.1</td>
<td>39.3</td>
<td>38.2</td>
</tr>
<tr>
<td>Black, %</td>
<td>40.1</td>
<td>40.7</td>
<td>39.4</td>
<td>40.0</td>
</tr>
<tr>
<td>Hispanic, %</td>
<td>34.4</td>
<td>33.1</td>
<td>36.2</td>
<td>34.5</td>
</tr>
<tr>
<td>Education, %</td>
<td></td>
<td></td>
<td>6.3</td>
<td></td>
</tr>
<tr>
<td>Less than high school</td>
<td>35.0</td>
<td>32.0</td>
<td>35.2</td>
<td>33.4</td>
</tr>
<tr>
<td>High school</td>
<td>34.5</td>
<td>36.2</td>
<td>35.5</td>
<td>35.7</td>
</tr>
<tr>
<td>College</td>
<td>30.5</td>
<td>31.7</td>
<td>29.3</td>
<td>30.8</td>
</tr>
<tr>
<td>Insurance, %</td>
<td></td>
<td></td>
<td>6.6</td>
<td></td>
</tr>
<tr>
<td>Any private</td>
<td>18.0</td>
<td>19.7</td>
<td>18.0</td>
<td>20.1</td>
</tr>
<tr>
<td>Public only</td>
<td>39.3</td>
<td>40.7</td>
<td>40.1</td>
<td>41.7</td>
</tr>
<tr>
<td>Uninsured</td>
<td>42.7</td>
<td>39.7</td>
<td>41.9</td>
<td>38.2</td>
</tr>
<tr>
<td>Conditions, %</td>
<td></td>
<td></td>
<td>4.5</td>
<td></td>
</tr>
<tr>
<td>0</td>
<td>41.2</td>
<td>39.1</td>
<td>40.6</td>
<td>38.1</td>
</tr>
<tr>
<td>1</td>
<td>21.4</td>
<td>21.7</td>
<td>21.5</td>
<td>22.0</td>
</tr>
<tr>
<td>≥2</td>
<td>37.4</td>
<td>39.2</td>
<td>37.9</td>
<td>39.9</td>
</tr>
</tbody>
</table>
### TABLE X

**FOUR-GROUP WEIGHTED STUDY SAMPLE DEMOGRAPHICS**

<table>
<thead>
<tr>
<th>Demographic</th>
<th>Medicaid Expanded</th>
<th>Non-Medicaid Expanded</th>
<th>Standardized Difference, %</th>
<th>Standardized Difference, %</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Pre (n=916.0)</td>
<td>Post (n=916.4)</td>
<td>Pre (n=910.8)</td>
<td>Post (n=910.4)</td>
</tr>
<tr>
<td>Poverty</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mean</td>
<td>68.6</td>
<td>68.6</td>
<td>68.9</td>
<td>68.2</td>
</tr>
<tr>
<td>SD</td>
<td>41.1</td>
<td>40.5</td>
<td>41.2</td>
<td>42.1</td>
</tr>
<tr>
<td>Age, years</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mean</td>
<td>38.7</td>
<td>38.6</td>
<td>38.4</td>
<td>38.5</td>
</tr>
<tr>
<td>SD</td>
<td>13.4</td>
<td>13.4</td>
<td>13.9</td>
<td>13.6</td>
</tr>
<tr>
<td>Male, %</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mean</td>
<td>37.7</td>
<td>37.9</td>
<td>38.5</td>
<td>38.1</td>
</tr>
<tr>
<td>SD</td>
<td>3.3</td>
<td>3.3</td>
<td>3.3</td>
<td>3.3</td>
</tr>
<tr>
<td>Black, %</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mean</td>
<td>41.3</td>
<td>41.3</td>
<td>41.0</td>
<td>39.4</td>
</tr>
<tr>
<td>SD</td>
<td>1.9</td>
<td>1.9</td>
<td>1.9</td>
<td>1.9</td>
</tr>
<tr>
<td>Hispanic, %</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Education, %</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Less than high school</td>
<td>37.2</td>
<td>37.0</td>
<td>37.1</td>
<td>35.8</td>
</tr>
<tr>
<td>High school</td>
<td>35.7</td>
<td>35.8</td>
<td>35.5</td>
<td>36.2</td>
</tr>
<tr>
<td>College</td>
<td>27.1</td>
<td>27.3</td>
<td>27.4</td>
<td>27.9</td>
</tr>
<tr>
<td>Insurance, %</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Any private</td>
<td>13.9</td>
<td>13.8</td>
<td>14.3</td>
<td>13.7</td>
</tr>
<tr>
<td>Public only</td>
<td>61.1</td>
<td>61.1</td>
<td>60.5</td>
<td>60.7</td>
</tr>
<tr>
<td>Uninsured</td>
<td>25.0</td>
<td>25.1</td>
<td>25.1</td>
<td>25.6</td>
</tr>
<tr>
<td>Conditions, %</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0</td>
<td>36.6</td>
<td>36.8</td>
<td>37.3</td>
<td>36.0</td>
</tr>
<tr>
<td>1</td>
<td>25.5</td>
<td>25.7</td>
<td>25.1</td>
<td>26.3</td>
</tr>
<tr>
<td>≥2</td>
<td>37.9</td>
<td>37.5</td>
<td>37.6</td>
<td>37.7</td>
</tr>
</tbody>
</table>
TABLE XI

SUMMARY STATISTICS OF PERCENT STANDARDIZED DIFFERENCES

<table>
<thead>
<tr>
<th></th>
<th>Medicaid Expanded</th>
<th>Non-Medicaid Expanded</th>
<th>Pre</th>
<th>Post</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Unweighted</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mean</td>
<td>4.3</td>
<td>4.3</td>
<td>15.4</td>
<td>15.7</td>
</tr>
<tr>
<td>SD</td>
<td>3.3</td>
<td>2.2</td>
<td>18.4</td>
<td>18.7</td>
</tr>
<tr>
<td>Minimum</td>
<td>0.4</td>
<td>0.6</td>
<td>3.8</td>
<td>2.6</td>
</tr>
<tr>
<td>Maximum</td>
<td>9.0</td>
<td>7.7</td>
<td>59.6</td>
<td>59.0</td>
</tr>
<tr>
<td><strong>Separate time weighted</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mean</td>
<td>3.4</td>
<td>3.8</td>
<td>2.4</td>
<td>2.8</td>
</tr>
<tr>
<td>SD</td>
<td>2.1</td>
<td>2.1</td>
<td>1.4</td>
<td>1.2</td>
</tr>
<tr>
<td>Minimum</td>
<td>1.1</td>
<td>1.3</td>
<td>0.3</td>
<td>1.3</td>
</tr>
<tr>
<td>Maximum</td>
<td>6.6</td>
<td>8.0</td>
<td>3.9</td>
<td>4.8</td>
</tr>
<tr>
<td><strong>Four-group weighted</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mean</td>
<td>0.5</td>
<td>2.1</td>
<td>1.2</td>
<td>1.6</td>
</tr>
<tr>
<td>SD</td>
<td>0.3</td>
<td>1.0</td>
<td>0.5</td>
<td>1.2</td>
</tr>
<tr>
<td>Minimum</td>
<td>0.1</td>
<td>0.5</td>
<td>0.6</td>
<td>0.5</td>
</tr>
<tr>
<td>Maximum</td>
<td>0.8</td>
<td>3.3</td>
<td>1.8</td>
<td>3.9</td>
</tr>
</tbody>
</table>
The distribution of propensity scores and derived weights in the subgroup are presented in Table XII. Similar to propensity scores and associated weights in the full sample, separate time weights had substantially greater variability. The main difference was again attributed to the Medicaid-expanded group. In the subgroup as well, separate time weighted and four-group weighted estimators offered vast improvement to covariate balance (data not shown).

Distributions of propensity scores from separate time weighted and four-group weighted models in the subgroup are displayed in Figures 8 and 9. Similar to the full sample, propensity scores in the separate time weighted model exhibit greater variability compared to the four-group weighted model. In turn, the propensity score weights in the separate time weighted model have greater variability compared to the four-group weighted model. Distributions of propensity score weights are shown in Figures 10 and 11.

Standardized differences of covariate balance across group and time periods in the subgroup are shown in Table XIII. Similar to the full sample, the standardized differences within treatment group, across pre and post time periods are much lower compared to within time period, across treatment groups. This indicates for the subgroup as well, the big gains to be made in covariate balance are within time period. Separate time weighted and four-group weighted models achieved sufficient covariate balance as indicated by all standardized differences less than 10%. The four-group weighted model resulted in better covariate balance compared to the separate time weighted model.

4.3.3 Prescription utilization and cost

Unweighted mean differences in number of prescriptions and total cost of Medicaid prescriptions are displayed in Table XIV. There was an increase of 3.95 prescriptions for individuals in Medicaid-expanded states \( (p=0.002) \), and no difference for individuals in non-
### TABLE XII

**DISTRIBUTION OF PROPENSITY SCORES AND DERIVED WEIGHTS IN SEPARATE TIME WEIGHTED AND FOUR-GROUP WEIGHTED MODELS IN THE SUBGROUP**

<table>
<thead>
<tr>
<th>Estimator</th>
<th>Period</th>
<th>n</th>
<th>Mean</th>
<th>Median</th>
<th>SD</th>
<th>25th Percentile</th>
<th>75th Percentile</th>
<th>IQR</th>
<th>Minimum</th>
<th>Maximum</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Propensity Scores</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Separate time weighted</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Expanded</td>
<td>Pre</td>
<td>401</td>
<td>0.36</td>
<td>0.36</td>
<td>0.14</td>
<td>0.26</td>
<td>0.47</td>
<td>0.21</td>
<td>0.07</td>
<td>0.68</td>
</tr>
<tr>
<td></td>
<td>Post</td>
<td>354</td>
<td>0.37</td>
<td>0.35</td>
<td>0.14</td>
<td>0.30</td>
<td>0.51</td>
<td>0.21</td>
<td>0.07</td>
<td>0.61</td>
</tr>
<tr>
<td>Nonexpanded</td>
<td>Pre</td>
<td>982</td>
<td>0.26</td>
<td>0.24</td>
<td>0.13</td>
<td>0.15</td>
<td>0.72</td>
<td>0.57</td>
<td>0.06</td>
<td>0.72</td>
</tr>
<tr>
<td></td>
<td>Post</td>
<td>855</td>
<td>0.26</td>
<td>0.28</td>
<td>0.14</td>
<td>0.13</td>
<td>0.34</td>
<td>0.21</td>
<td>0.07</td>
<td>0.61</td>
</tr>
<tr>
<td>Four-group weighted</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Expanded</td>
<td>Pre</td>
<td>401</td>
<td>0.19</td>
<td>0.19</td>
<td>0.07</td>
<td>0.14</td>
<td>0.24</td>
<td>0.10</td>
<td>0.04</td>
<td>0.38</td>
</tr>
<tr>
<td></td>
<td>Post</td>
<td>354</td>
<td>0.19</td>
<td>0.19</td>
<td>0.07</td>
<td>0.14</td>
<td>0.23</td>
<td>0.09</td>
<td>0.04</td>
<td>0.38</td>
</tr>
<tr>
<td>Nonexpanded</td>
<td>Pre</td>
<td>982</td>
<td>0.14</td>
<td>0.13</td>
<td>0.07</td>
<td>0.08</td>
<td>0.19</td>
<td>0.11</td>
<td>0.03</td>
<td>0.38</td>
</tr>
<tr>
<td></td>
<td>Post</td>
<td>855</td>
<td>0.14</td>
<td>0.14</td>
<td>0.07</td>
<td>0.08</td>
<td>0.19</td>
<td>0.11</td>
<td>0.03</td>
<td>0.38</td>
</tr>
<tr>
<td><strong>Propensity Score Weights</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Separate time weighted</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Expanded</td>
<td>Pre</td>
<td>401</td>
<td>3.50</td>
<td>2.74</td>
<td>2.23</td>
<td>2.13</td>
<td>3.82</td>
<td>1.69</td>
<td>1.48</td>
<td>14.19</td>
</tr>
<tr>
<td></td>
<td>Post</td>
<td>354</td>
<td>3.44</td>
<td>2.86</td>
<td>2.22</td>
<td>1.96</td>
<td>3.33</td>
<td>1.37</td>
<td>1.65</td>
<td>13.38</td>
</tr>
<tr>
<td>Nonexpanded</td>
<td>Pre</td>
<td>982</td>
<td>1.41</td>
<td>1.31</td>
<td>0.31</td>
<td>1.17</td>
<td>1.56</td>
<td>0.39</td>
<td>1.06</td>
<td>3.52</td>
</tr>
<tr>
<td></td>
<td>Post</td>
<td>855</td>
<td>1.41</td>
<td>1.38</td>
<td>0.32</td>
<td>1.15</td>
<td>1.52</td>
<td>0.37</td>
<td>1.08</td>
<td>2.57</td>
</tr>
<tr>
<td>Four-group weighted</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Expanded</td>
<td>Pre</td>
<td>401</td>
<td>1.00</td>
<td>1.00</td>
<td>0.00</td>
<td>1.00</td>
<td>1.00</td>
<td>0.00</td>
<td>1.00</td>
<td>1.00</td>
</tr>
<tr>
<td></td>
<td>Post</td>
<td>354</td>
<td>1.13</td>
<td>1.10</td>
<td>0.32</td>
<td>0.89</td>
<td>1.32</td>
<td>0.43</td>
<td>0.54</td>
<td>2.56</td>
</tr>
<tr>
<td>Nonexpanded</td>
<td>Pre</td>
<td>982</td>
<td>0.41</td>
<td>0.32</td>
<td>0.31</td>
<td>0.17</td>
<td>0.56</td>
<td>0.39</td>
<td>0.06</td>
<td>0.46</td>
</tr>
<tr>
<td></td>
<td>Post</td>
<td>855</td>
<td>0.47</td>
<td>0.40</td>
<td>0.31</td>
<td>0.22</td>
<td>0.63</td>
<td>0.41</td>
<td>0.08</td>
<td>2.23</td>
</tr>
</tbody>
</table>
Figure 8. Distribution of propensity scores using a separate time weighted estimator in the subgroup
Figure 9. Distribution of propensity scores using a four-group weighted estimator in the subgroup.
Figure 10. Distribution of propensity score weights using a separate time weighted estimator in the subgroup
Figure 11. Distribution of propensity score weights using a four-group weighted estimator in the subgroup.
### TABLE XIII

**SUMMARY STATISTICS OF PERCENT STANDARDIZED DIFFERENCES IN THE SUBGROUP**

<table>
<thead>
<tr>
<th></th>
<th>Medicaid Expanded</th>
<th>Non-Medicaid Expanded</th>
<th>Pre</th>
<th>Post</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Unweighted</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mean</td>
<td>7.2</td>
<td>4.1</td>
<td>14.9</td>
<td>13.4</td>
</tr>
<tr>
<td>SD</td>
<td>3.6</td>
<td>1.7</td>
<td>18.5</td>
<td>21.7</td>
</tr>
<tr>
<td>Minimum</td>
<td>1.4</td>
<td>2</td>
<td>0.9</td>
<td>0.1</td>
</tr>
<tr>
<td>Maximum</td>
<td>11</td>
<td>6.9</td>
<td>58.9</td>
<td>63.2</td>
</tr>
<tr>
<td><strong>Separate time weighted</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mean</td>
<td>4.7</td>
<td>4.5</td>
<td>3.1</td>
<td>2.5</td>
</tr>
<tr>
<td>SD</td>
<td>2.6</td>
<td>2.6</td>
<td>1.4</td>
<td>0.9</td>
</tr>
<tr>
<td>Minimum</td>
<td>1</td>
<td>0.6</td>
<td>0.5</td>
<td>1.5</td>
</tr>
<tr>
<td>Maximum</td>
<td>8</td>
<td>7.8</td>
<td>4.5</td>
<td>4.2</td>
</tr>
<tr>
<td><strong>Four-group weighted</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mean</td>
<td>1.2</td>
<td>2.5</td>
<td>1.6</td>
<td>2.1</td>
</tr>
<tr>
<td>SD</td>
<td>0.6</td>
<td>1.5</td>
<td>1.1</td>
<td>0.9</td>
</tr>
<tr>
<td>Minimum</td>
<td>0.6</td>
<td>0.4</td>
<td>0.1</td>
<td>0.7</td>
</tr>
<tr>
<td>Maximum</td>
<td>2.6</td>
<td>5.3</td>
<td>3.4</td>
<td>3.3</td>
</tr>
</tbody>
</table>
TABLE XIV

UNWEIGHTED MEANS FOR NUMBER OF PRESCRIPTIONS, COST OF MEDICAID PRESCRIPTIONS, AND TOTAL CHARGES

<table>
<thead>
<tr>
<th></th>
<th>Pre</th>
<th>Post</th>
<th>Post-Pre Policy Difference (p-value)</th>
<th>DID (p-value&lt;sup&gt;a&lt;/sup&gt;)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of prescriptions&lt;sup&gt;b&lt;/sup&gt;</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Expanded</td>
<td>14.30</td>
<td>18.25</td>
<td>3.95 (0.002)</td>
<td>3.82 (0.01)</td>
</tr>
<tr>
<td>Nonexpanded</td>
<td>10.82</td>
<td>10.95</td>
<td>0.13 (0.86)</td>
<td></td>
</tr>
<tr>
<td>Cost of prescriptions&lt;sup&gt;b&lt;/sup&gt;</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Expanded</td>
<td>769.8</td>
<td>936.8</td>
<td>166.9 (0.21)</td>
<td>168.5 (0.28)</td>
</tr>
<tr>
<td>Nonexpanded</td>
<td>349.2</td>
<td>347.6</td>
<td>-1.6 (0.98)</td>
<td></td>
</tr>
<tr>
<td>Total charges&lt;sup&gt;c&lt;/sup&gt;</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Expanded</td>
<td>2077.7</td>
<td>2575.5</td>
<td>497.8 (0.006)</td>
<td>505.9 (0.02)</td>
</tr>
<tr>
<td>Nonexpanded</td>
<td>2125.1</td>
<td>2117.0</td>
<td>-8.1 (0.94)</td>
<td></td>
</tr>
</tbody>
</table>

<sup>a</sup>standard errors are estimated by linear regression

<sup>b</sup>full sample (n=6,830)

<sup>c</sup>subgroup of individuals with low medical charges ($1-$9,999) and without private insurance (n=2,592)

Medicaid-expanded states following the 2014 Medicaid expansion. The DID in number of prescriptions was 3.82 (<i>p</i>=0.01). This can be interpreted as the effect of Medicaid expansion policy was associated with an increase of nearly four prescription medications per year, per individual after accounting for background differences. In terms of the total cost in Medicaid prescriptions, individuals in Medicaid-expanded states observed an increase of $167, though this increase was non-significant (<i>p</i>=0.21). Individuals in non-Medicaid-expanded states also observed non-significant changes in cost of prescriptions (-$1.6, <i>p</i>=0.98) following the rollout of
the 2014 Medicaid expansion. The DID in total cost of Medicaid prescriptions of nearly $169, while substantial, was not significant ($p=0.28).

It is interesting to note that individuals in Medicaid-expanded states acquired more medications and spent over double the amount of money for Medicaid prescriptions compared to non-Medicaid-expanded states following the 2014 ACA’s Medicaid expansion. This is likely due to the near double proportion of individuals on public insurance in Medicaid-expanded states. Increases in prescription acquisition and cost were observed in Medicaid-expanded states following the policy rollout. Conversely, the prescription market in non-Medicaid-expanded states remained constant during this period. The final conclusion gleaned from unadjusted, unweighted estimates is such that Medicaid expansion was associated with an increase of nearly four additional medications per individual, while not significantly influencing cost.

Absolute mean differences for total charges in the subgroup of individuals with low medical charges ($1-$9,999) and without private insurance are also displayed in Table XIV. The Medicaid-expanded group observed a $500 increase in medical charges following the introduction of Medicaid expansion ($p=0.006). The non-Medicaid-expanded group did not observe any change in medical charges during this period ($p=0.94). The DID in total charges for the subgroup was $506 ($p=0.02). This can be interpreted as the effect of Medicaid expansion policy was associated with a $500 yearly spending increase per individual among individuals with modest utilization and non-private insurance.

Both outcome measures exhibited considerable positive skewness. This is typical of healthcare utilization and cost data. Oftentimes few individuals incur very high costs while the majority of individuals incur minimal costs. Differences of sample means in skewed data can be
inefficient and sensitive to individual observations especially when the sample sizes between groups are vastly different (Griswold et al., 2004).

A modified Park test (Park, 1966) was conducted to select a particular family for each of the three outcome distributions. The test recommended the Poisson family with log link for the prescription count ($\lambda=1.45$), Poisson or Gamma family with log link for the Medicaid prescription cost ($\lambda=1.34$), and inverse Gaussian (Wald family) with log link for total charges ($\lambda=4.84$). In final models, number of prescriptions was represented by the Poisson family with log link. Medicaid prescription cost was represented by the Gamma family with log link as research has demonstrated the Gamma distribution is an excellent choice for modeling healthcare cost data by offering flexibility and robustness to the mean estimate (Basu and Manning, 2009). In Gamma generalized linear models of the Medicaid prescription cost data, a small positive value was added to the outcome. This small positive value was chosen by first adding $1$ to each cost outcome and incrementally increasing by $1$ until the model estimates converged. The final constant added to each outcome was determined to be $15$. Total charges were modeled via an inverse Gaussian generalized linear model. Difference-in-differences of generalized linear models were represented as the marginal effect of the interaction term between treatment and post indicators with delta method calculated standard errors using the function \textit{marg} in R package \texttt{modmarg} (Gold et al., 2017).

Regression coefficient estimates from covariate adjusted linear and generalized linear naïve models are displayed in Table XV. We begin by examining the coefficient estimates in the linear models. The DID in number and cost of prescriptions, as measured by the interaction term between Medicaid-expanded and post indicators, in adjusted models was similar to the unadjusted estimates in Table XIV. Adjusted models indicate Medicaid expansion was
TABLE XV
REGRESSION COEFFICIENT ESTIMATES FROM COVARAITE ADJUSTED TWO-WAY
FIXED EFFECTS MODELS

<table>
<thead>
<tr>
<th>Variable</th>
<th>Linear Number of prescriptions</th>
<th>Linear Cost of prescriptions</th>
<th>Generalized Linear Model Number of prescriptions&lt;sup&gt;a&lt;/sup&gt;</th>
<th>Generalized Linear Model Cost of prescriptions&lt;sup&gt;b&lt;/sup&gt;</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Estimate</td>
<td>p-value</td>
<td>Estimate</td>
<td>p-value</td>
</tr>
<tr>
<td>Medicaid Expanded</td>
<td>-0.50</td>
<td>0.57</td>
<td>173.5</td>
<td>0.11</td>
</tr>
<tr>
<td>Post</td>
<td>-0.71</td>
<td>0.26</td>
<td>-34.0</td>
<td>0.66</td>
</tr>
<tr>
<td>Medicaid Expanded x Post</td>
<td>3.99</td>
<td>0.001</td>
<td>174.1</td>
<td>0.26</td>
</tr>
<tr>
<td>Poverty</td>
<td>0.02</td>
<td>0.001</td>
<td>0.80</td>
<td>0.33</td>
</tr>
<tr>
<td>Age</td>
<td>0.45</td>
<td>&lt;0.001</td>
<td>9.37</td>
<td>0.002</td>
</tr>
<tr>
<td>Male</td>
<td>-0.59</td>
<td>0.30</td>
<td>36.0</td>
<td>0.61</td>
</tr>
<tr>
<td>Black</td>
<td>-1.98</td>
<td>0.003</td>
<td>116.8</td>
<td>0.16</td>
</tr>
<tr>
<td>Hispanic</td>
<td>-2.22</td>
<td>0.002</td>
<td>165.1</td>
<td>0.07</td>
</tr>
<tr>
<td>Education</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Less than high school</td>
<td>ref</td>
<td></td>
<td>ref</td>
<td></td>
</tr>
<tr>
<td>High school</td>
<td>0.17</td>
<td>0.80</td>
<td>-58.1</td>
<td>0.48</td>
</tr>
<tr>
<td>College</td>
<td>1.02</td>
<td>0.15</td>
<td>64.7</td>
<td>0.46</td>
</tr>
<tr>
<td>Insurance</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Any private</td>
<td>ref</td>
<td></td>
<td>ref</td>
<td></td>
</tr>
<tr>
<td>Public only</td>
<td>11.77</td>
<td>&lt;0.001</td>
<td>760.1</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>Uninsured</td>
<td>-1.59</td>
<td>0.05</td>
<td>-116.3</td>
<td>0.24</td>
</tr>
<tr>
<td>Conditions</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0</td>
<td>ref</td>
<td></td>
<td>ref</td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>-0.64</td>
<td>0.38</td>
<td>16.6</td>
<td>0.86</td>
</tr>
<tr>
<td>≥2</td>
<td>16.1</td>
<td>&lt;0.001</td>
<td>634.3</td>
<td>&lt;0.001</td>
</tr>
</tbody>
</table>

<sup>a</sup>Poisson generalized linear model, log link

<sup>b</sup>Gamma generalized linear model, log link
associated with a significant increase of four additional prescription medications \( (p=0.001) \) and a non-significant increase of $174 in Medicaid prescription spending \( (p=0.26) \). Less poor (higher poverty variable) and older individuals were associated with higher prescription acquisition. Every two additional years in age of an individual was associated with nearly one additional yearly prescription medication \( (p<0.001) \). Black and Hispanic individuals were associated with less medication prescriptions. Persons exclusively with public insurance were associated with an additional 12 yearly prescriptions compared to persons with private insurance \( (p<0.001) \). Having two or more conditions was associated with an additional 16 yearly prescriptions \( (p<0.001) \). Every additional year in age was associated with an increased cost of $9.37 in Medicaid prescriptions \( (p=0.002) \). Having exclusively public insurance versus private and at least two conditions was associated with considerable additional cost in Medicaid prescriptions, $760 \( (p<0.001) \) and $634 \( (p<0.001) \), respectively.

Regression parameter estimates from a Poisson model indicate different relationships between covariates and prescription count. A Poisson regression model of prescription counts preserved the significant associations of the linear model specification while expressing several new, highly significant associations. Poverty and age still exhibit significant positive associations, as well as Black and Hispanic still exhibit significant negative associations with prescription count \( (all \ p<0.001) \). Having exclusively public insurance versus private and at least two conditions remained strongly associated with additional prescription medications. However in Poisson prescription count models, uninsured versus private insurance and one condition were now highly significant predictors \( (both \ p<0.001) \). Uninsured individuals versus those with private insurance acquired less prescription medications \( (p<0.001) \). Furthermore, education in Poisson models was significant, indicating more education was associated with greater
 prescription acquisition (all levels, \( p<0.001 \)). The post expansion indicator and gender were also now significant predictors (both \( p<0.001 \)), while in linear models they were not.

The Gamma generalized linear models of prescription cost likewise revealed some new interesting associations. Similar to prescription count models, Gamma cost models observed a significant association between uninsured versus private insurance and having at least one priority condition, while in linear models, they were not. The Medicaid-expanded indicator was significant in the Gamma model, while it was non-significant in the linear model. Directional relationships and significance were preserved among poverty, age, male, Black, Hispanic, and education variables. Linear and Gamma generalized linear models concluded a non-significant effect for the coefficient of interest, the Medicaid-expanded and post interaction term. Careful consideration need be taken into the model specification as different specifications might indicate different associative relationships.

Difference-in-differences using the 13 estimators in linear model specifications were computed for the acquisition and total cost of Medicaid prescriptions following the 2014 Medicaid expansion. Five hundred replications were used in bootstrap standard error estimates. Table XVI presents these DID estimates and associated 95% confidence intervals. Figure 12 presents these values visually in forest plots. Grey vertical lines were drawn in forest plots to represent the unadjusted, unweighted mean DID estimates of MEPS data (3.8 medications and $169). All 95% confidence intervals of the 13 estimators covered the unweighted mean prescription count and cost DID estimates. There is a large variation of DID point estimates using the 13 estimators, ranging from 3.58 to 4.12 prescriptions and $88 to $220 in total cost of Medicaid prescriptions. The substantial range of DID point estimates in Medicaid prescription cost is due to its non-significance, resulting in the erratic accuracy of point estimates. The naïve
two-way fixed effects. When covariates were included in the naïve two-way fixed effects model (estimator 2), the coverage length for number of prescriptions was shortened. The DID estimate for the total cost of Medicaid prescriptions remained similar to the naïve estimate not including covariates ($174 versus $169), although both were non-significant. Separate time weighted models with model-derived standard errors (estimators 3 and 4) have the lowest point estimates compared to all other models. When covariates are included in separate time weighted models (estimator 4), the standard error was reduced. The bootstrap median estimate of the DID in separate time weighted estimators including covariates (estimator 4b and 4c) increased the DID point estimate, exceeding naïve estimates. Ninety-five percent coverage length and location using basic and percentile bootstrap standard errors are similar in prescription count models. The four-group weighted model not including covariates (estimator 5) produced results similar to the two-way fixed effects model (estimator 1). When covariates were included in the four-group weighted model (estimator 6a) the DID point estimate for number of prescriptions increased and coverage was tightened, especially exhibited for the outcome number of prescriptions. Bootstrap methods in the four-group weighted model (estimators 6b and 6c) produced point estimates similar to that of the model-based point estimate (estimator 6a) though with wider confidence intervals. The AIPW DID bootstrap (estimators 7a and 7b) median estimate was consistent with the naïve point estimate for number of prescriptions, and shifted to the left for cost of prescriptions. Coverage from the AIPW was similar to other estimators. The AIPW estimator produced shorter coverage intervals compared to naïve estimates in prescription count models.

Difference-in-differences estimates from the 13 estimators in generalized linear models were appended to the linear model estimates of Table XVI. Forest plots in Figure 13 present
# TABLE XVI

**DIFFERENCE-IN-DIFFERENCES ESTIMATES OF NUMBER AND COST OF PRESCRIPTIONS USING 13 ESTIMATORS**

<table>
<thead>
<tr>
<th>Estimator</th>
<th>Linear</th>
<th>Generalized Linear Model</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Number of prescriptions</td>
<td>Cost of Medicaid prescriptions</td>
</tr>
<tr>
<td></td>
<td>DID</td>
<td>95% CI</td>
</tr>
<tr>
<td>Two-way fixed effect</td>
<td></td>
<td></td>
</tr>
<tr>
<td>(1)</td>
<td>3.82</td>
<td>(0.93, 6.70)</td>
</tr>
<tr>
<td>(2): incl. X</td>
<td>3.99</td>
<td>(1.57, 6.41)</td>
</tr>
<tr>
<td>Separate time weighted</td>
<td></td>
<td></td>
</tr>
<tr>
<td>(3)</td>
<td>3.58</td>
<td>(1.01, 6.14)</td>
</tr>
<tr>
<td>(4a): incl. X</td>
<td>3.72</td>
<td>(1.58, 5.86)</td>
</tr>
<tr>
<td>(4b): incl. X; basic BS SE</td>
<td>3.70</td>
<td>(1.46, 6.16)</td>
</tr>
<tr>
<td>(4c): incl. X; percentile BS SE</td>
<td>3.70</td>
<td>(1.24, 5.94)</td>
</tr>
<tr>
<td>Four-group weighted</td>
<td></td>
<td></td>
</tr>
<tr>
<td>(5)</td>
<td>3.89</td>
<td>(1.03, 6.76)</td>
</tr>
<tr>
<td>(6a): incl. X</td>
<td>4.12</td>
<td>(1.76, 6.49)</td>
</tr>
<tr>
<td>(6b): incl. X; basic BS SE</td>
<td>4.01</td>
<td>(1.13, 7.11)</td>
</tr>
<tr>
<td>(6c): incl. X; percentile BS SE</td>
<td>4.01</td>
<td>(0.09, 6.89)</td>
</tr>
<tr>
<td>AIPWc</td>
<td></td>
<td></td>
</tr>
<tr>
<td>(7a): sandwich SEc</td>
<td>3.92</td>
<td>(1.37, 6.46)</td>
</tr>
<tr>
<td>(7b): basic BS SEc</td>
<td>3.98</td>
<td>(1.41, 6.46)</td>
</tr>
<tr>
<td>(7c): percentile BS SEc</td>
<td>3.98</td>
<td>(1.49, 6.55)</td>
</tr>
</tbody>
</table>

*aPoisson generalized linear model, log link

*bGamma generalized linear model, log link

*cabbreviations: AIPW=augmented inverse probability weighted estimator, BS=bootstrap, SE=standard error
Figure 12. Forest plots comparing difference-in-differences estimators with linear model specification

- The grey vertical lines represent the unadjusted, unweighted DID estimates

- Abbreviations: sep=separate, wts=weights, perc=percentile, bs=bootstrap, sand=sandwich
Figure 13. Forest plots comparing difference-in-differences estimators with generalized linear model specification

*a* the grey vertical lines represent the unadjusted, unweighted DID estimates

*b* abbreviations: sep=separate, wts=weights, perc=percentile, bs=bootstrap, sand=sandwich
estimates (estimator 1) were exactly the same as the unadjusted mean DID estimate presented prior in Table XIV. These values visually. Gray vertical lines were overlaid in forest plots to indicate the adjusted, unweighted mean estimate in Poisson and Gamma models (3.1 medications and $103). All models indicate a significant increase in prescription medications from individuals in Medicaid-expanded states following the expansion, as noted by the absence of DID confidence intervals including zero. Difference-in-differences estimates in Poisson models for number of prescriptions tended to be attenuated compared to those with linear model specification, around three prescriptions in Poisson models compared to four in linear models. This is with the exception of the separate time and four-group weighted models not including covariates (estimators 3 and 5) whose DID estimates were larger than those in similarly specified linear models. Noteworthy from forest plots was the difference in confidence interval length between some of the estimators. Confidence intervals of existing methods using model-derived standard errors (estimators 1, 2, 3, 4a, 5, and 6a) had much smaller confidence interval length compared to the proposed estimators. This is likely due to the overdispersion of the outcome variable. Estimators using model-based standard error estimates in generalized linear models might not be appropriate for making DID inferences in such overdispersed data.

Difference-in-differences of Gamma prescription cost models are presented in Table XVI and presented visually in Figure 13. In general, Gamma cost models also attenuated the DID cost estimates downward. Similar to the linear model specification, results from Gamma cost models conclude a clear, non-significant relationship between the cost of Medicaid prescriptions and Medicaid expansion. Estimates actually changed sign in the two-way fixed effects and separate time weighted models including covariates (estimators 2 and 4a). Three of the 13 models (models 2, 4a, and 4b) did not cover the unweighted, linear mean estimate of $169. Including
covariates to the outcome model typically attenuated the DID estimate and decreased its standard error. Regression coefficient estimates for total charges in linear and inverse Gaussian models for the subgroup are displayed in Table XVII. In the linear model, the adjusted DID of $449 in total charges was similar to the unadjusted estimate of $506 in Table XIV. Regarding the term of interest, the Medicaid-expanded and post indicator, significance and associative relationship were preserved across model specifications. Uninsured individuals were associated with $758 less medical charges compared to those on public only insurance. The majority of relationships between covariates and total charges were congruent between model specifications. Gender and college versus less than high school education were significant in the generalized linear model, while in the linear model they were non-significant. Changing the model specification from linear to inverse Gaussian for the outcome total charges preserved the relationship of the variable of interest, though not for all other covariates.

Linear and inverse Gaussian model DID estimates of total charges from the 13 estimators in the subgroup are displayed in Table XVIII and further visually in Figure 14. Difference-in-differences estimates were generally larger from inverse Gaussian models compared to linear models. In some of the proposed estimators with bootstrap standard errors, the inverse Gaussian model specification actually reversed the significance of the DID estimate. This is an important case where the model and estimator selection played a direct role in the conclusion of an effect. For some estimators, the analyst would conclude Medicaid expansion contributed to an increase in total medical charges, while for other estimators, the analyst would conclude a non-significant effect of Medicaid expansion on total medical charges in this subgroup.

Poisson models of prescription count indicate 2014 Medicaid expansion policy attributed to an increase of three to four yearly prescription medications among low-income individuals in
TABLE XVII

REGRESSION COEFFICIENT ESTIMATES FROM COVARAITE ADJUSTED TWO-WAY FIXED EFFECTS MODELS OF TOTAL CHARGES IN THE SUBGROUP

<table>
<thead>
<tr>
<th>Variable</th>
<th>Total Charges</th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Linear</td>
<td>p-value</td>
<td>Inverse Gaussian</td>
<td>Estimate</td>
<td>p-value</td>
<td>Estimate</td>
<td>p-value</td>
</tr>
<tr>
<td>Medicaid Expanded</td>
<td>-186.6</td>
<td>0.20</td>
<td>-0.132</td>
<td>0.07</td>
<td>-0.132</td>
<td>0.07</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Post</td>
<td>-44.3</td>
<td>0.69</td>
<td>-0.018</td>
<td>0.75</td>
<td>-0.018</td>
<td>0.75</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Medicaid Expanded x Post</td>
<td>448.8</td>
<td>0.03</td>
<td>0.249</td>
<td>0.03</td>
<td>0.249</td>
<td>0.03</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Poverty</td>
<td>0.23</td>
<td>0.84</td>
<td>0.000</td>
<td>0.79</td>
<td>0.000</td>
<td>0.79</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Age</td>
<td>3.96</td>
<td>0.33</td>
<td>-0.001</td>
<td>0.58</td>
<td>-0.001</td>
<td>0.58</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Male</td>
<td>-185.7</td>
<td>0.07</td>
<td>-0.175</td>
<td>&lt;0.001</td>
<td>-0.175</td>
<td>&lt;0.001</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Black</td>
<td>153.1</td>
<td>0.18</td>
<td>0.119</td>
<td>0.06</td>
<td>0.119</td>
<td>0.06</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Hispanic</td>
<td>-66.9</td>
<td>0.59</td>
<td>-0.050</td>
<td>0.44</td>
<td>-0.050</td>
<td>0.44</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Education</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Less than high school</td>
<td>ref</td>
<td>ref</td>
<td>ref</td>
<td>ref</td>
<td>ref</td>
<td>ref</td>
<td></td>
<td></td>
</tr>
<tr>
<td>High school</td>
<td>131.7</td>
<td>0.24</td>
<td>0.080</td>
<td>0.19</td>
<td>0.080</td>
<td>0.19</td>
<td></td>
<td></td>
</tr>
<tr>
<td>College</td>
<td>-112.6</td>
<td>0.35</td>
<td>-0.134</td>
<td>0.03</td>
<td>-0.134</td>
<td>0.03</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Insurance</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Public only</td>
<td>ref</td>
<td>ref</td>
<td>ref</td>
<td>ref</td>
<td>ref</td>
<td>ref</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Uninsured</td>
<td>-757.6</td>
<td>&lt;0.001</td>
<td>-0.363</td>
<td>&lt;0.001</td>
<td>-0.363</td>
<td>&lt;0.001</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Conditions</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0</td>
<td>ref</td>
<td>ref</td>
<td>ref</td>
<td>ref</td>
<td>ref</td>
<td>ref</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>324.8</td>
<td>0.01</td>
<td>0.258</td>
<td>&lt;0.001</td>
<td>0.258</td>
<td>&lt;0.001</td>
<td></td>
<td></td>
</tr>
<tr>
<td>≥2</td>
<td>988.6</td>
<td>&lt;0.001</td>
<td>0.521</td>
<td>&lt;0.001</td>
<td>0.521</td>
<td>&lt;0.001</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
## TABLE XVIII

DID ESTIMATES OF TOTAL CHARGES USING 13 ESTIMATORS IN THE SUBGROUP

<table>
<thead>
<tr>
<th>Estimator</th>
<th>Total Charges(^a)</th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Linear DID 95% CI</td>
<td>Inverse Gaussian DID 95% CI</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Two-way fixed effect</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(1)</td>
<td>506</td>
<td>(88, 924)</td>
<td>515</td>
<td>(7, 1023)</td>
<td></td>
</tr>
<tr>
<td>(2): incl. X</td>
<td>449</td>
<td>(45, 852)</td>
<td>601</td>
<td>(21, 1182)</td>
<td></td>
</tr>
<tr>
<td>Separate time weighted</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(3)</td>
<td>575</td>
<td>(196, 954)</td>
<td>635</td>
<td>(160, 1111)</td>
<td></td>
</tr>
<tr>
<td>(4a): incl. X</td>
<td>590</td>
<td>(226, 953)</td>
<td>960</td>
<td>(369, 1551)</td>
<td></td>
</tr>
<tr>
<td>(4b): incl. X; basic BS SE(^a)</td>
<td>595</td>
<td>(117, 1071)</td>
<td>1073</td>
<td>(-125, 1980)</td>
<td></td>
</tr>
<tr>
<td>(4c): incl. X; percentile BS SE(^a)</td>
<td>595</td>
<td>(119, 1073)</td>
<td>1073</td>
<td>(165, 2270)</td>
<td></td>
</tr>
<tr>
<td>Four-group weighted</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(5)</td>
<td>443</td>
<td>(62, 824)</td>
<td>461</td>
<td>(35, 888)</td>
<td></td>
</tr>
<tr>
<td>(6a): incl. X</td>
<td>455</td>
<td>(90, 819)</td>
<td>619</td>
<td>(120, 1118)</td>
<td></td>
</tr>
<tr>
<td>(6b): incl. X; basic BS SE(^a)</td>
<td>458</td>
<td>(48, 881)</td>
<td>633</td>
<td>(-159, 1275)</td>
<td></td>
</tr>
<tr>
<td>(6c): incl. X; percentile BS SE(^a)</td>
<td>458</td>
<td>(34, 867)</td>
<td>633</td>
<td>(-8, 1426)</td>
<td></td>
</tr>
<tr>
<td>AIPW</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(7a): sandwich SE(^a)</td>
<td>597</td>
<td>(133, 1061)</td>
<td>577</td>
<td>(118, 1036)</td>
<td></td>
</tr>
<tr>
<td>(7b): basic BS SE(^a)</td>
<td>599</td>
<td>(313, 894)</td>
<td>589</td>
<td>(318, 903)</td>
<td></td>
</tr>
<tr>
<td>(7c): percentile BS SE(^a)</td>
<td>599</td>
<td>(305, 886)</td>
<td>589</td>
<td>(276, 860)</td>
<td></td>
</tr>
</tbody>
</table>

\(^a\)abbreviations: AIPW=augmented inverse probability weighted estimator, BS=bootstrap, SE=standard error
Figure 14. Forest plots comparing estimators for total charges

1. Naive
2. Naive with X
3. Sep time wts
4a. Sep time wts+X
4b. Sep time wts+X. basic bs
4c. Sep time wts+X. perc bs
5. 4-group wts
6a. 4-group wts+X
6b. 4-group wts+X. basic bs
6c. 4-group wts+X. perc bs
7a. AIPW. sand
7b. AIPW. basic bs
7c. AIPW. perc bs

---

\(^a\) the grey vertical lines represent the unadjusted, unweighted DID estimates

\(^b\) abbreviations: sep=separate, wts=weights, perc=percentile, bs=bootstrap, sand=sandwich
Medicaid-expanded states. Results further indicate no difference to the total cost of Medicaid prescriptions attributed to the expansion. The effect of Medicaid expansion on total medical charges among individuals with non-private insurance and some, limited (<$10,000) charges remains inconclusive. Medicaid expansion was attributed with an additional three to four yearly prescription medications for individuals in Medicaid-expanded states while not significantly influencing prescription cost.

4.4 Discussion

This study had several limitations. Northeast and South census regions were not comprised entirely of expanded and nonexpanded states, respectively. As such, derived estimates here not suitable for policy decisions. Furthermore, contained results are not generalizable for the entire United States, rather only Northeast and South census regions. The study used the year of Medicaid expansion (2014) for the post period, as the 2015 MEPS cost data was not yet available. It might take some time for utilization trends and Medicaid drug spending to stabilize following the introduction of the new policy. Future research could confirm whether the trends observed here are a product of the actual policy or transitory shock following the policy introduction.

Priority conditions were self-reported as opposed to obtained through medical records. There is a possibility this method did not capture all diagnoses or concluded diagnoses that were not actually present. For example, an individual might not have wanted to reveal a specific diagnosis or refused to answer a question when a diagnosis was present. Such individuals were subsequently excluded from the study. Conversely, self-diagnoses might have been over-reported. In a study examining agreement between self-reported medical conditions and electronic medical records, the prevalence of comorbidities was found to be lower in the medical
records (Smith et al., 2008). Explanations include “poor patient-clinician communication, self-diagnosis in the absence of a satisfactory explanation for symptoms, or the ‘health literacy’ of the patient” (Smith et al., 2008). Agreement between self-reported diagnoses and electronic medical records identified diagnoses increased with increased surveillance time. The MEPS two year surveillance period may have contributed to enhanced data integrity of self-reported outcomes.

Our results indicate that state implementation of Medicaid expansion increased drug prescribing while not affecting Medicaid drug spending. A study from Wen et al. (2016) found similar conclusions. Potential reasons the authors cite to explain these findings are as follows. First, according to a budget survey of Medicaid officials in 2014 and 2015 by the Henry J. Kaiser Family Foundation and Health Management Associates, expanded states were more than twice likely to implement pharmacy management initiatives compared to nonexpanded states. Examples of these initiatives include prior authorization programs, incentives to use generic drugs, preferred drug lists, pharmacy benefit carve outs, and reduced reimbursements for certain drugs (Smith et al., 2014). These programs were aimed at reducing the cost of prescription medications. Second, several states were pushing to increase enrollment in risk-based managed care programs. Examples of this include a compulsory enrollment in managed care for new enrollees, expanding compulsory managed care to additional groups, and establishing new managed care programs. Lastly, Medicaid-expanded states already had a large share of managed care market penetration. This could have helped soften the impact of additional enrollees to similar programs.

A change of four medications per individual, per year after switching to an insurance type with a different medication access structure was found in previous studies. Stroupe and colleagues (2017) found that Veterans acquired four less medications from the VA system after
enrolling in Part D Medicare, an alternate source of medication acquisition. Khan and colleagues (2018) found that prescription assistance programs were associated with three to four medications per person. An effective healthcare policy involving new access routes to medication acquisition might expect to see a policy usage of three to four medications per person, per year.

Medicaid expansion is the hot topic in current ACA reform. Lawmakers are trying to determine a way to roll back Medicaid expansion without severely disrupting access to care. Current studies such as ours, on the impact of Medicaid expansion on pharmacy utilization and cost are timely and critical. Model specification and estimator selection play a crucial role in terms of location, precision, and associative attributes of the DID effect estimate. Careful consideration need be taken into DID estimation procedures to avoid incorrect conclusions, as such conclusions could potentially have a broad impact, including on the United States’ healthcare system.
5. SIMULATION STUDIES

To assess the performance of DID estimators under various scenarios of selection bias, two simulation studies were conducted. The first is a motivating simulation study using artificial data under small to moderate size covariate confounding. The second is a comprehensive simulation study devised around the real data and trends of the MEPS dataset. True DID effects were either specified for the simulation with artificial data, or computed for the simulation devised around real data. Details regarding data generation, simulation algorithms, estimator evaluation, and results from the two simulation studies are given in sections 5.1 and 5.2. Section 5.3 presents a discussion of the overall estimator performance across both simulation studies. A code supplement is provided in the Appendix.

5.1 Simulation using artificial data

A simple motivating simulation was conducted using artificial, idealized data. We followed the simulation parameters and DID selectivity scenarios akin to Stuart et al. (2014) in effort to directly compare results between the proposed and existing methods. The simulated datasets consisted of 500 observations in each of the four groups (treatment pre, treatment post, control pre, control post) for a total of 2,000 total observations per dataset.

A normally distributed covariate $X$ was used to introduce covariate confounding. Settings of covariate confounding represented different kinds of selection bias likely to occur using observational data. The four scenarios of covariate confounding were: no difference, treatment groups differ at baseline but do not change over time, treatment groups do not differ at baseline but change over time, and treatment groups both differ at baseline and change over time. The following model was used to introduce selection bias in the data generation process,

$$Y = \alpha + \beta_P P + \beta_Z Z + \beta_X X + \beta_{PZ} PZ + \beta_{PX} PX + \beta_{ZX} ZX + \beta_{PZX} PZX + u$$
Confounding effect sizes of 0.2-0.3 SD, which reflect small to moderate size covariate imbalance were used (Stuart, 2010). These simulation settings are displayed in Table XIX.

**TABLE XIX**

SIMULATION SETTINGS OF SELECTION BIAS

<table>
<thead>
<tr>
<th>Setting</th>
<th>How groups differ in the covariate X</th>
<th>Covariate difference at baseline</th>
<th>Change in X in control group</th>
<th>Change in X in treatment group</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>No difference</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>2</td>
<td>Group difference</td>
<td>0.3</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>3</td>
<td>Group and time difference</td>
<td>0.3</td>
<td>0.2</td>
<td>0.2</td>
</tr>
<tr>
<td>4</td>
<td>Group by time difference</td>
<td>0.3</td>
<td>0.1</td>
<td>-0.2</td>
</tr>
</tbody>
</table>

5.1.1 **Algorithm**

The following algorithm was implemented for the simulation study using artificial data:

1. Generate a normally distributed covariate X for each of the four groups including any desired selection bias, \( X \sim N(0 + \text{selection bias}, 1) \)

2. Generate the outcome data via the model,

\[
Y = \alpha + \beta_P P + \beta_Z Z + \beta_X X + \beta_P PZ + \beta_P PX + \beta_Z ZX + \beta_P ZX + \beta_X PZ + u
\]

\[u \sim N(0,1)\]

where \( \beta_P = 0.1, \beta_Z = 0.3, \beta_X = 0.2, \beta_PZ = 0.1, \beta_PX = 0.2, \beta_ZX = 0.1, \beta_PZX = 0.15 \)

3. Calculate the true DID effect, \( \text{DID} = \text{mean}(\beta_PZ + \beta_PZX X_{PZ}) \) among the post treated observations
4. Estimate propensity score model parameters and compute individual weights

5. Estimate the DID using the 13 estimators

6. Record the accuracy and precision of DID estimates

7. Repeat process 5,000 times

8. Repeat steps 1-7 for the four settings in covariate confounding

9. Calculate well-accepted measures of accuracy and precision

Five thousand datasets under each of the four scenarios in covariate confounding scenarios were created for a total of 20,000 generated datasets. Each estimator was applied to the same 20,000 datasets. Five hundred bootstrap replicates were drawn to compute confidence intervals. Estimators were compared by established, desirable properties of unbiasedness, consistency, and efficiency (Amemiya, 1985). R software version 3.4.2 was used for the simulations using artificial data. Augmented inverse probability weighted estimates were calculated from software code adapted from function `estimate.ATE` in package CausalGAM (Glynn and Quinn, 2010).

5.1.2 Evaluation criteria

Performance of statistical estimators was evaluated under three well-accepted accuracy and precision measures: mean bias, RMSE, and coverage rate.

The mean bias is the mean of all distances between the parameter estimate and its actual value. For a parameter of interest, $\theta$,

$$\text{mean bias}(\hat{\theta}) = E_{\theta}(\hat{\theta} - \theta) = \frac{1}{n} \sum_{i=1}^{n}(\hat{\theta}_i - \theta).$$

This measure has also been called the mean error, mean deviation, mean difference, or bias (Walther and Moore, 2005). It indicates whether an estimator consistently over- or underestimates the parameter of interest (Walther and Moore, 2005). A mean bias close to zero is
desirable.

Root-mean-square error is the square root of the combined variance and square bias. It is formally defined as,

$$\text{RMSE}(\hat{\theta}) = \sqrt{E_\theta[(\hat{\theta} - \theta)^2]} = \sqrt{\frac{1}{n} \sum_{i=1}^{n} (\hat{\theta} - \theta)^2}.$$ 

The RMSE is likely the optimal criterion for evaluating the combined accuracy and precision (Demirtas and Yavuz, 2015). To our knowledge there is no guideline for an acceptable level of RMSE. A small RMSE is desirable.

The coverage rate is the percentage of times the true parameter value lies inside the confidence interval. Ideally, the actual coverage should be close to the nominal rate, so Type I and II errors are properly represented. It is important not to evaluate the coverage rate alone, as a high variance can lead to high coverage rate. Coverage was considered poor if its rate dropped below 90% (Demirtas and Yavuz, 2015). Coverage was considered good if the rate fell between two standard errors of the nominal coverage probability \( p \), \( \text{SE}(p) = \sqrt{p(1-p)/B} \) for \( B \) simulations (Tang et al., 2005). In this investigation using 5,000 simulated datasets, a coverage rate between 0.944-0.956 was considered good, 0.900-0.943 or 0.957-0.999 was considered fair, and less than 0.900 or greater than 0.990 was considered poor.

Bias is a measure of accuracy and RMSE and coverage rate are hybrid measures of accuracy and precision (Demirtas, 2007b). Since this study was focused on standard error estimation procedures, RMSE and coverage were given more importance in the evaluation of estimators. While a desirable estimator will have low RMSE and a coverage rate close to the nominal value, it needs to also exhibit acceptable bias. Good precision is worthless without sufficient accuracy (Demirtas, 2007b).
5.1.3 Results

Simulation results of the 13 estimators in 5,000 simulated datasets across each of the four settings are presented in Table XX. Coverage results are presented visually in Figure 15. A black vertical line on coverage plots represent the 0.95 nominal level, and results shaded in red represent estimators proposed in this dissertation.

All estimator specifications perform well in the case of no covariate confounding. All covariate adjusted estimators display good accuracy and precision and would be a great choice in the case of no covariate confounding. When confounding is introduced, bias and RMSE become inflated in naïve estimators, and coverage deteriorates.

Covariate adjustment resulted in reduced variance, a slight reduction in RMSE, and similar bias. The reduction in variance lead to a downward shift in coverage rates. Regression adjustment contributed information towards variance explanation of the outcome which translated to greater confidence (smaller standard errors) in the produced treatment effect estimate. The distribution of the covariate $X$ was dependent on the invented confounding in the treatment group and time. While the covariate $X$ was related to the treatment effect estimate of interest, $PZ$, the greater strength in the association between $X$ and the outcome $Y$ was enough to make covariate adjustment worthwhile in the respect that it did not inflate bias estimates. Despite the reduction to standard errors from regression adjustment, and coverage improvement in some settings, it was not enough to achieve good coverage in all settings. In separate time weighted models, regression adjustment improved results in all settings except for the group-only confounded setting. In four-group weighted models, regression adjustment improved the no bias Medicaid-expanded states. Results further indicate no difference to the total cost of Medicaid and group-only bias settings, but worsened coverage rates in the group and time and group by time
### TABLE XX

**SIMULATION RESULTS IN A STUDY USING ARTIFICIAL DATA**

<table>
<thead>
<tr>
<th>Setting</th>
<th>Two-way FE&lt;sup&gt;b&lt;/sup&gt;</th>
<th>Separate time weighted</th>
<th>Four-group weighted</th>
<th>AIPW&lt;sup&gt;b&lt;/sup&gt;</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>(1)</td>
<td>(2): incl. X</td>
<td>(3)</td>
<td>(4a): incl. X</td>
</tr>
<tr>
<td>Mean Bias</td>
<td>0.0026</td>
<td>0.0027</td>
<td>0.0025</td>
<td>0.0026</td>
</tr>
<tr>
<td>2</td>
<td>0.0595</td>
<td>0.0587</td>
<td>-0.0237</td>
<td>-0.0236</td>
</tr>
<tr>
<td>3</td>
<td>0.0828</td>
<td>0.0822</td>
<td>-0.0004</td>
<td>-0.0004</td>
</tr>
<tr>
<td>4</td>
<td>-0.0782</td>
<td>0.0377</td>
<td>-0.0037</td>
<td>-0.0036</td>
</tr>
<tr>
<td>RMSE</td>
<td>0.0944</td>
<td>0.0899</td>
<td>0.0897</td>
<td>0.0897</td>
</tr>
<tr>
<td>2</td>
<td>0.1118</td>
<td>0.1067</td>
<td>0.0932</td>
<td>0.0931</td>
</tr>
<tr>
<td>3</td>
<td>0.1257</td>
<td>0.1215</td>
<td>0.0900</td>
<td>0.0899</td>
</tr>
<tr>
<td>4</td>
<td>0.1236</td>
<td>0.0981</td>
<td>0.0905</td>
<td>0.0905</td>
</tr>
<tr>
<td>95% Confidence Interval Coverage</td>
<td>0.958</td>
<td>0.953</td>
<td>0.967</td>
<td>0.954</td>
</tr>
<tr>
<td>2</td>
<td>0.910</td>
<td>0.906</td>
<td>0.958</td>
<td>0.943</td>
</tr>
<tr>
<td>3</td>
<td>0.870</td>
<td>0.850</td>
<td>0.967</td>
<td>0.954</td>
</tr>
<tr>
<td>4</td>
<td>0.875</td>
<td>0.933</td>
<td>0.964</td>
<td>0.949</td>
</tr>
</tbody>
</table>

<sup>a</sup>areas shaded in grey represent the proposed methods in this dissertation

<sup>b</sup>abbreviations: FE=fixed effects, AIPW=augmented inverse probability weighted estimator, BS=bootstrap, SE=standard error, perc=percentile, sand=sandwich
Figure 15. Coverage rates in a study using artificial data

1. Naive
2. Naive with X
3. Sep time wts
4a. Sep time wts+X
4b. Sep time wts+X. basic bs
4c. Sep time wts+X. perc bs
5. 4-grp wts
6a. 4-grp wts+X
6b. 4-grp wts+X. basic bs
6c. 4-grp wts+X. perc bs
7a. AIPW. sand
7b. AIPW. bs basic
7c. AIPW. bs perc

Abbreviations: sep=separate, wts=weights, grp=group, perc=percentile, bs=bootstrap, sand=sandwich
bias settings. It is clear that regression adjustment alone in this artificial simulation was not enough to generate good coverage for all settings.

Separate time weighted estimators had excellent accuracy. Examining the effect of the separate time weights (comparing models 1 versus 3, and 2 versus 4a) we see a reduction in bias and RMSE with the separate time weights. There is an especially great reduction to the bias in the group and time setting as well as the group by time setting, from 0.822 to -0.0004 and 0.0377 to -0.0036 in covariate adjusted models. The separate time weighted estimator which includes the covariate $X$ in the outcome model greatly improved confidence interval coverage by reducing it closer to the 95% nominal rate. This improvement to the coverage can be seen across all settings except in the group-only confounded setting. This is due to the fact that the separate time weighted estimator makes a great effort to balance groups within each time period, but contributes less effort to balancing groups across time periods. Since this estimator contributes the most energy to balancing time periods, it performs especially well in time-confounded settings. The estimator performs well in terms of bias in the group and time confounded setting (setting 3), as well as in the group by time confounded setting (setting 4). It makes sense that it performs better in the setting where confounding was introduced separately in group and time, instead of combined in a group by time setting. Overall, the separate time weighted model was an improvement to the two-way fixed effects model in DID estimation, especially in settings of a time confounding.

The four-group weighted design exhibits greater bias and RMSE compared to the separate time weighted estimator (estimators 5 versus 3, and 6a versus 4a) across all settings except in a group-only bias (setting 2). The four-group weighted design does a fantastic job at reducing bias in the group-only bias setting. This is likely because the four-group weighted design goes to
great lengths to put both groups at any point in time, on level playing field, by balancing the
groups in terms of treatment probabilities relative to the same reference group. This is similar to
what was observed earlier in the separate time weighted estimator providing improved mean bias
in the presence of a time bias. Confidence interval coverage in settings of a time bias (settings 3
and 4) were good in the model not including the covariate $X$. However, in the four-group
weighted model including the covariate $X$, a likely specification choice of a well-informed
analyst, only fair coverage was achieved (0.933-0.935). The regression adjusted four-group
weighted model exhibits excellent coverage in the no bias setting and group-only bias setting.
The excellent accuracy in these settings help place the confidence interval at a good location to
obtain fantastic coverage rates.

Bootstrap standard errors maintain the gains made in accuracy of the weighted estimators,
while refining coverage rates to excellent levels. These estimators represent the proposed
procedure of this dissertation of using bootstrap median and standard error estimates instead of
the usual model-based estimates. Bootstrap estimates exhibit similarly excellent accuracy
compared to their model-based counterparts (estimators 4b versus 3, 4c versus 4a, 6b versus 5,
6c versus 6a). The real benefit of this approach is what bootstrap is known for: a precise
confidence interval estimate. Coverage rates in bootstrap separate time weighted models are
excellent across all settings of covariate confounding. This separate time weighted estimator with
bootstrap standard errors is the only estimator to produce excellent coverage (0.947-0.956)
across all settings of covariate confounding. The four-group weighted estimator exhibits
excellent coverage in the no bias and group-only bias settings (0.949-0.953), and slight
undercoverage in the presence of a time bias (0.936-0.941). In this simulation using artificially
generated data, bootstrap standard errors offer a slight refinement to model-based estimates.
Basic and percentile bootstrap confidence intervals exhibit similar performance in both separate time weighted and four-group weighted models. Data generated in this experiment was sampled from a normal distribution. The desired DID statistic in this study was also likely to be more or less symmetric and normally distributed. Since the percentile bootstrap coverage rate degrades as the distribution of its estimates move away from symmetry, this could offer an explanation as to the similarity between basic and percentile bootstrap methods in this particular simulation using normally simulated data.

The AIPW had excellent mean bias and RMSE. Similar to the separate time weighted estimator, the mean bias was somewhat inflated in the group-only confounded setting, though still acceptable. This is likely due to the similarity between the AIPW and the separate time weighted estimator with their focus on balance within time period, rather than group. Confidence interval coverage of the AIPW was excellent in three-quarters of the settings (0.944-0.953), and exhibited fair coverage (0.938-0.942) in the remainder of settings. The AIPW estimator holds promise in combining good accuracy and precision.

Confidence interval coverage rates were greatly improved in weighted estimators versus naïve two-way fixed effects estimators by accounting for the selection bias of group membership and time. Propensity score weighting techniques were successful in balancing treatment groups and time periods. The separate time weighted estimator performed superior to the four-group weighted estimator in terms of mean bias and RMSE in the presence of a time bias. However, in the presence of a group-only bias, the four-group weighted estimator performed superior to the separate time weighted estimator in terms of mean bias and RMSE. Likewise, confidence interval coverage degraded for the four-group weighted estimator in the presence of a time selection bias. Mean bias, RMSE, and 95% confidence interval coverage were excellent in
separate time weighted estimators with bootstrap standard errors. Weighed models with bootstrap standard errors resulted in an improved precision compared to model-based standard errors in the same model. The proposed empirical bootstrap standard error in separate time weighted and four-group weighted estimators resulted in a major improvement to the precision over model-based standard error estimates. The separate time weighted estimator with bootstrap standard errors and the AIPW estimator consistently performed superior to other estimators in all types of selection bias.

5.2 Simulation devised around real data

A simulation study was conducted with data generated similar to a real data example. The MEPS dataset provided structural guidance to the simulated datasets. The goal was to emulate the MEPS’s real distributional, marginal, and associative characteristics. Simulated datasets which resemble reality will be more generalizable and credible (Burton et al., 2006). Covariate summary measures and statistical testing were performed on the simulated datasets to ensure simulated datasets resembled the actual MEPS data (Burton et al., 2006). Estimator performance was evaluated under various types and magnitudes of covariate confounding plausible to be present in the MEPS dataset. Methods which consistently perform superior in these plausible scenarios will be more trustworthy and robust in comparison to those which only perform well in a few specific situations.

5.2.1 Data generation

Variables included in the MEPS dataset consisted of two outcomes measures and eight covariate predictors. The two outcome variables consisted of one count (number of prescriptions) and one right-skewed continuous (total healthcare charges) variable. The eight predictor covariates consisted of two continuous (poverty, age), three dichotomous (gender, Black,
Hispanic), and three ordinal (education, insurance, conditions) variables. Simulated datasets were created using the same sample size as the MEPS dataset, including the sample sizes of groups (treatment pre: \( n=916 \), treatment post: \( n=871 \), control pre: \( n=2,636 \), control post: \( n=2,407 \)).

In the generation of the eight covariates, the actual MEPS dataset was first divided into four groups (treatment pre, treatment post, control pre, control post). Marginal distributions and correlations were determined separately within each of the four groups (Demirtas, 2007a; Demirtas and Hedeker, 2011; Demirtas, 2014; Demirtas and Hedeker, 2016). The eight covariates were then generated simultaneously in each of the four groups. The four simulated datasets were appended together for the final dataset of predictor covariates which included post and treatment group indicators representing the four appended groups. Power polynomials were used to generate the nonnormal, continuous variables poverty and age (Demirtas et al., 2012). R package PoisBinOrdNonNor was used for multivariate data generation (Demirtas et al., 2017). Random seeds were used to construct each of the simulated datasets to ensure results were reproducible.

Outcome measures number of prescription medications (\( RXTOT \)) and total charges (\( TOTTCCH \)) were the two outcomes examined in the simulation study. We did not include total cost of Medicaid prescriptions (\( RXMCD \)) as there was not a significant DID treatment effect and any comparison of estimators under this outcome would be useless. It was important to include a continuous outcome for which the treatment effect of the variable of interest was significant to represent liberal reflections of the dangers in covariate confounding. Failing to adjust for the estimation of the propensity scores will have the greatest consequences with a continuous outcome (Williamson et al., 2014). This is attributed to the greatest variance reduction for a continuous outcome.
Generation of outcome measures was performed in such a way to accommodate the actual distribution of the outcome variables. Below are the algorithms used in the data generation process of outcome measures:

1. Construct a generalized linear regression model for each outcome using the original MEPS dataset and retain coefficient estimates

\[ R_{XTOT} \sim \text{Poisson}(\mu) \]

\[ TOTTCH \sim \text{inverse Gaussian}(\mu, \lambda) \]

\[
g(\mu, \lambda) = \beta_0 + \beta_P P + \beta_Z Z + \beta_PZ PZ + \beta_{Poverty} Poverty + \beta_{Age} Age + \beta_{Male} Male
\]

\[ + \beta_{Black} Black + \beta_{Hispanic} Hispanic + \beta_{Education} Education \]

\[ + \beta_{Insurance} Insurance + \beta_{Conditions} Conditions \]

2. Generate the 8 covariates

3. Predict the outcome on the simulated set of covariates using the regression relationship identified in step 1

4. Round down generated outcomes to the nearest integer

### 5.2.2 Design elements

Estimator performance was evaluated across several scenarios of continuous covariate confounding plausible to occur in observational data. A two (confounded covariates) by four (types of covariate confounding) by three (magnitudes of covariate confounding) factorial design was used for both outcome measures, \( R_{XTOT} \) and \( TOTTCH \). Covariate confounding was created \textit{after} the generation of simulated datasets.

Two continuous variables were used to create four settings in covariate confounding: no bias, group-only bias, group and time bias, and group by time bias. Three effect sizes in covariate confounding of small, medium, and large were incorporated into the settings of covariate
confounding. Three effect sizes in covariate confounding of 0.1-0.2, 0.35-0.5, and 0.25-0.8 were considered as small, medium, and large, respectively (Cohen, 1988). Cohen’s $d$ was used as a metric for effect size. Table XXI shows the desired covariate confounding effect sizes in terms of Cohen’s $d$ for each setting of selection bias, across the three levels of effect size. The poverty variable took continuous values between 0 and 138, where lower values correspond to greater poverty. The age variable in the final MEPS dataset took integer values of 18-64. To create the confounded poverty and age variables, new variables were calculated using the desired effect size multiplier then rounded down to the nearest integer in the case of age, and truncated to match the distributional assumptions of the original MEPS dataset. However, the rounding and truncating procedures of the newly created confounded variables attenuated the calculated effect size, $d$, below desired levels. To solve this, effect size multiplier tuning parameters were found via trial and error by increasing in increments of 0.1 from the desired effect size multiplier until the actual effect size difference after rounding and truncating was met or exceeded. Table XXII shows the final effect size multiplier tuning parameters that were used in the creation of confounded poverty and age variables.

The goal of this simulation study was to examine the performance of standard error estimation procedures of the DID in plausible scenarios of covariate confounding. Similar to DID estimation procedures in the application to the MEPS dataset, the marginal effect was calculated as the discrete-difference of the interaction term between post and treatment group indicators and its delta method derived standard error. Software function $\text{marg}$ in package modmarg (Gold et al., 2017) was used to calculate the marginal effect and associated delta method derived standard error.

There is the underlying truth in the MEPS dataset, which we will never know, but we
<table>
<thead>
<tr>
<th>Setting</th>
<th>How groups differ in the covariate $X$</th>
<th>Covariate difference at baseline</th>
<th>Change in $X$ in control group</th>
<th>Change in $X$ in treatment group</th>
</tr>
</thead>
<tbody>
<tr>
<td>Small effect size, $d=0.1-0.2$</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>No difference</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>2</td>
<td>Group difference</td>
<td>0.2</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>3</td>
<td>Group and time difference</td>
<td>0.2</td>
<td>0.15</td>
<td>0.15</td>
</tr>
<tr>
<td>4</td>
<td>Group by time difference</td>
<td>0.2</td>
<td>0.1</td>
<td>-0.15</td>
</tr>
<tr>
<td>Medium effect size, $d=0.35-0.5$</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>Group difference</td>
<td>0.5</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>3</td>
<td>Group and time difference</td>
<td>0.5</td>
<td>0.35</td>
<td>0.35</td>
</tr>
<tr>
<td>4</td>
<td>Group by time difference</td>
<td>0.5</td>
<td>0.15</td>
<td>-0.35</td>
</tr>
<tr>
<td>Large effect size, $d=0.25-0.8$</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>Group difference</td>
<td>0.8</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>3</td>
<td>Group and time difference</td>
<td>0.8</td>
<td>0.55</td>
<td>0.55</td>
</tr>
<tr>
<td>4</td>
<td>Group by time difference</td>
<td>0.8</td>
<td>0.25</td>
<td>-0.55</td>
</tr>
</tbody>
</table>
TABLE XXII

SIMULATION SETTINGS OF ACTUAL EFFECT SIZE MULTIPLIERS

<table>
<thead>
<tr>
<th>Setting</th>
<th>Age</th>
<th></th>
<th></th>
<th></th>
<th>Poverty</th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Group 1</td>
<td>Group 2</td>
<td>Group 3</td>
<td>Group 4</td>
<td>Group 1</td>
<td>Group 2</td>
<td>Group 3</td>
</tr>
<tr>
<td>Small effect size, $d=0.1-0.2$</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1 No difference</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>2 Group difference</td>
<td>0.26</td>
<td>0.26</td>
<td>0</td>
<td>0</td>
<td>0.21</td>
<td>0.21</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>3 Group and time difference</td>
<td>0.26</td>
<td>0.43</td>
<td>0</td>
<td>0.18</td>
<td>0.21</td>
<td>0.37</td>
<td>0</td>
<td>0.17</td>
</tr>
<tr>
<td>4 Group by time difference</td>
<td>0.24</td>
<td>0.09</td>
<td>0</td>
<td>0.18</td>
<td>0.21</td>
<td>0.06</td>
<td>0</td>
<td>0.18</td>
</tr>
<tr>
<td>Medium effect size, $d=0.35-0.5$</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2 Group difference</td>
<td>0.55</td>
<td>0.55</td>
<td>0</td>
<td>0</td>
<td>0.52</td>
<td>0.52</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>3 Group and time difference</td>
<td>0.55</td>
<td>0.95</td>
<td>0</td>
<td>0.44</td>
<td>0.52</td>
<td>0.90</td>
<td>0</td>
<td>0.38</td>
</tr>
<tr>
<td>4 Group by time difference</td>
<td>0.55</td>
<td>0.20</td>
<td>0</td>
<td>0.18</td>
<td>0.52</td>
<td>0.17</td>
<td>0</td>
<td>0.16</td>
</tr>
<tr>
<td>Large effect size, $d=0.25-0.8$</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2 Group difference</td>
<td>0.87</td>
<td>0.86</td>
<td>0</td>
<td>0</td>
<td>0.83</td>
<td>0.83</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>3 Group and time difference</td>
<td>0.87</td>
<td>1.45</td>
<td>0</td>
<td>0.61</td>
<td>0.83</td>
<td>1.40</td>
<td>0</td>
<td>0.58</td>
</tr>
<tr>
<td>4 Group by time difference</td>
<td>0.87</td>
<td>0.31</td>
<td>0</td>
<td>0.28</td>
<td>0.83</td>
<td>0.28</td>
<td>0</td>
<td>0.27</td>
</tr>
</tbody>
</table>
were able to obtain an estimate of the truth which represents the overall effect of simulated trends. A huge dataset, ten times the size of the original MEPS dataset was generated with similar marginal and associational trends to the MEPS dataset. The truth was calculated as the discrete-difference marginal effect including any introduced covariate confounding in this huge dataset. This value served as the benchmark to gauge estimator performance.

The discrete-difference marginal effect was obtained as follows. In the situation where we wish to introduce confounding through the covariate age, a regression model with linear component,

$$X \beta = \beta_0 + \beta_P P + \beta_Z Z + \beta_PZPZ + \beta_{Age} Age + \beta_{PAge} PAge + \beta_{ZAge} ZAge + \beta_{PZAge} PZAge$$

$$+ \beta_{Poverty} Poverty + \beta_{Male} Male + \beta_{Black} Black + \beta_{Hispanic} Hispanic$$

$$+ \beta_{Education} Education + \beta_{Insurance} Insurance + \beta_{Comorbidity} Conditions$$

was first evaluated. Poisson and Gamma generalized linear models were used for outcomes $RXTOT$ and $RXMCD$, respectively. Discrete-differences of observations in group 2 (i.e. treated post) were calculated as,

$$\Delta E[Y|X]/\Delta PZ = \exp(X \beta) - \exp(\beta_PZ - \beta_{PZAge} Age)$$

as well as discrete-differences of observations not in group 2,

$$\Delta E[Y|X]/\Delta PZ = \exp(X \beta + \beta_PZ + \beta_{PZAge} Age) - \exp(X \beta).$$

These calculated discrete-differences were then averaged across all observations for the true DID marginal effect (Hilbe, 2011). Discrete-difference marginal effects from estimators were compared against the true DID effect estimate to determine estimator performance.

Mean bias, RMSE, and coverage rate were used as metrics for accuracy and precision to evaluate the performance of the estimators. True values were calculated under the large simulated dataset, ten times the size of the original MEPS dataset (original: $n=6,930$; big:
In total, there were 40 combinations of simulation specifications (two confounded variables, by three levels of confounding, under three effect sizes, plus the no confounding setting, across two outcomes). For each scenario, 1,000 simulated datasets were generated and the 13 estimators were applied to each of the simulated datasets. Two hundred bootstrap replications were used in each bootstrap estimate of the standard error (estimators 4b, 4c, 6b, 6c, 7b, 7c). The same evaluation criterion was applied in the simulation study using artificial data. For 1,000 simulated datasets and a 95% nominal coverage rate, acceptable coverages from an estimator would be between 0.936-0.964. A coverage rate between 0.936-0.964 was considered good, 0.900-0.935 or 0.965-0.990 was considered fair, and less than 0.90 or greater than 0.99 was considered poor.

A unique seed was used in the generation of each dataset and each bootstrap resampling procedure to ensure datasets were independent and results, reproducible. Software package glm2, which uses a modified default fitting algorithm for greater model stability, was used to estimate the generalized linear models (Marschner, 2017).

5.2.3 **Algorithm**

The following algorithm was used in the simulation study devised around the MEPS data:

1. Generate a big simulated dataset of $RXTOT$ outcome and the eight predictor variables, ten times the size of the original MEPS dataset, $n=69,300$.
2. Create the newly confounded covariate, $X'$, where $d$ is the specified effect size multiplier tuning parameter and $\sigma_X$ is the SD of $X$,

$$X' = (d \cdot \sigma_X) + X$$

3. Round and truncate the newly created confounded variable $X'$ to match the distributional assumptions of the MEPS data.
4. Construct a Poisson model using the big simulated dataset \((n=69,300)\) including interactions between post and treatment and the confounded variable of interest, \(X'\),

\[ RXTOT \sim \text{Poisson}(\mu) \]

\[
g(\mu) = \beta_0 + \beta_P P + \beta_Z Z + \beta_{PZ} PZ + \beta_{X'} X' + \beta_{PX'} PX' + \beta_{ZX'} ZX' + \beta_{PZX'} PZX' \\
+ \beta_{\text{Poverty}} \text{Poverty} + \beta_{\text{Male}} \text{Male} + \beta_{\text{Black}} \text{Black} + \beta_{\text{Hispanic}} \text{Hispanic} \\
+ \beta_{\text{Education}} \text{Education} + \beta_{\text{Insurance}} \text{Insurance} \\
+ \beta_{\text{Conditions}} \text{Conditions} \]

5. Calculate the true DID effect

6. Generate a simulated dataset of \(RXTOT\) outcome and eight predictor variables the same size as the original MEPS dataset, \(n=69,30\)

7. Estimate the propensity score model parameters and compute weights

8. Estimate the DID and associated standard error using the 13 estimators

9. Record accuracy and precision measures of DID estimates, comparing the true DID estimate from the big dataset with the finite-difference marginal DID effect estimate from the generated dataset

10. Repeat steps 6-9, 9,999 times

11. Calculate well-accepted accuracy and precision measures

12. Repeat steps 1-11 for each simulation scenario

Simulation experiments for the subgroup were conducted in a similar manner. The subgroup was first formed from the original data, and procedures were carried out the same as above. The only difference in simulations involving the subgroup was the use of an inverse Gaussian generalized linear model in place of Poisson and size of generated datasets \((n=2,592)\). Software R version 3.4.2 was used for all data generation and simulations.
5.2.4 Examination of simulated datasets

Descriptive and graphical tools were used to demonstrate the proximity of simulated datasets to the MEPS dataset. Formal statistical testing was used to test for distributional differences, correlational comparisons were made through visuals, and trends from regression models were compared between the actual and simulated MEPS datasets.

Simulated datasets were examined to determine the similarity between the actual MEPS dataset and simulated datasets. Distributions of variables between the actual and simulated datasets exhibit good concordance. Table XXIII presents a comparison of distributional characteristics between the actual MEPS dataset \(n=6,830\) and the big simulated dataset \(n=68,300\). There did not exist any significant differences in predictor variables between the actual MEPS dataset and the big simulated MEPS dataset (all \(p \geq 0.51\), except for age \(p=0.02\)).

There was a significant difference between the outcome \(RXTOT\) in the actual and simulated datasets. This was not surprising. While we did our best to devise a data generation procedure similar to the true underlying mechanism, the erratic behavior of the total prescriptions outcome made it difficult. This distribution was heavily right-skewed, with several individuals acquiring no medications.

Variable correlations in the actual and big simulated MEPS datasets are presented visually in Figure 16. The size of the circle represents the magnitude of the association and the color represents the direction. Blue indicates a positive correlation between variables and red indicates a negative correlation. Visual representations of these correlations are further stratified by group and displayed in Figures 17-20. It can be seen that the correlational relationships between variables were retained in the data generation procedures.

Regression estimates comparing the actual MEPS datasets, the big simulated dataset, and
TABLE XXIII

COMPARISON OF THE ACTUAL AND BIG SIMULATED DATASETS

<table>
<thead>
<tr>
<th>Variable</th>
<th>Actual (n=6,830)</th>
<th>Simulated (n=68,300)</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of Prescriptions</td>
<td></td>
<td></td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>Mean</td>
<td>12.3</td>
<td>10.1</td>
<td></td>
</tr>
<tr>
<td>SD</td>
<td>26.9</td>
<td>14.3</td>
<td></td>
</tr>
<tr>
<td>Skewness</td>
<td>3.7</td>
<td>2.0</td>
<td></td>
</tr>
<tr>
<td>Kurtosis</td>
<td>19.4</td>
<td>4.5</td>
<td></td>
</tr>
<tr>
<td>Poverty</td>
<td></td>
<td></td>
<td>0.51</td>
</tr>
<tr>
<td>Mean</td>
<td>70.4</td>
<td>70.7</td>
<td></td>
</tr>
<tr>
<td>SD</td>
<td>41.5</td>
<td>35.7</td>
<td></td>
</tr>
<tr>
<td>Skewness</td>
<td>-0.3</td>
<td>-0.1</td>
<td></td>
</tr>
<tr>
<td>Kurtosis</td>
<td>-1.1</td>
<td>-0.6</td>
<td></td>
</tr>
<tr>
<td>Age, years</td>
<td></td>
<td></td>
<td>0.02</td>
</tr>
<tr>
<td>Mean</td>
<td>38.2</td>
<td>37.8</td>
<td></td>
</tr>
<tr>
<td>SD</td>
<td>13.3</td>
<td>11.7</td>
<td></td>
</tr>
<tr>
<td>Skewness</td>
<td>0.3</td>
<td>0.3</td>
<td></td>
</tr>
<tr>
<td>Kurtosis</td>
<td>-1.1</td>
<td>-0.5</td>
<td></td>
</tr>
<tr>
<td>Male, %</td>
<td>38.6</td>
<td>38.7</td>
<td>0.88</td>
</tr>
<tr>
<td>Black, %</td>
<td>39.6</td>
<td>40.0</td>
<td>0.58</td>
</tr>
<tr>
<td>Hispanic, %</td>
<td>35.7</td>
<td>35.4</td>
<td>0.72</td>
</tr>
<tr>
<td>Education, %</td>
<td></td>
<td></td>
<td>0.97</td>
</tr>
<tr>
<td>Less than high school</td>
<td>34.6</td>
<td>34.5</td>
<td></td>
</tr>
<tr>
<td>High school</td>
<td>35.6</td>
<td>35.4</td>
<td></td>
</tr>
<tr>
<td>College</td>
<td>29.9</td>
<td>30.0</td>
<td></td>
</tr>
<tr>
<td>Insurance, %</td>
<td></td>
<td></td>
<td>0.95</td>
</tr>
<tr>
<td>Any private</td>
<td>19.0</td>
<td>19.1</td>
<td></td>
</tr>
<tr>
<td>Public only</td>
<td>41.0</td>
<td>41.0</td>
<td></td>
</tr>
<tr>
<td>Uninsured</td>
<td>40.0</td>
<td>39.9</td>
<td></td>
</tr>
<tr>
<td>Conditions, %</td>
<td></td>
<td></td>
<td>0.85</td>
</tr>
<tr>
<td>0</td>
<td>39.4</td>
<td>39.7</td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>21.7</td>
<td>21.4</td>
<td></td>
</tr>
<tr>
<td>≥2</td>
<td>38.9</td>
<td>38.8</td>
<td></td>
</tr>
</tbody>
</table>
Figure 16. Correlational relationships among variables in actual and big simulated datasets: all data.
Figure 17. Correlational relationships among variables in actual and big simulated datasets: group 1
Figure 18. Correlational relationships among variables in actual and big simulated datasets: group 2
Figure 19. Correlational relationships among variables in actual and big simulated datasets: group 3
Figure 20. Correlational relationships among variables in actual and big simulated datasets: group 4
the first five randomly generated datasets of prescription count models are presented in Table XXIV. The magnitudes and directions of covariate-outcome relationships were maintained in the data generation process. The only difference in statistical significance was between the Medicaid-expanded indicator, where it was non-significant in the actual dataset and significant in the simulated datasets. The magnitude and direction of this variable was the same between the actual and simulated datasets. Of note, was the preservation of the $PZ$ parameter estimate in terms of magnitude, direction, and significance, as this was the estimand of interest. Regression relationships between the actual and simulated datasets preserved the trends of the actual MEPS data.

Distributional comparisons of variables from the actual (n=2,592) and big generated dataset (n=25,920) in the subgroup are presented in Table XXV. There was no difference in any of the predictor variables between the actual and big simulated datasets (all $p \geq 0.16$). The location of the actual and generated total charges outcome were similar ($2,176$ versus $2,200$). The generated total charges outcome failed to obtain a similar spread to the actual total charges outcome (SD=$2,466$ versus $753$). Testing for a difference between the actual and generated $TOTTC'H$ outcome revealed no difference between the actual and simulated datasets ($p=0.62$). The simulation algorithm did an excellent job in generating the total charges outcome and covariates in the big simulated subgroup.

Correlational relationships of variates in the subgroup are presented visually in Figures 21-25. The size of the circle represents the magnitude of the association, and colors represent the direction of the association. Blue represents a positive correlation and red a negative correlation between variables. Overall, associative relationships were preserved in the simulated datasets. The strength of relationships between some of the variates increased in the data generation...
### TABLE XXIV

REGRESSION COEFFICIENT ESTIMATES FROM ACTUAL AND SIMULATED DATASETS FOR NUMBER OF PRESCRIPTIONS

<table>
<thead>
<tr>
<th>Variable</th>
<th>Actual (n=6,830)</th>
<th>Big (n=68,300)</th>
<th>1 (n=6,830)</th>
<th>2 (n=6,830)</th>
<th>3 (n=6,830)</th>
<th>4 (n=6,830)</th>
<th>5 (n=6,830)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Medicaid Expanded</td>
<td>-0.012</td>
<td>-0.012**</td>
<td>-0.012**</td>
<td>-0.012**</td>
<td>-0.012**</td>
<td>-0.012**</td>
<td>-0.012**</td>
</tr>
<tr>
<td>Post</td>
<td>-0.056***</td>
<td>-0.060***</td>
<td>-0.060***</td>
<td>-0.060***</td>
<td>-0.060***</td>
<td>-0.060***</td>
<td>-0.060***</td>
</tr>
<tr>
<td>Medicaid Expanded x Post</td>
<td>0.232***</td>
<td>0.240***</td>
<td>0.240***</td>
<td>0.240***</td>
<td>0.240***</td>
<td>0.240***</td>
<td>0.240***</td>
</tr>
<tr>
<td>Poverty</td>
<td>0.001***</td>
<td>0.002***</td>
<td>0.002***</td>
<td>0.002***</td>
<td>0.002***</td>
<td>0.002***</td>
<td>0.002***</td>
</tr>
<tr>
<td>Age</td>
<td>0.033***</td>
<td>0.034***</td>
<td>0.034***</td>
<td>0.034***</td>
<td>0.034***</td>
<td>0.034***</td>
<td>0.034***</td>
</tr>
<tr>
<td>Male</td>
<td>-0.135***</td>
<td>-0.142***</td>
<td>-0.142***</td>
<td>-0.142***</td>
<td>-0.142***</td>
<td>-0.142***</td>
<td>-0.142***</td>
</tr>
<tr>
<td>Black</td>
<td>-0.148***</td>
<td>-0.154***</td>
<td>-0.154***</td>
<td>-0.154***</td>
<td>-0.154***</td>
<td>-0.154***</td>
<td>-0.154***</td>
</tr>
<tr>
<td>Hispanic</td>
<td>-0.183***</td>
<td>-0.193***</td>
<td>-0.193***</td>
<td>-0.193***</td>
<td>-0.193***</td>
<td>-0.193***</td>
<td>-0.193***</td>
</tr>
</tbody>
</table>

**Education**

| Less than high school  | ref              | ref              | ref          | ref          | ref          | ref          | ref          |
| High school            | 0.048***         | 0.050***         | 0.050***     | 0.050***     | 0.050***     | 0.050***     | 0.050***     |
| College                | 0.130***         | 0.136***         | 0.136***     | 0.136***     | 0.136***     | 0.136***     | 0.136***     |

**Insurance, %**

| Any private            | ref              | ref              | ref          | ref          | ref          | ref          | ref          |
| Public only            | 0.781***         | 0.816***         | 0.816***     | 0.816***     | 0.816***     | 0.816***     | 0.816***     |
| Uninsured              | -0.414***        | -0.476***        | -0.476***    | -0.476***    | -0.476***    | -0.476***    | -0.476***    |

**Conditions**

| 0                      | ref              | ref              | ref          | ref          | ref          | ref          | ref          |
| 1                      | 0.832***         | 1.164***         | 1.164***     | 1.164***     | 1.164***     | 1.164***     | 1.164***     |
| ≥2                     | 2.199***         | 2.612***         | 2.612***     | 2.612***     | 2.612***     | 2.612***     | 2.612***     |

*p-value: ***≤0.001, **≤0.01, *≤0.05*
## TABLE XXV

### COMPARISON OF THE ACTUAL AND BIG SIMULATED DATASETS IN THE SUBGROUP

<table>
<thead>
<tr>
<th>Variable</th>
<th>Actual (n=2,592)</th>
<th>Simulated (n=25,920)</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Total Charges</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mean</td>
<td>2176</td>
<td>2200</td>
<td>0.62</td>
</tr>
<tr>
<td>SD</td>
<td>2466</td>
<td>753</td>
<td></td>
</tr>
<tr>
<td>Skewness</td>
<td>1.4</td>
<td>0.4</td>
<td></td>
</tr>
<tr>
<td>Kurtosis</td>
<td>0.9</td>
<td>-0.6</td>
<td></td>
</tr>
<tr>
<td><strong>Poverty</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mean</td>
<td>67.6</td>
<td>68.3</td>
<td>0.38</td>
</tr>
<tr>
<td>SD</td>
<td>40.2</td>
<td>34.8</td>
<td></td>
</tr>
<tr>
<td>Skewness</td>
<td>-0.2</td>
<td>-0.1</td>
<td></td>
</tr>
<tr>
<td>Kurtosis</td>
<td>-1.0</td>
<td>-0.1</td>
<td></td>
</tr>
<tr>
<td><strong>Age, years</strong></td>
<td></td>
<td></td>
<td>0.16</td>
</tr>
<tr>
<td>Mean</td>
<td>39.1</td>
<td>38.7</td>
<td></td>
</tr>
<tr>
<td>SD</td>
<td>13.4</td>
<td>11.5</td>
<td></td>
</tr>
<tr>
<td>Skewness</td>
<td>0.1</td>
<td>0.2</td>
<td></td>
</tr>
<tr>
<td>Kurtosis</td>
<td>-1.2</td>
<td>-0.5</td>
<td></td>
</tr>
<tr>
<td><strong>Male, %</strong></td>
<td></td>
<td></td>
<td>0.79</td>
</tr>
<tr>
<td><strong>Black, %</strong></td>
<td></td>
<td></td>
<td>0.97</td>
</tr>
<tr>
<td><strong>Hispanic, %</strong></td>
<td></td>
<td></td>
<td>0.71</td>
</tr>
<tr>
<td><strong>Education, %</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Less than high school</td>
<td>37.7</td>
<td>38.1</td>
<td>0.92</td>
</tr>
<tr>
<td>High school</td>
<td>34.9</td>
<td>34.7</td>
<td></td>
</tr>
<tr>
<td>College</td>
<td>27.4</td>
<td>27.2</td>
<td></td>
</tr>
<tr>
<td><strong>Insurance, %</strong></td>
<td></td>
<td></td>
<td>0.80</td>
</tr>
<tr>
<td>Public only</td>
<td>57.3</td>
<td>57.6</td>
<td></td>
</tr>
<tr>
<td>Uninsured</td>
<td>42.7</td>
<td>42.4</td>
<td></td>
</tr>
<tr>
<td><strong>Conditions, %</strong></td>
<td></td>
<td></td>
<td>0.84</td>
</tr>
<tr>
<td>0</td>
<td>31.0</td>
<td>30.9</td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>23.0</td>
<td>23.6</td>
<td></td>
</tr>
<tr>
<td>≥2</td>
<td>46.0</td>
<td>45.5</td>
<td></td>
</tr>
</tbody>
</table>
Figure 21. Correlational relationships among variables in actual and big simulated datasets: all subgroup
Figure 22. Correlational relationships among variables in actual and big simulated datasets: subgroup, group 1
Figure 23. Correlational relationships among variables in actual and big simulated datasets: subgroup, group 2
Figure 24. Correlational relationships among variables in actual and big simulated datasets: subgroup, group 3
Figure 25. Correlational relationships among variables in actual and big simulated datasets: subgroup, group 4
process. None of the associative trends in the big simulated datasets became weaker in the data generation process.

Table XXVI shows the regression relationship between total charges and covariates in the actual and first five simulated datasets. The magnitude and direction of the regression relationships were well maintained through the data generation procedures. Some covariates gained associative strength through the data generation procedures. Variates that were significant in the actual data remained significant in the generated data, including the covariate of interest, Medicaid expanded and post indicator.

Distributions of simulated datasets were similar in distributional characteristics, marginal and associational relationships, as well as observed trends to the actual MEPS dataset. There were some differences between the actual and big generated datasets in terms of associative strength between variates. The big generated dataset had ten times the number of observations which explain the strengthening of some measured relationships. None of the relationships got weaker. The idiosyncratic nature of outcome variables made it challenging to replicate a similar data generation process for these variables. A minority of individuals comprised the majority of medications and cost. The most prevalent outcome was of zero prescriptions and zero cost. We have overcome these challenges and successfully simulated datasets which were sufficiently close and exhibit similar trends on average to the actual MEPS dataset.

5.2.5 Results

Results of estimator performance under four types, and three sizes of confounding in two continuous variables are displayed in Tables XXVII-XXXII for number of prescriptions. Coverage rates from these results are presented visually in Figures 26-31. Similarly, full results and visual representations of coverage rates for total charges are displayed in Tables XXXIII-
### TABLE XXVI

REGRESSION COEFFICIENT ESTIMATES FROM ACTUAL AND SIMULATED DATASETS FOR TOTAL CHARGES

<table>
<thead>
<tr>
<th>Variable</th>
<th>Actual (n=2,592)</th>
<th>Big (n=25,920)</th>
<th>1 (n=2,592)</th>
<th>2 (n=2,592)</th>
<th>3 (n=2,592)</th>
<th>4 (n=2,592)</th>
<th>5 (n=2,592)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Medicaid Expanded</td>
<td>-0.132</td>
<td>-0.133***</td>
<td>-0.133***</td>
<td>-0.133***</td>
<td>-0.132***</td>
<td>-0.133***</td>
<td></td>
</tr>
<tr>
<td>Post</td>
<td>-0.018</td>
<td>-0.018***</td>
<td>-0.018***</td>
<td>-0.018***</td>
<td>-0.018***</td>
<td>-0.018***</td>
<td></td>
</tr>
<tr>
<td>Medicaid Expanded x Post</td>
<td>0.249*</td>
<td>0.250***</td>
<td>0.249***</td>
<td>0.250***</td>
<td>0.250***</td>
<td>0.250***</td>
<td></td>
</tr>
<tr>
<td>Poverty</td>
<td>0.0002</td>
<td>0.0002***</td>
<td>0.0002***</td>
<td>0.0002***</td>
<td>0.0002***</td>
<td>0.0002***</td>
<td></td>
</tr>
<tr>
<td>Age</td>
<td>-0.001</td>
<td>-0.001***</td>
<td>-0.001***</td>
<td>-0.001***</td>
<td>-0.001***</td>
<td>-0.001***</td>
<td></td>
</tr>
<tr>
<td>Male</td>
<td>-0.175***</td>
<td>-0.175***</td>
<td>-0.175***</td>
<td>-0.175***</td>
<td>-0.175***</td>
<td>-0.175***</td>
<td></td>
</tr>
<tr>
<td>Black</td>
<td>0.119</td>
<td>0.119***</td>
<td>0.119***</td>
<td>0.119***</td>
<td>0.119***</td>
<td>0.119***</td>
<td></td>
</tr>
<tr>
<td>Hispanic</td>
<td>-0.050</td>
<td>-0.050***</td>
<td>-0.050***</td>
<td>-0.050***</td>
<td>-0.050***</td>
<td>-0.050***</td>
<td></td>
</tr>
<tr>
<td>Education</td>
<td>-0.132</td>
<td>-0.133***</td>
<td>-0.133***</td>
<td>-0.133***</td>
<td>-0.132***</td>
<td>-0.133***</td>
<td></td>
</tr>
<tr>
<td>Less than high school</td>
<td>ref</td>
<td>ref</td>
<td>ref</td>
<td>ref</td>
<td>ref</td>
<td>ref</td>
<td></td>
</tr>
<tr>
<td>High school</td>
<td>0.080</td>
<td>0.080***</td>
<td>0.080***</td>
<td>0.080***</td>
<td>0.080***</td>
<td>0.080***</td>
<td></td>
</tr>
<tr>
<td>College</td>
<td>-0.134*</td>
<td>-0.134***</td>
<td>-0.134***</td>
<td>-0.135***</td>
<td>-0.134***</td>
<td>-0.134***</td>
<td></td>
</tr>
<tr>
<td>Insurance</td>
<td>ref</td>
<td>ref</td>
<td>ref</td>
<td>ref</td>
<td>ref</td>
<td>ref</td>
<td></td>
</tr>
<tr>
<td>Public only</td>
<td>-0.363***</td>
<td>-0.363***</td>
<td>-0.363***</td>
<td>-0.363***</td>
<td>-0.363***</td>
<td>-0.363***</td>
<td></td>
</tr>
<tr>
<td>Uninsured</td>
<td>ref</td>
<td>ref</td>
<td>ref</td>
<td>ref</td>
<td>ref</td>
<td>ref</td>
<td></td>
</tr>
<tr>
<td>Conditions</td>
<td>ref</td>
<td>ref</td>
<td>ref</td>
<td>ref</td>
<td>ref</td>
<td>ref</td>
<td></td>
</tr>
<tr>
<td>0</td>
<td>0.258***</td>
<td>0.258***</td>
<td>0.258***</td>
<td>0.258***</td>
<td>0.258***</td>
<td>0.258***</td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>0.521***</td>
<td>0.522***</td>
<td>0.522***</td>
<td>0.522***</td>
<td>0.522***</td>
<td>0.522***</td>
<td></td>
</tr>
<tr>
<td>≥2</td>
<td>0.521***</td>
<td>0.522***</td>
<td>0.522***</td>
<td>0.522***</td>
<td>0.522***</td>
<td>0.522***</td>
<td></td>
</tr>
</tbody>
</table>

*P-value: ***≤0.001, **≤0.01, *≤0.05*
<table>
<thead>
<tr>
<th>Setting</th>
<th>Two-way FE\textsuperscript{b}</th>
<th>Separate time weighted</th>
<th>Four-group weighted</th>
<th>AIPW\textsuperscript{b}</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>BS SE\textsuperscript{ab}</td>
<td>SE\textsuperscript{ab}</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>(4b): incl. X, basic</td>
<td>BS SE\textsuperscript{ab}</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>perc BS SE\textsuperscript{ab}</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>(5) incl. X, basic</td>
<td>BS SE\textsuperscript{ab}</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>perc BS SE\textsuperscript{ab}</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>(6a): incl. X, basic</td>
<td>BS SE\textsuperscript{ab}</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>perc BS SE\textsuperscript{ab}</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>(6b): incl. X, basic</td>
<td>BS SE\textsuperscript{ab}</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>perc BS SE\textsuperscript{ab}</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>(6c): incl. X, basic</td>
<td>BS SE\textsuperscript{ab}</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>perc BS SE\textsuperscript{ab}</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>(7a): sand SE</td>
<td>BS SE\textsuperscript{ab}</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>perc BS SE\textsuperscript{ab}</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>(7b): basic BS</td>
<td>SE\textsuperscript{ab}</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>perc BS SE\textsuperscript{ab}</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>(7c): perc BS</td>
<td>SE\textsuperscript{ab}</td>
</tr>
<tr>
<td>Mean Bias</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>-0.2860 0.0130</td>
<td>-0.0398 0.0383</td>
<td>0.0386 0.0386</td>
<td>0.3631 0.0109 0.0115 0.0115</td>
</tr>
<tr>
<td>2</td>
<td>-0.2866 0.0132</td>
<td>-0.0689 0.0376</td>
<td>0.0380 0.0380</td>
<td>0.3961 0.0109 0.0115 0.0115</td>
</tr>
<tr>
<td>3</td>
<td>-0.2957 0.0126</td>
<td>-0.0718 0.0358</td>
<td>0.0362 0.0362</td>
<td>0.3622 0.0074 0.0080 0.0080</td>
</tr>
<tr>
<td>4</td>
<td>-0.4413 0.0142</td>
<td>-0.0880 0.0394</td>
<td>0.0398 0.0398</td>
<td>0.4282 0.0099 0.0106 0.0106</td>
</tr>
<tr>
<td>RMSE</td>
<td>0.8790 0.0549</td>
<td>0.3413 0.0679</td>
<td>0.0685 0.0685</td>
<td>0.7279 0.0508 0.0510 0.0510</td>
</tr>
<tr>
<td>2</td>
<td>0.8792 0.0548</td>
<td>0.3431 0.0675</td>
<td>0.0680 0.0680</td>
<td>0.7570 0.0505 0.0507 0.0507</td>
</tr>
<tr>
<td>3</td>
<td>0.8822 0.0548</td>
<td>0.3437 0.0666</td>
<td>0.0670 0.0670</td>
<td>0.7376 0.0501 0.0503 0.0503</td>
</tr>
<tr>
<td>4</td>
<td>0.9411 0.0572</td>
<td>0.3532 0.0703</td>
<td>0.0708 0.0708</td>
<td>0.7848 0.0524 0.0527 0.0527</td>
</tr>
<tr>
<td>95% Confidence Interval Coverage</td>
<td>0.326 1.000</td>
<td>0.558 1.000</td>
<td>0.877 0.878</td>
<td>0.524 1.000 0.943 0.942</td>
</tr>
<tr>
<td>2</td>
<td>0.325 1.000</td>
<td>0.544 1.000</td>
<td>0.877 0.877</td>
<td>0.508 1.000 0.942 0.947</td>
</tr>
<tr>
<td>3</td>
<td>0.319 1.000</td>
<td>0.544 1.000</td>
<td>0.884 0.882</td>
<td>0.523 1.000 0.948 0.954</td>
</tr>
<tr>
<td>4</td>
<td>0.293 1.000</td>
<td>0.552 1.000</td>
<td>0.877 0.880</td>
<td>0.493 1.000 0.948 0.950</td>
</tr>
</tbody>
</table>

\textsuperscript{a}areas shaded in grey represent the proposed methods in this dissertation

\textsuperscript{b}abbreviations: FE=fixed effects, AIPW=augmented inverse probability weighted estimator, BS=bootstrap, SE=standard error, perc=percentile, sand=sandwich
## TABLE XXVIII

**SIMULATION RESULTS IN 3 TYPES OF MODERATE POVERTY COVARIATE CONFOUNDING FOR NUMBER OF PRESCRIPTIONS**

<table>
<thead>
<tr>
<th>Setting</th>
<th>Two-way FE&lt;sup&gt;b&lt;/sup&gt;</th>
<th>Separate time weighted</th>
<th>Four-group weighted</th>
<th>AIPW&lt;sup&gt;b&lt;/sup&gt;</th>
</tr>
</thead>
<tbody>
<tr>
<td>(1)</td>
<td>(2): incl. X</td>
<td>(3)</td>
<td>(4a): incl. X</td>
<td>(4b): incl. X, basic BS Se&lt;sup&gt;a,b&lt;/sup&gt;</td>
</tr>
<tr>
<td>Mean Bias</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>-0.2876</td>
<td>0.0135</td>
<td>-0.1136</td>
<td>0.0362</td>
</tr>
<tr>
<td>3</td>
<td>-0.2849</td>
<td>0.0348</td>
<td>-0.1112</td>
<td>0.0440</td>
</tr>
<tr>
<td>4</td>
<td>-0.6060</td>
<td>0.0117</td>
<td>-0.1652</td>
<td>0.0389</td>
</tr>
<tr>
<td>RMSE</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>0.8795</td>
<td>0.0551</td>
<td>0.3894</td>
<td>0.0674</td>
</tr>
<tr>
<td>3</td>
<td>0.8787</td>
<td>0.0643</td>
<td>0.3898</td>
<td>0.0721</td>
</tr>
<tr>
<td>4</td>
<td>1.0287</td>
<td>0.0590</td>
<td>0.3956</td>
<td>0.0722</td>
</tr>
<tr>
<td>95% Confidence Interval Coverage</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>0.325</td>
<td>1.000</td>
<td>0.483</td>
<td>1.000</td>
</tr>
<tr>
<td>3</td>
<td>0.329</td>
<td>1.000</td>
<td>0.481</td>
<td>1.000</td>
</tr>
<tr>
<td>4</td>
<td>0.261</td>
<td>1.000</td>
<td>0.485</td>
<td>1.000</td>
</tr>
</tbody>
</table>

<sup>a</sup>areas shaded in grey represent the proposed methods in this dissertation

<sup>b</sup>abbreviations: FE=fixed effects, AIPW=augmented inverse probability weighted estimator, BS=bootstrap, SE=standard error, perc=percentile, sand=sandwich
TABLE XXIX

SIMULATION RESULTS IN 3 TYPES OF LARGE POVERTY COVARIATE CONFOUNDING FOR NUMBER OF PRESCRIPTIONS

<table>
<thead>
<tr>
<th>Setting</th>
<th>Two-way FE (^b)</th>
<th>Separate time weighted</th>
<th>Four-group weighted</th>
<th>AIPW (^b)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>(1)</td>
<td>(2): incl. X</td>
<td>(3)</td>
<td>(4a): incl. X</td>
</tr>
<tr>
<td>Mean Bias</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>-0.2885</td>
<td>0.0146</td>
<td>-0.1578</td>
<td>0.0348</td>
</tr>
<tr>
<td>3</td>
<td>-0.2773</td>
<td>0.0956</td>
<td>-0.1492</td>
<td>0.0723</td>
</tr>
<tr>
<td>4</td>
<td>-0.8081</td>
<td>0.0011</td>
<td>-0.2713</td>
<td>0.0325</td>
</tr>
</tbody>
</table>

RMSE

|                    |                  |                        |                     |             |                                     |                                     | 0.9615 | 0.0515 | 0.0518 | 0.0518 | 0.0744 | 0.0750 | 0.0750 |
| 2       | 0.8798          | 0.0558                | 0.4866              | 0.0683      | 0.0687 | 0.0687 | 0.9615 | 0.0515 | 0.0518 | 0.0518 | 0.0744 | 0.0750 | 0.0750 |
| 3       | 0.8762          | 0.1109                | 0.4893              | 0.0939      | 0.0949 | 0.0949 | 0.8874 | 0.0558 | 0.0565 | 0.0565 | 0.0804 | 0.0809 | 0.0809 |
| 4       | 1.1593          | 0.0609                | 0.4898              | 0.0721      | 0.0726 | 0.0726 | 1.0492 | 0.0611 | 0.0608 | 0.0608 | 0.4883 | 0.4884 | 0.4884 |

95% Confidence Interval Coverage

|                    |                  |                        |                     |             |                                     |                                     | 0.420  | 1.000  | 0.948  | 0.951  | 0.924  | 0.911  | 0.912  |
| 2       | 0.324           | 1.000                 | 0.377               | 1.000       | 0.889  | 0.895  | 0.420  | 1.000  | 0.948  | 0.951  | 0.924  | 0.911  | 0.912  |
| 3       | 0.333           | 1.000                 | 0.375               | 1.000       | 0.761  | 0.745  | 0.448  | 1.000  | 0.932  | 0.932  | 0.890  | 0.875  | 0.885  |
| 4       | 0.205           | 1.000                 | 0.384               | 1.000       | 0.906  | 0.905  | 0.393  | 1.000  | 0.937  | 0.939  | 0.000  | 0.000  | 0.000  |

\(^a\)areas shaded in grey represent the proposed methods in this dissertation

\(^b\)abbreviations: FE=fixed effects, AIPW=augmented inverse probability weighted estimator, BS=bootstrap, SE=standard error, perc=percentile, sand=sandwich


TABLE XXX

SIMULATION RESULTS IN 4 TYPES OF SMALL AGE COVARIATE CONFOUNDING FOR NUMBER OF PRESCRIPTIONS

<table>
<thead>
<tr>
<th>Setting</th>
<th>Two-way FE(^b)</th>
<th>Separate time weighted</th>
<th>Four-group weighted</th>
<th>AIPW(^b)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>(1) incl. X</td>
<td>(2):</td>
<td>(3)</td>
<td>(4): incl. X, basic BS SE(^a,b)</td>
</tr>
<tr>
<td></td>
<td>(2):</td>
<td>(3):</td>
<td>(4):</td>
<td>(5): incl. X, basic BS SE(^a,b)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(3): incl. X</td>
<td>(4): incl. X</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>(4): perc BS</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mean Bias</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>-0.2916 0.0074</td>
<td>-0.0454 0.0326</td>
<td>0.0330 0.0330</td>
<td>0.3574 0.0053</td>
</tr>
<tr>
<td>2</td>
<td>-0.3184 -0.0174</td>
<td>-0.0814 0.0116</td>
<td>0.0119 0.0119</td>
<td>0.5493 -0.0254</td>
</tr>
<tr>
<td>3</td>
<td>-0.3064 0.0298</td>
<td>-0.0585 0.0089</td>
<td>0.0091 0.0091</td>
<td>0.7798 -0.0648</td>
</tr>
<tr>
<td>4</td>
<td>-1.4314 0.0051</td>
<td>0.1784 0.0345</td>
<td>0.0356 0.0356</td>
<td>1.0545 -0.0326</td>
</tr>
<tr>
<td>RMSE</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>0.8808 0.0538</td>
<td>0.3420 0.0650</td>
<td>0.0654 0.0654</td>
<td>0.7251 0.0499</td>
</tr>
<tr>
<td>2</td>
<td>0.8901 0.0617</td>
<td>0.3248 0.0618</td>
<td>0.0623 0.0623</td>
<td>0.9510 0.0634</td>
</tr>
<tr>
<td>3</td>
<td>0.8859 0.0750</td>
<td>0.3219 0.0674</td>
<td>0.0678 0.0678</td>
<td>1.0576 0.0907</td>
</tr>
<tr>
<td>4</td>
<td>1.6552 0.0694</td>
<td>0.4170 0.0797</td>
<td>0.0805 0.0805</td>
<td>1.3112 0.0745</td>
</tr>
<tr>
<td>95% Confidence Interval Coverage</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>0.320 1.000</td>
<td>0.559 1.000</td>
<td>0.899 0.897</td>
<td>0.524 1.000</td>
</tr>
<tr>
<td>2</td>
<td>0.313 1.000</td>
<td>0.581 1.000</td>
<td>0.937 0.943</td>
<td>0.408 1.000</td>
</tr>
<tr>
<td>3</td>
<td>0.318 1.000</td>
<td>0.585 1.000</td>
<td>0.945 0.940</td>
<td>0.349 1.000</td>
</tr>
<tr>
<td>4</td>
<td>0.087 1.000</td>
<td>0.487 1.000</td>
<td>0.911 0.911</td>
<td>0.260 1.000</td>
</tr>
</tbody>
</table>

\(^a\)areas shaded in grey represent the proposed methods in this dissertation

\(^b\)abbreviations: FE=fixed effects, AIPW=augmented inverse probability weighted estimator, BS=bootstrap, SE=standard error, perc=percentile, sand=sandwich
<table>
<thead>
<tr>
<th>Setting</th>
<th>Two-way FE&lt;sup&gt;b&lt;/sup&gt; (1)</th>
<th>Separate time weighted (3)</th>
<th>Four-group weighted (5)</th>
<th>AIPW&lt;sup&gt;b&lt;/sup&gt; (7a): sand SE</th>
<th>AIPW&lt;sup&gt;b&lt;/sup&gt; (7b): basic BS SE&lt;sup&gt;a,b&lt;/sup&gt;</th>
<th>AIPW&lt;sup&gt;b&lt;/sup&gt; (7c): perc BS SE&lt;sup&gt;a,b&lt;/sup&gt;</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Mean Bias</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>-0.3491</td>
<td>-0.0461</td>
<td>-0.1057</td>
<td>-0.0215</td>
<td>0.7403</td>
<td>-0.0213</td>
</tr>
<tr>
<td>3</td>
<td>-0.1895</td>
<td>0.3292</td>
<td>0.1287</td>
<td>0.1079</td>
<td>1.8823</td>
<td>-0.1049</td>
</tr>
<tr>
<td>4</td>
<td>-2.7170</td>
<td>-0.0956</td>
<td>0.0564</td>
<td>-0.0137</td>
<td>2.2592</td>
<td>-0.2436</td>
</tr>
<tr>
<td><strong>RMSE</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>0.9015</td>
<td>0.0932</td>
<td>0.3620</td>
<td>0.0751</td>
<td>1.2234</td>
<td>0.1074</td>
</tr>
<tr>
<td>3</td>
<td>0.8525</td>
<td>0.3491</td>
<td>0.3747</td>
<td>0.1450</td>
<td>2.0736</td>
<td>0.1405</td>
</tr>
<tr>
<td>4</td>
<td>2.8413</td>
<td>0.1346</td>
<td>0.4161</td>
<td>0.0951</td>
<td>2.5687</td>
<td>0.2624</td>
</tr>
<tr>
<td><strong>95% Confidence Interval Coverage</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>0.320</td>
<td>1.000</td>
<td>0.529</td>
<td>0.999</td>
<td>0.331</td>
<td>1.000</td>
</tr>
<tr>
<td>3</td>
<td>0.350</td>
<td>0.674</td>
<td>0.508</td>
<td>0.925</td>
<td>0.052</td>
<td>1.000</td>
</tr>
<tr>
<td>4</td>
<td>0.003</td>
<td>0.999</td>
<td>0.532</td>
<td>0.998</td>
<td>0.081</td>
<td>0.993</td>
</tr>
</tbody>
</table>

<sup>a</sup>areas shaded in grey represent the proposed methods in this dissertation

<sup>b</sup>abbreviations: FE=fixed effects, AIPW=augmented inverse probability weighted estimator, BS=bootstrap, SE=standard error, perc=percentile, sand=sandwich
# TABLE XXXII

SIMULATION RESULTS IN 3 TYPES OF LARGE AGE COVARIATE CONFOUNDING FOR NUMBER OF PRESCRIPTIONS

<table>
<thead>
<tr>
<th>Setting</th>
<th>Two-way FE&lt;sup&gt;b&lt;/sup&gt;</th>
<th>Separate time weighted</th>
<th>Four-group weighted</th>
<th>AIPW&lt;sup&gt;b&lt;/sup&gt;</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>(1)</td>
<td>(2): incl. X</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mean Bias</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>-0.3696</td>
<td>-0.0649</td>
<td>-0.1099</td>
<td>0.9585</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>0.0988</td>
<td>1.0374</td>
<td>0.6058</td>
<td>3.2085</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>-4.6751</td>
<td>-3.8544</td>
<td>-4.1391</td>
<td>-3.2821</td>
</tr>
<tr>
<td>RMSE</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>0.9096</td>
<td>0.1361</td>
<td>0.4690</td>
<td>1.6022</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>0.8370</td>
<td>1.0524</td>
<td>0.7652</td>
<td>3.1517</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>4.7485</td>
<td>3.8548</td>
<td>4.1599</td>
<td>3.4029</td>
</tr>
<tr>
<td>95% Confidence Interval Coverage</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>0.318</td>
<td>0.997</td>
<td>0.446</td>
<td>0.281</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>0.358</td>
<td>0.000</td>
<td>0.219</td>
<td>0.008</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.004</td>
</tr>
</tbody>
</table>

<sup>a</sup>areas shaded in grey represent the proposed methods in this dissertation

<sup>b</sup>abbreviations: FE=fixed effects, AIPW=augmented inverse probability weighted estimator, BS=bootstrap, SE=standard error, perc=percentile, sand=sandwich
Figure 26. Coverage results under small poverty covariate confounding for number of Medicaid prescriptions

1. Naive
2. Naive with X
3. Sep time wts
4a. Sep time wts+X
4b. Sep time wts+X. basic bs
4c. Sep time wts+X. perc bs
5. 4-grp wts
6a. 4-grp wts+X
6b. 4-grp wts+X. basic bs
6c. 4-grp wts+X. perc bs
7a. AIPW. sand
7b. AIPW. bs basic
7c. AIPW. bs perc

aabbreviations: sep=separate, wts=weights, grp=group, perc=percentile, bs=bootstrap, sand=sandwich
Figure 27. Coverage results under medium poverty covariate confounding for number of Medicaid prescriptions

1. Naive
2. Naive with X
3. Sep time wts
4a. Sep time wts+X
4b. Sep time wts+X, basic bs
4c. Sep time wts+X, perc bs
5. 4-grp wts
6a. 4-grp wts+X
6b. 4-grp wts+X, basic bs
6c. 4-grp wts+X, perc bs
7a. AIPW, sand
7b. AIPW, bs basic
7c. AIPW, bs perc

*a abbreviations: sep=separate, wts=weights, grp=group, perc=percentile, bs=bootstrap, sand=sandwich
Figure 28. Coverage results under large poverty covariate confounding for number of Medicaid prescriptions

1. Naive
2. Naive with X
3. Sep time wts
4a. Sep time wts+X
4b. Sep time wts+X. basic bs
4c. Sep time wts+X. perc bs
5. 4-grp wts
6a. 4-grp wts+X
6b. 4-grp wts+X. basic bs
6c. 4-grp wts+X. perc bs
7a. AIPW. sand
7b. AIPW. bs basic
7c. AIPW. bs perc

^ abbreviations: sep=separate, wts=weights, grp=group, perc=percentile, bs=bootstrap, sand=sandwich
Figure 29. Coverage results under small age covariate confounding for number of Medicaid prescriptions “a”

1. Naive
2. Naive with X
3. Sep time wts
4a. Sep time wts+X
4b. Sep time wts+X. basic bs
4c. Sep time wts+X. perc bs
5. 4-grp wts
6a. 4-grp wts+X
6b. 4-grp wts+X. basic bs
6c. 4-grp wts+X. perc bs
7a. AIPW. sand
7b. AIPW. bs basic
7c. AIPW. bs perc

“a” abbreviations: sep=separate, wts=weights, grp=group, perc=percentile, bs=bootstrap, sand=sandwich
Figure 30. Coverage results under medium age covariate confounding for number of Medicaid prescriptions

1. Naive
2. Naive with X
3. Sep time wts
4a. Sep time wts+X
4b. Sep time wts+X. basic bs
4c. Sep time wts+X. perc bs
5. 4-grp wts
6a. 4-grp wts+X
6b. 4-grp wts+X. basic bs
6c. 4-grp wts+X. perc bs
7a. AIPW. sand
7b. AIPW. bs basic
7c. AIPW. bs perc

Abbreviations: sep=separate, wts=weights, grp=group, perc=percentile, bs=bootstrap, sand=sandwich
Figure 31. Coverage results under large age covariate confounding for number of Medicaid prescriptions$^a$

1. Naive
2. Naive with X
3. Sep time wts
4a. Sep time wts+X
4b. Sep time wts+X. basic bs
4c. Sep time wts+X. perc bs
5. 4-grp wts
6a. 4-grp wts+X
6b. 4-grp wts+X. basic bs
6c. 4-grp wts+X. perc bs
7a. AIPW. sand
7b. AIPW. bs basic
7c. AIPW. bs perc

$^a$abbreviations: sep=separate, wts=weights, grp=group, perc=percentile, bs=bootstrap, sand=sandwich
XXXVIII and Figures 32-37. A black vertical line on coverage plots represent the 0.95 nominal level, and results shaded in red represent the estimators proposed in this dissertation.

Regression adjustment greatly improved bias and RMSE for all outcomes, comparing all covariate adjusted estimators to its unadjusted counterpart (2 versus 1, 4a versus 3, 6a versus 5). Our hypothesis involving regression adjustment was that it would result in an improvement to the precision. For the outcome $RXTOT$, regression adjustment did not notably influence the length of confidence intervals. Conversely, for the outcome $TOTTCCH$, regression adjustment hugely reduced the standard error. While bias was good for both outcomes after regression adjustment, the effect to the standard error influenced coverage in different ways. Regression adjustment in $RXTOT$ produced a negligible bias but no notable effect to the standard error, and resulted in overcoverage. Regression adjustment in $TOTTCCH$ produced a negligible bias as well, but detrimentally small standard errors which resulted in undercoverage. We expected the main gains from covariate adjustment would be to the precision. Included covariates must have captured some of the bias that otherwise would have been confounded with treatment (Murray, 1998). However, an unexpected result from covariate adjustment was the great improvement to bias. Covariate adjustment in unweighted and weighted models had the ability to reduce standard errors, though it was erratic and extreme. Model-based standard errors for all outcomes, in regression-adjusted, weighted, and unweighted models all resulted in disastrous precision.

It was hypothesized the separate time weighted estimator would improve bias, at the cost of perhaps an increased variance. To test this hypothesis, we compared the separate time weighted estimator with covariate adjustment to the two-way fixed effects estimator with covariate adjustment (estimators 4a versus 2). The separate time weighted estimator showed little improvement to the bias except in some cases of a group and time bias (setting 3). Separate time
### TABLE XXXIII

SIMULATION RESULTS IN 4 TYPES OF SMALL POVERTY COVARIATE CONFOUNDING FOR TOTAL CHARGES

<table>
<thead>
<tr>
<th>Setting</th>
<th>Two-way FE&lt;sup&gt;b&lt;/sup&gt;</th>
<th>Separate time weighted</th>
<th>Four-group weighted</th>
<th>AIPW&lt;sup&gt;b&lt;/sup&gt;</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>(1)</td>
<td>(2): incl. X</td>
<td>(3)</td>
<td>(4a): incl. X</td>
</tr>
<tr>
<td>Mean Bias</td>
<td>80.58 0.61</td>
<td>-13.06 0.63</td>
<td>0.62 0.62</td>
<td>-2.49 0.62</td>
</tr>
<tr>
<td>1</td>
<td>80.65 0.64</td>
<td>0.30 0.65</td>
<td>0.65 0.65</td>
<td>-7.90 0.62</td>
</tr>
<tr>
<td>2</td>
<td>80.49 0.71</td>
<td>-0.26 0.69</td>
<td>0.69 0.69</td>
<td>6.08 0.64</td>
</tr>
<tr>
<td>3</td>
<td>77.10 0.62</td>
<td>-12.36 0.64</td>
<td>0.64 0.64</td>
<td>-16.06 0.61</td>
</tr>
<tr>
<td>RMSE</td>
<td>109.88 3.86</td>
<td>27.40 3.87</td>
<td>3.89 3.89</td>
<td>31.16 3.86</td>
</tr>
<tr>
<td>1</td>
<td>109.93 3.87</td>
<td>25.23 3.87</td>
<td>3.89 3.89</td>
<td>32.19 3.87</td>
</tr>
<tr>
<td>2</td>
<td>109.82 3.88</td>
<td>25.11 3.88</td>
<td>3.90 3.90</td>
<td>34.09 3.87</td>
</tr>
<tr>
<td>3</td>
<td>107.35 3.89</td>
<td>27.28 3.89</td>
<td>3.91 3.91</td>
<td>34.59 3.89</td>
</tr>
<tr>
<td>4</td>
<td>107.81 3.89</td>
<td>27.28 3.89</td>
<td>3.91 3.91</td>
<td>34.59 3.89</td>
</tr>
<tr>
<td>95% Confidence Interval Coverage</td>
<td>0.861 0.015</td>
<td>1.000 0.016</td>
<td>0.938 0.942</td>
<td>1.000 0.014</td>
</tr>
<tr>
<td>1</td>
<td>0.860 0.022</td>
<td>1.000 0.020</td>
<td>0.936 0.941</td>
<td>1.000 0.019</td>
</tr>
<tr>
<td>2</td>
<td>0.861 0.031</td>
<td>1.000 0.024</td>
<td>0.936 0.941</td>
<td>1.000 0.025</td>
</tr>
<tr>
<td>3</td>
<td>0.868 0.020</td>
<td>1.000 0.019</td>
<td>0.936 0.942</td>
<td>1.000 0.018</td>
</tr>
</tbody>
</table>

<sup>a</sup>areas shaded in grey represent the proposed methods in this dissertation

<sup>b</sup>abbreviations: FE=fixed effects, AIPW=augmented inverse probability weighted estimator, BS=bootstrap, SE=standard error, perc=percentile, sand=sandwich
### TABLE XXXIV

**SIMULATION RESULTS IN 3 TYPES OF MODERATE POVERTY COVARIATE CONFOUNDING FOR TOTAL CHARGES**

<table>
<thead>
<tr>
<th>Setting</th>
<th>Two-way FE&lt;sup&gt;b&lt;/sup&gt; (1) incl. X</th>
<th>Separate time weighted</th>
<th>Four-group weighted</th>
<th>AIPW&lt;sup&gt;b&lt;/sup&gt;</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>(2):</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>incl. X</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>80.73 0.69</td>
<td>20.09 0.68</td>
<td>-15.50 0.62</td>
<td>-53.07 0.59</td>
</tr>
<tr>
<td>3</td>
<td>80.81 1.20</td>
<td>21.29 0.94</td>
<td>18.63 0.70</td>
<td>-53.00 0.67</td>
</tr>
<tr>
<td>4</td>
<td>73.49 0.59</td>
<td>-6.02 0.66</td>
<td>-36.32 0.49</td>
<td>-60.31 0.67</td>
</tr>
<tr>
<td>Mean Bias</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>80.73 0.69</td>
<td>20.09 0.68</td>
<td>-15.50 0.62</td>
<td>-53.07 0.59</td>
</tr>
<tr>
<td>3</td>
<td>80.81 1.20</td>
<td>21.29 0.94</td>
<td>18.63 0.70</td>
<td>-53.00 0.67</td>
</tr>
<tr>
<td>4</td>
<td>73.49 0.59</td>
<td>-6.02 0.66</td>
<td>-36.32 0.49</td>
<td>-60.31 0.67</td>
</tr>
<tr>
<td>RMSE</td>
<td>109.99 3.88</td>
<td>39.18 3.88</td>
<td>38.32 3.87</td>
<td>53.19 3.88</td>
</tr>
<tr>
<td>2</td>
<td>110.05 4.00</td>
<td>40.10 3.93</td>
<td>45.71 3.88</td>
<td>53.11 3.93</td>
</tr>
<tr>
<td>3</td>
<td>104.79 3.91</td>
<td>27.97 3.92</td>
<td>50.09 3.89</td>
<td>60.41 3.91</td>
</tr>
<tr>
<td>95% Confidence Interval Coverage</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>0.858 0.030</td>
<td>1.00 0.028</td>
<td>1.00 0.042</td>
<td>0.000 0.000</td>
</tr>
<tr>
<td>3</td>
<td>0.857 0.055</td>
<td>1.00 0.049</td>
<td>0.993 0.050</td>
<td>0.000 0.000</td>
</tr>
<tr>
<td>4</td>
<td>0.883 0.029</td>
<td>1.00 0.024</td>
<td>0.993 0.037</td>
<td>0.000 0.000</td>
</tr>
</tbody>
</table>

<sup>a</sup>areas shaded in grey represent the proposed methods in this dissertation

<sup>b</sup>abbreviations: FE=fixed effects, AIPW=augmented inverse probability weighted estimator, BS=bootstrap, SE=standard error, perc=percentile, sand=sandwich
### TABLE XXXV

SIMULATION RESULTS IN 3 TYPES OF LARGE POVERTY COVARIATE CONFOUNDING FOR TOTAL CHARGES

<table>
<thead>
<tr>
<th>Setting</th>
<th>Two-way FE&lt;sup&gt;b&lt;/sup&gt;</th>
<th>Separate time weighted</th>
<th>Four-group weighted</th>
<th>AIPW&lt;sup&gt;b&lt;/sup&gt;</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>(1)</td>
<td>(2): incl. X</td>
<td>(3)</td>
<td>(4a): incl. X</td>
</tr>
<tr>
<td>Mean Bias</td>
<td>80.78 0.74</td>
<td>39.17 0.69</td>
<td>-22.38 0.58</td>
<td>0.56 0.56</td>
</tr>
<tr>
<td>3</td>
<td>80.88 2.35</td>
<td>44.21 1.46</td>
<td>32.13 0.84</td>
<td>0.82 0.82</td>
</tr>
<tr>
<td>4</td>
<td>69.14 0.49</td>
<td>-3.05 0.63</td>
<td>-55.45 0.19</td>
<td>0.16 0.16</td>
</tr>
<tr>
<td>4</td>
<td>101.79 3.92</td>
<td>33.70 3.94</td>
<td>69.74 3.90</td>
<td>3.91 3.91</td>
</tr>
</tbody>
</table>

<sup>a</sup>areas shaded in grey represent the proposed methods in this dissertation

<sup>b</sup>abbreviations: FE=fixed effects, AIPW=augmented inverse probability weighted estimator, BS=bootstrap, SE=standard error, perc=percentile, sand=sandwich
<table>
<thead>
<tr>
<th>Setting</th>
<th>Two-way FE&lt;sup&gt;b&lt;/sup&gt;</th>
<th>Separate time weighted</th>
<th>Four-group weighted</th>
<th>AIPW&lt;sup&gt;b&lt;/sup&gt;</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>(1) incl. X</td>
<td>(2): incl. X</td>
<td>(3)</td>
<td>(4a): incl. X</td>
</tr>
<tr>
<td></td>
<td>(4b): incl. X, basic BS</td>
<td>Se&lt;sub&gt;a,b&lt;/sub&gt;</td>
<td></td>
<td>(4c): incl. X, perc BS SE&lt;sub&gt;a,b&lt;/sub&gt;</td>
</tr>
<tr>
<td>Mean Bias</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>80.59</td>
<td>-13.06</td>
<td>0.62</td>
<td>-2.49</td>
</tr>
<tr>
<td>2</td>
<td>77.46</td>
<td>-19.54</td>
<td>0.65</td>
<td>-2.26</td>
</tr>
<tr>
<td>3</td>
<td>77.32</td>
<td>-19.89</td>
<td>0.55</td>
<td>3.39</td>
</tr>
<tr>
<td>4</td>
<td>89.97</td>
<td>-16.49</td>
<td>0.63</td>
<td>4.16</td>
</tr>
<tr>
<td>RMSE</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>109.88</td>
<td>27.40</td>
<td>3.89</td>
<td>31.16</td>
</tr>
<tr>
<td>2</td>
<td>107.61</td>
<td>35.58</td>
<td>3.91</td>
<td>33.66</td>
</tr>
<tr>
<td>3</td>
<td>107.51</td>
<td>35.87</td>
<td>3.90</td>
<td>35.14</td>
</tr>
<tr>
<td>4</td>
<td>116.94</td>
<td>30.55</td>
<td>3.83</td>
<td>32.61</td>
</tr>
<tr>
<td>95% Confidence Interval Coverage</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>0.861</td>
<td>1.000</td>
<td>0.938</td>
<td>1.000</td>
</tr>
<tr>
<td>2</td>
<td>0.865</td>
<td>1.000</td>
<td>0.937</td>
<td>1.000</td>
</tr>
<tr>
<td>3</td>
<td>0.866</td>
<td>1.000</td>
<td>0.938</td>
<td>1.000</td>
</tr>
<tr>
<td>4</td>
<td>0.825</td>
<td>1.000</td>
<td>0.937</td>
<td>1.000</td>
</tr>
</tbody>
</table>

<sup>a</sup>Areas shaded in grey represent the proposed methods in this dissertation

<sup>b</sup>Abbreviations: FE=fixed effects, AIPW=augmented inverse probability weighted estimator, BS=bootstrap, SE=standard error, perc=percentile, sand=sandwich
TABLE XXXVII

SIMULATION RESULTS IN 3 TYPES OF MODERATE AGE COVARIATE CONFOUNDING FOR TOTAL CHARGES

<table>
<thead>
<tr>
<th>Setting</th>
<th>Two-way FE&lt;sup&gt;b&lt;/sup&gt;</th>
<th>Separate time weighted</th>
<th>Four-group weighted</th>
<th>AIPW&lt;sup&gt;b&lt;/sup&gt;</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>(1) incl. X</td>
<td>(2): incl. X</td>
<td>(3) incl. X</td>
<td>(4a): incl. X, basic BS Se&lt;sup&gt;a,b&lt;/sup&gt;</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(4b): incl. X, perc BS BS SE&lt;sup&gt;a,b&lt;/sup&gt;</td>
<td>(4c): incl. X, perc BS BS SE&lt;sup&gt;a,b&lt;/sup&gt;</td>
<td>(5) incl. X</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>(6a): incl. X, basic BS BS SE&lt;sup&gt;a,b&lt;/sup&gt;</td>
<td>(6b): incl. X, perc BS BS SE&lt;sup&gt;a,b&lt;/sup&gt;</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>(6c): incl. X, perc BS BS SE&lt;sup&gt;a,b&lt;/sup&gt;</td>
<td>(7a): sand SE</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>(7b): basic BS BS SE&lt;sup&gt;a,b&lt;/sup&gt;</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>(7c): perc BS BS SE&lt;sup&gt;a,b&lt;/sup&gt;</td>
</tr>
</tbody>
</table>

| Mean Bias | | | | | |
|-----------|----------------------|----------------------|----------------------|----------------------|----------------------|----------------------|----------------------|
| 2         | 80.62                | 0.65                 | -29.44               | 0.69                 | 0.69                 | 0.55                 | 0.72                 | 0.70                 | 0.70                 | -53.18               | -53.18               | -53.18               |
| 3         | 79.77                | -0.82                | -31.65               | -0.11                | -0.13                | 24.46                | 0.34                 | 0.31                 | 0.31                 | -54.04               | -54.04               | -54.04               |
| 4         | 99.44                | 0.83                 | -31.95               | 0.70                 | 0.70                 | 15.09                | 1.05                 | 1.03                 | 1.03                 | -34.36               | -34.36               | -34.36               |

| RMSE | | | | | |
|------|----------------------|----------------------|----------------------|----------------------|----------------------|----------------------|----------------------|
| 2    | 109.91               | 3.88                 | 51.70                | 3.88                 | 3.90                 | 45.13                | 3.89                 | 3.90                 | 3.90                 | 53.29                | 53.30                | 53.30                |
| 3    | 109.28               | 3.90                 | 53.80                | 3.82                 | 3.84                 | 52.54                | 3.84                 | 3.86                 | 3.86                 | 54.15                | 54.15                | 54.15                |
| 4    | 124.37               | 3.80                 | 46.46                | 3.77                 | 3.79                 | 47.66                | 3.86                 | 3.87                 | 3.87                 | 34.54                | 34.54                | 34.54                |

| 95% Confidence Interval Coverage | | | | | |
|---------------------------------|----------------------|----------------------|----------------------|----------------------|----------------------|----------------------|----------------------|
| 2                               | 0.860                | 0.068                | 0.993                | 0.069                | 0.938                | 0.941                | 0.995                | 0.089                | 0.939                | 0.937                | 0.000                | 0.000                | 0.000                |
| 3                               | 0.863                | 0.107                | 0.989                | 0.110                | 0.942                | 0.946                | 0.984                | 0.108                | 0.948                | 0.942                | 0.000                | 0.000                | 0.000                |
| 4                               | 0.785                | 0.065                | 0.997                | 0.055                | 0.936                | 0.942                | 0.993                | 0.069                | 0.936                | 0.935                | 0.000                | 0.000                | 0.000                |

<sup>a</sup>Areas shaded in grey represent the proposed methods in this dissertation

<sup>b</sup>Abbreviations: FE=fixed effects, AIPW=augmented inverse probability weighted estimator, BS=bootstrap, SE=standard error, perc=percentile, sand=sandwich
### TABLE XXXVIII

**SIMULATION RESULTS IN 3 TYPES OF LARGE AGE COVARIATE CONFOUNDING FOR TOTAL CHARGES**

<table>
<thead>
<tr>
<th>Setting</th>
<th>Two-way FE&lt;sup&gt;b&lt;/sup&gt;</th>
<th>Separate time weighted</th>
<th>Four-group weighted</th>
<th>AIPW&lt;sup&gt;b&lt;/sup&gt;</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>(4b): incl. X, basic BS SE&lt;sup&gt;a,b&lt;/sup&gt;</td>
<td>(4c): incl. X, perc BS SE&lt;sup&gt;a,b&lt;/sup&gt;</td>
</tr>
<tr>
<td>Mean Bias</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>77.39</td>
<td>0.41</td>
<td>-28.89</td>
<td>0.62</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0.62</td>
</tr>
<tr>
<td>3</td>
<td>78.78</td>
<td>-3.03</td>
<td>-38.91</td>
<td>-1.31</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>-1.35</td>
</tr>
<tr>
<td>4</td>
<td>108.92</td>
<td>40.52</td>
<td>38.31</td>
<td>40.68</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>40.68</td>
</tr>
<tr>
<td>RMSE</td>
<td></td>
<td></td>
<td>67.30</td>
<td>3.88</td>
</tr>
<tr>
<td>2</td>
<td>107.56</td>
<td>3.87</td>
<td>67.30</td>
<td>3.88</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>3.90</td>
</tr>
<tr>
<td>3</td>
<td>108.57</td>
<td>4.88</td>
<td>71.93</td>
<td>4.04</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>4.08</td>
</tr>
<tr>
<td>4</td>
<td>132.07</td>
<td>40.71</td>
<td>51.58</td>
<td>40.86</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>40.86</td>
</tr>
<tr>
<td>95% Confidence Interval Coverage</td>
<td></td>
<td></td>
<td>61.00</td>
<td>3.92</td>
</tr>
<tr>
<td>2</td>
<td>0.865</td>
<td>0.118</td>
<td>0.956</td>
<td>0.102</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0.936</td>
</tr>
<tr>
<td>3</td>
<td>0.863</td>
<td>0.126</td>
<td>0.937</td>
<td>0.127</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0.936</td>
</tr>
<tr>
<td>4</td>
<td>0.753</td>
<td>0.000</td>
<td>1.000</td>
<td>0.000</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0.000</td>
</tr>
</tbody>
</table>

<sup>a</sup>areas shaded in grey represent the proposed methods in this dissertation

<sup>b</sup>abbreviations: FE=fixed effects, AIPW=augmented inverse probability weighted estimator, BS=bootstrap, SE=standard error, perc=percentile, sand=sandwich
Figure 32. Coverage results under small poverty covariate confounding for total charges

1. Naive
2. Naive with X
3. Sep time wts
4a. Sep time wts+X
4b. Sep time wts+X. basic bs
4c. Sep time wts+X. perc bs
5. 4-grp wts
6a. 4-grp wts+X
6b. 4-grp wts+X. basic bs
6c. 4-grp wts+X. perc bs
7a. AIPW. sand
7b. AIPW. bs basic
7c. AIPW. bs perc

*a abbreviations: sep=separate, wts=weights, grp=group, perc=percentile, bs=bootstrap, sand=sandwich
Figure 33. Coverage results under medium poverty covariate confounding for total charges

1. Naive
2. Naive with X
3. Sep time wts
4a. Sep time wts+X
4b. Sep time wts+X. basic bs
4c. Sep time wts+X. perc bs
5. 4-grp wts
6a. 4-grp wts+X
6b. 4-grp wts+X. basic bs
6c. 4-grp wts+X. perc bs
7a. AIPW. sand
7b. AIPW. bs basic
7c. AIPW. bs perc

▲ Setting 2 △ Setting 3 □ Setting 4

*aabbreviations: sep=separate, wts=weights, grp=group, perc=percentile, bs=bootstrap, sand=sandwich
Figure 34. Coverage results under large poverty covariate confounding for total charges

1. Naive
2. Naive with X
3. Sep time wts
4a. Sep time wts+X
4b. Sep time wts+X. basic bs
4c. Sep time wts+X. perc bs
5. 4-grp wts
6a. 4-grp wts+X
6b. 4-grp wts+X. basic bs
6c. 4-grp wts+X. perc bs
7a. AIPW. sand
7b. AIPW. bs basic
7c. AIPW. bs perc

*abbreviations: sep=separate, wts=weights, grp=group, perc=percentile, bs=bootstrap, sand=sandwich*
Figure 35. Coverage results under small age covariate confounding for total charges

1. Naive
2. Naive with X
3. Sep time wts
4a. Sep time wts+X
4b. Sep time wts+X. basic bs
4c. Sep time wts+X. perc bs
5. 4-grp wts
6a. 4-grp wts+X
6b. 4-grp wts+X. basic bs
6c. 4-grp wts+X. perc bs
7a. AIPW. sand
7b. AIPW. bs basic
7c. AIPW. bs perc

*abbreviations: sep=separate, wts=weights, grp=group, perc=percentile, bs=bootstrap, sand=sandwich
Figure 36. Coverage results under medium age covariate confounding for total charges

1. Naive
2. Naive with X
3. Sep time wts
4a. Sep time wts+X
4b. Sep time wts+X. basic bs
4c. Sep time wts+X. perc bs
5. 4-grp wts
6a. 4-grp wts+X
6b. 4-grp wts+X. basic bs
6c. 4-grp wts+X. perc bs
7a. AIPW. sand
7b. AIPW. bs basic
7c. AIPW. bs perc

\[^{a}\text{abbreviations: sep=separate, wts=weights, grp=group, perc=percentile, bs=bootstrap, sand=sandwich}\]
Figure 37. Coverage results under large age covariate confoundig for total charges

1. Naive
2. Naive with X
3. Sep time wts
4a. Sep time wts+X
4b. Sep time wts+X, basic bs
4c. Sep time wts+X, perc bs
5. 4-grp wts
6a. 4-grp wts+X
6b. 4-grp wts+X, basic bs
6c. 4-grp wts+X, perc bs
7a. AIPW sand
7b. AIPW bs basic
7c. AIPW bs perc

a abbreviations: sep=separate, wts=weights, grp=group, perc=percentile, bs=bootstrap, sand=sandwich
weighting reduced the variance by one-third in models involving the count outcome \( RXTOT \). The reduction in variance lead to an increase in the coverage, though not enough to reach desired 0.95 levels. In models involving the outcome \( TOTCH \), the variance was not reduced. As the bias and variance remained similar to the two-way fixed effect model, there was little change to the coverage. Separate time weighted estimators did not improve bias as was expected, but did reduce the variance in some cases. Overall, separate time weighting with model-based standard errors delivered poor coverage rates.

The four-group weighed estimator exhibited mixed results in terms of accuracy compared to the separate time weighted and naïve two-way fixed effect estimators. In scenarios involving a confounded poverty variable, the four-group weighted estimator demonstrated superior accuracy compared to other estimators. However, when confounding was introduced through age, while still at acceptable levels, accuracy diminished relative to other estimators. Coverage was extremely poor across all settings, and in all scenarios. While accuracy was good, the model could not come across a reliable standard error to result in an acceptable coverage rate. While the separate time weighted estimator with model-based standard errors delivered good accuracy, coverage was terrible.

Bootstrap standard errors provided the much-needed improvement to the coverage from model-based methods. Bootstrap standard errors performed better with the continuous outcome, total charges. Under scenarios of varying effect size, weighted estimator type, and setting, 88.8% (71/80) of coverage rates in estimators using bootstrap standard errors (estimators 4b, 4c, 6b, and 6c) were good (0.936-0.948). This is excellent considering only 3% (3/120) of model-based standard errors (estimators 1, 2, 3, 4a, 5, and 6a) were good for the same outcome. In models involving prescription count, gains from bootstrap standard errors were less pronounced. In the
separate time weighted estimator, 10% (5/40) of coverage rates were good. In the four-group weighted estimator with confounded poverty, 90% (18/20) of coverage rates were good, compared to 10% (2/20) of coverage rates with confounded age. These rates are still an improvement to estimators using model-based standard errors, of which none (0/120) of the scenarios exhibited a good coverage rate. Bootstrap standard errors resulted in a dramatic improvement to the coverage from common-practice, model-based methods.

Percentile bootstrap coverage rates edged out basic bootstrap coverage rates in the competition to get closer to nominal levels. Of the 80 scenarios, percentile bootstrap methods had the most precise coverage rate 53% of the time, basic bootstrap methods 36% of the time, and it was a tie 11% of the time. While percentile bootstrap rates had the greater proportion of precision wins, coverage rates between the two methods were quite similar. From a gross perspective, coverage from percentile and bootstrap methods were comparable.

The AIPW estimator exhibited erratic results in the simulation. In models involving the count outcome number of prescriptions, measures of accuracy were similar to the separate time weighted model for all scenarios except in the setting of group by time confounding. In this setting, accuracy began to degrade. In models involving the continuous outcome total charges, accuracy degraded completely. For both outcomes, confidence interval length was similar to well-performing bootstrap estimators (estimators 4b, 4c, 6b, and 6c). However the collapse of accuracy in several scenarios, resulted in the subsequent collapse of coverage.

Despite all available covariate adjustments and weighting techniques, there remained confounding inherent in the data. Proof that there exists hidden confounding can be determined from the imperfect results of the no bias setting (setting 1). In this setting given no existing confounding, estimators should have been spot on in terms of accuracy and coverage. Further,
conditional and marginal estimates should have equated in the case of the continuous outcome, no unmeasured confounding, and knowledge of the true outcome model (Austin, 2011). Both of which did not happen. As the confounding effect size increased, accuracy and precision diminished. Coverage rates even in the best-performing estimators became erratic for large levels of covariate confounding. Confounding still present in the data likely included both a group and time bias. Group bias might arise from individualistic reasons governing the geographic location of individuals. There are several factors which determine where an individual or family chooses to live, and these attributes might be confounders which we have not accounted for in our limited covariate list. A time bias is further likely to exist, especially in the situation of few time points. The introduction of the ACA brought several changes to the healthcare climate. There were changes on the side of individuals seeking care and also on the side of those offering the care. A multitude of possibilities exist for confounding in these situations. Other time confounders could include attributes influenced by political, natural disasters, and employment rates. In nonexperimental designs, it is difficult to remove all sources of confounding entirely. While we did our best to use the most comprehensive set variates and techniques, there still existed unmeasured confounding influencing the results.

It could have been the case that a large proportion of the data was same-person data. We treated groups as independent, or non-same-person data. If there were a lot of individuals in both the pre and post groups, then failing to account for this would have downwardly biased standard error estimates. This would have led to misleadingly narrow confidence intervals. We did observe misleadingly narrow confidence intervals for the outcome TOTTC in the subgroup, but not for the outcome RXTOT in the full sample. Since the subgroup was a distinct subset of the
full sample, it can be determined that the potential issue of non-same-person data driving confidence intervals artificially downward was not an issue in this investigation.

There existed overdispersion in Poisson models. For the prescription count outcome, the variance was much larger than the mean (mean=10.1, variance=204.5). Large overdispersion can lead to grossly deflated standard errors and in turn inflated test statistics (Cameron and Trivedi, 2005). Therefore it was particularly important to use a robust variance estimate. Estimators not using a robust variance estimate actually had a greater standard error which led to overcoverage (estimators 4b and 4c versus 4a, 6b and 6c versus 6a in \(RXTOT\) models). Potential issues from overdispersion were not an issue in prescription count models.

Properties of the confounded variables themselves played a role in the estimator performance. Poverty had more variability relative to its mean compared to age. This relative variability is one-half for poverty (mean=71, SD=36) versus one-third for age (mean=38, SD=12). The greater variability in poverty gave the model more information to work with in outcome explanation. In return, models involving poverty had more stable results between settings in covariate confounding because the attributes inherent to the variables themselves were lending help to the variance explanation.

Proposed empirical standard error techniques performed dramatically better compared to existing methods. In the simulation devised around the MEPS data, it is crystal clear that the proposed variance estimation methods are critical to obtain a good precision.
6. DISCUSSION

The central interest of this dissertation was to identify an improved standard error estimation strategy of a DID using propensity score weighting. Considerations in how propensity score weighting might affect the variance of the treatment effect in DID estimation, before now, have not been investigated to the best of our knowledge. Building upon prior work in DID estimation and propensity score approaches, we have developed a framework for estimating the variance of a DID treatment effect using propensity score weights. In addition, we have proposed a new AIPW DID estimator. The performance of existing and proposed estimators were evaluated under a broad range of settings in covariate confounding plausible to be present in observational data. Techniques presented herein offer a dramatic improvement to the precision of existing estimators.

It is imperative to obtain the correct standard error of a policy effect in order to make informed policy decisions. Precise variance estimation is critically important for produced confidence intervals to be representative of advertised Type I error rates. For example, suppose the existing four-group weighted estimator is applied to an experimental healthcare policy and concludes a significant effect of the policy in terms of a $200 savings per individual at the 0.05 level. When the same four-group weighted estimator is calculated using appropriate bootstrap standard errors, the resulting policy effect becomes non-significant. Per existing methods, policymakers might chose to expand the policy to a larger population from the understanding it would save a significant amount of money. However in actuality, the policy does not save a significant sum of money and potentially wastes resources. Imprecise standard errors in DID estimation can have a broad impact on healthcare policy decisions. While it is important for any candidate estimator to exhibit good accuracy, a precise estimator will produce the most honest \( p \)-
values ultimately contributing to the most informed decisions. Our focus is on an estimation procedure of the standard error. Excellent properties in bias have already been established for existing methods. Existing estimators however do not account for the sampling variability of propensity scores and can in turn exhibit unacceptable coverage rates.

Difference-in-differences standard error estimation in non-same-subject data has largely been overlooked. The samples were thought to be independent and thus, model-based standard errors were deemed appropriate. More specifically, the DID was obtained as the point estimate and associated standard error of the interaction term between post and treatment indicators. With the growing need for marginal effect estimates in DID, propensity score methods have offered hope. In newer propensity score based approaches, researchers continue to use model-based standard errors. However, we have demonstrated that this approach to the variance estimation can be disastrous. Methods using propensity scores in DID estimation to account for distributional differences in time and treatment group are new, and poorly understood. There is limited insight in the literature as to how propensity score weighting methods might affect the standard error of the estimated treatment effect.

Regression adjustment in weighted and unweighted models reduced the variance of the estimate. While it moved coverage rates closer to nominal values in the simulation using artificial data and a Gaussian model, we saw in the simulation devised around real data that the effects to the variance can be erratic and extreme. Regression adjustment should theoretically improve the precision in most cases. But when covariates are strong predictors of the outcome, or worse, strong predictors of treatment, overconfidence can lead to an over shrinking of the variance estimates. Consistent with both outcome models regardless of propensity score
weighting, was the unacceptable job of covariate adjustment alone in obtaining an adequate coverage rate.

The separate time weighted estimator was successful in removing bias in the simulation study involving artificial data. However, in the simulation study devised around real data, improvements to the bias from separate time weighting were minimal and limited to specific settings in covariate confounding. The greatest improvements to the bias were in the presence of a time bias. This was especially pronounced when the time bias was introduced independently as a time and group bias, as opposed to when it was introduced dependently as a time by group bias. In the simulation using artificial data, coverage rates were improved due to better accuracy. However, in the simulation devised around real data, while there was a reduction to the variance for some models, this reduction was not enough and coverage rates remained unacceptable. Model-based standard errors in the separate time weighted estimator did not deliver good coverage.

The four-group weighted estimator delivered excellent accuracy in both simulation studies. The estimator particularly shined in the setting of a group-only bias. While coverage was good for some settings in the simulation using artificial data, coverage results completely deteriorated in the simulation devised around real data. The greatest improvements to the bias caused the variance to shoot up, resulting in similar coverage rates to the separate time weighted estimator. Model-based standard errors in the four-group weighted model should be avoided.

Regression adjustment can reduce the variance of the treatment effect estimate, however bias might arise when the model is misspecified or fails to include needed nonlinear or interaction terms (Rubin 1973). On the other hand, propensity score techniques can reduce the bias by balancing covariates. However, this improvement to the bias comes at the sacrifice of an
inflated variance. There are two competing mechanisms at work, both embodying the bias-
variance tradeoff in opposite directions. Golinelli and colleagues (2012) recommend allowing
regression adjustment to resolve minor covariate imbalances after weighting instead of
aggressively pursuing balance. Combining regression adjustment and propensity scores in a
parsimonious weighted regression model offers the opportunity to improve both accuracy and
precision. It also opens the door to dangerous pitfalls when methods are applied haphazardly
without care.

Currently there is not an expression which allows us to directly estimate the variance of
an IPW DID. Contemporary methods use the standard error of the interaction term in a
propensity-weighted regression. What is incorrect about this method is that it treats the weights
as known values, when in fact they are not. The weights directly depend on the estimated
propensity scores. Propensity scores introduce at least two additional components of variability:
(1) variability in the estimation of propensity scores and (2) variability in the formation of
propensity score weights. Resampling the weighted observations allows us to capture the
sampling distribution of the weighted observations. The resamples simulate the sampling
distribution of the weighted observations, which in turn allows the analyst to capture both
sources of unknown variability without estimating directly. Bootstrap standard errors offers the
opportunity to bypass the uncertainty introduced by the propensity score weights.

Utilizing propensity score weighting introduces additional sources of variability to the
estimand. Variability introduced by the propensity score weights alters the effective sample size
(ESS). The variance of a weighted mean is $\sigma^2/ESS$ where $ESS = n/(1 + Var(w))$. For weights
equal to 1, $Var(w) = 0$ and the ESS will equal and actual sample size. Here the variance of the
weighed mean is simply the usual variance estimate, $\sigma^2/n$. When using propensity score
weights, $Var(w) > 0$, and the ESS is reduced. This causes the variance of the weighted mean to increase. The variance of the propensity score weights exactly affects the weighted mean.

The proposed bootstrap procedures offer a substantial improvement to the coverage of contemporary methods. In both simulation experiments, empirical standard errors produced the much-needed refinement to model-based standard errors. Our hypothesis that model-based standard errors would yield too wide confidence intervals did not hold up. We observed the changes to coverage after incorporating propensity score weights were colorful functions involving multifaceted aspects of the propensity score sampling variability. There does not yet exist a closed-form expression for the variance of a propensity-weighted DID estimator to account for this colorful dynamic. The two-part sampling variability of the propensity score weights are then used as an input into an additional layer of sampling variability in the DID effect estimate. Of course sampling from the weighted distributions will still not yield estimates with perfect coverage. There remain aspects to the sampling distribution of the DID effect estimate, some known and some unknown. Resampling the weighted observations can bring us closer.

Basic and percentile bootstrap methods were comparable in the simulation studies. In the future use of these estimators, choice of bootstrap estimation procedure could be guided by knowledge of the effect estimate or model specification. The basic bootstrap method is a better choice if the analyst suspects an asymmetric sampling distribution of the DID. If it is a transformation of the DID that is desired, the percentile bootstrap method is a better choice as it is transformation respecting while the basic bootstrap method is not. The bias-corrected and accelerated bootstrap method offers the opportunity to mitigate the symmetry short-coming of the percentile method, however the computation can be torturous (Carpenter and Bithell, 2000).
This method could be investigated in the future to determine if the refinement to the coverage is worth the computational cost.

In comparing propensity score weights between the separate time and four-group weighted estimators, the weights from the separate time estimator exhibited greater variability. The difference in weight variability was completely attributed to the weights of the treatment group. The treated group in the pre period of the four-group weighted model all had a weight of one, and thus no variability. The treated group in the post period of the four-group weighted estimator similarly had small variability of weights. The weights from the separate time weighted model had greater variability and thus experienced a greater reduction to the ESS. This contributed to a larger variance of the separate time weighted estimates compared to the four-group weighted estimates. The four-group weights are advantageous by their limitation to the variance of the formed weights, which in turn limits the reduction to the ESS.

The current AIPW estimator performed well under a fully correct model specification. This led to better results in the simulation using artificial data, where variates were generated in a simple manner and models were specified more closely to the underlying mechanism. Of course in the real world we never know the true model specification. Accuracy of the AIPW completely degraded in the simulation devised around real data. The estimator continued to produce confidence intervals of similar length to well-performing estimators. Sadly, the collapse in accuracy resulted in the decimation of coverage. If accuracy was able to be improved in this estimator, it would deliver coverage estimates in the ballpark of nominal levels, perhaps even in the center bullseye. The AIPW DID estimator holds great potential if a refinement to the accuracy can be found.
A potential criticism might be that scenarios harnessed in these simulations do not reflect actual scenarios exhibited in real life. More specifically, simulations in this study might not be sufficiently complex to address the true underlying mechanisms and trends of the MEPS data. While real data are nearly always richer than artificial data, we have tried several approaches and given thoughtful consideration into the construction of the data generation procedures. Procedures were selected to represent those similar to the underlying mechanism. In final, we were able to invent a mechanism of data generation, producing datasets sufficiently similar on average, in marginal and associational relationships, as well as on observable trends to the MEPS data. We argue these simulated datasets in conjunction with the comprehensive simulation experiment contribute to the credibility and generalizability of the conclusions. We acknowledge that simulations performed herein correspond to a small subset of situations and at worst, are artificially constructed and superficially contrived. A simulation cannot substitute for a theoretical proof, making the findings here limited to the simulations performed. There is always a chance with simulations that the findings could be spurious and specific to the scenarios produced. We hope, and argue, that estimators which consistently perform well in a myriad of scenarios in covariate confounding will continue to perform well in new scenarios.

This dissertation is not strongly theoretical. Hypotheses were based on large sample properties from other IPW estimators including a single difference (Lunceford and Davidian, 2004), risk difference, and risk ratio (Williamson et al., 2012). The variance estimation in more complex IPW estimators is non-trivial. A simulation study offers the opportunity to assess estimator performance when an expression for the variance does not exist. Future work could include the theoretical derivation of a large-sample IPW DID marginal variance estimator.

A limited number of settings in covariate confounding were examined for two continuous
variables. Estimator performance might vary in cases of dichotomous and ordinal variable confounding. A beta-binomial regression might be considered in this type of variable generation where the properties of a beta distribution could be manipulated to generate different properties of discrete variables. Furthermore, introduction of covariate confounding could be incorporated into the simulation process, instead of introduced after. Such methods would need to be carefully constructed to preserve the marginal and associational relationships between variables. Lastly, we barely scratched the surface in the setting of group by time covariate confounding. We examined the situation where the starting value of the treatment group begins larger than the control group and the trajectories of post estimates cross paths. It is possible the control group starts larger than the treatment group, and trajectories cross paths. Additionally, trends of both groups could increase or decrease together without crossing paths. There are a total of six general scenarios in group by time covariate confounding which involve group starting levels their rate of change. Future studies could examine these alternate scenarios in varying magnitude. Presented results only represent one such case and do not provide a full picture of this setting. Estimator performance may vary in alternative scenarios of group by time confounding. Future studies should examine estimator performance under more scenarios in covariate confounding.

Robustness of estimation methods was not examined critically in this investigation. In the simulation devised around the MEPS data, the increase in confounding effect size can be viewed as a partial investigation of robustness. With increasing effect size, the models move further away from correct specification. For instance, the effect of interaction terms involving the confounded variable, treatment indicator, and post indicator become more important. The models represented in the estimators are inherently misspecified by not including these interaction terms. A well informed analyst who suspects a certain type of bias may include some of the interactions
in the model, though we consider the situation where models are inherently misspecified. In the future, robustness could be further tested by gauging estimator performance under additional forms of outcome model and propensity score model specification.

Despite these limitations, this dissertation has several strengths. Its strengths include an application to a real world dataset, simulation using artificial data, and a rigorous, comprehensive simulation study devised around the real data example. We have introduced two new approaches in the standard error estimation of a DID as well as a new estimator with three variations in standard error estimation. Most importantly, it is practical, useful, and broadly applicable. This dissertation contributes to the state of the art of how propensity score weighting in regression models affects the variance of an estimated a treatment effect. This is an area desperate for attention as applied researchers are increasingly employing these methods with little guidance of how these combined approaches might alter produced estimates.

Future research should include detailed examination of extreme propensity scores and propensity score models specification. Estimator stability might be sensitive to extreme propensity scores and associated weights. Extreme propensity score weights can contribute to bias and lead to an artificially inflated variance. In turn, this might have altered results and conclusions of weighted estimators. Propensity score weights can be especially problematic in situations where the propensity score model is misspecified. Extreme weights are not the main issue in comparison to propensity score models "missing important interactions and non-linearities in pre-treatment covariates” which can lead to large bias (Lee et al. 2011). Lee and colleagues (2011) recommend moving the focus away from propensity score trimming into correct specification of the propensity score model. The sensitivity of estimator performance
under propensity score trimming and alternate specifications of the propensity score model could be further explored.

Estimator coverage should be further evaluated under the case of no bias. Current coverage results rely on the DID point estimate to provide the location of the confidence interval. By setting the point estimate to the true value (zero bias situation), new coverage results would more purely reflect variance performance. For example, a shift in confidence interval location might dramatically improve the coverage of the AIPW in cases where it exhibited detrimentally large bias. As the main focus of this dissertation is an improved variance estimate of the DID, an alternate examination of coverage, non-dependent on an estimators' bias, could permit a more direct comparison of competing variance estimands.

While methods in DID estimation and bootstrap alone are not novel, the use of bootstrap in these DID estimators is, to the best of our knowledge. This study is the first of its kind to examine the precision in DID estimation with propensity score weighting. Our study extends propensity score weighting strategies in DID estimation. Our study further corroborates with previous work by Stuart et al. (2014) and Song et al. (2012) by underlining the critical use of propensity score techniques for accuracy refinement in DID.

Standard error estimation in DID estimators is not limited to healthcare policy estimation. It can be translated to virtually any pre-post/treatment-control scenario or any model including an interaction term for that matter. Results contained can be immediately used by individuals in academia, government, applied research, methodology, pharmaceutical industry, health policy administration, medical science, and the social sciences. The included code supplement in the Appendix ensures the proposed techniques are readily accessible to researchers. Even if DID estimation is not exactly what one is after, there are likely overlapping aspects to a diverse range
of research interests. For instance, any investigation which includes an interaction term in a regression model might find this research useful. Not limited to interaction terms in regression models, we have further generated thought regarding any regression modeling incorporating propensity score weights. Our results demonstrate that the analyst should exercise skepticism when using standard errors pulled directly from the regression model.

The separate time weighted estimator delivers better accuracy in settings of a group and time or group by time bias. The four-group weighted estimator works better in the setting of a group-only bias. Both estimators will perform well given they incorporate the proposed bootstrap standard error methods. Usage of existing methods using model-based standard errors will result in disastrous precision. Selection of an estimator should be determined by the suspected bias inherent to the data. If the analyst is unsure of the type of bias contained in the data, the news is still descent. Either estimator using the proposed variance estimation method will deliver acceptable accuracy and precision.

Difference-in-differences and propensity scores can be incorporated to determine the marginal effect of a policy in observational data. Contemporary methods fail to deliver a precise variance estimate of a DID due to unaccounted for sampling variability of the propensity score weights. We have proposed a way which corrects the precision issue in leading propensity score weighted DID estimators. This correction offers the much needed refinement to the precision while retaining the bias improvements made from leading estimators. Bootstrap standard errors in DID regression modeling using propensity score weights are necessary for obtaining a precise estimate.
7. SUMMARY OF CONTRIBUTIONS

Gauging the effect of a healthcare policy in terms of how much it costs the individual, how much it costs the state, or how many prescriptions the individual was able to have filled due to the policy, for example, is vital for guiding the future direction of a healthcare system.

Advances in econometric and biostatistical fields have yielded tools to measure a policy’s effect and mitigate a deficiency that arises from the measurement of this effect, respectively. Difference-in-differences is an approach from econometrics used to determine the effect of a policy, and propensity score weighting is an approach from biostatistics which can be used to put policy-affected and policy-unaffected groups on a level playing field to arrive at a DID estimate closer to the truth. However, while propensity scores can bring an estimate closer to the real answer (better accuracy), a tradeoff can be the imprecise alteration of the closeness of these estimates in repeats of the experiment (worsened precision). An approach could be to obtain the uncertainty of the DID estimate by resampling the propensity-weighted observations.

Combining DID and propensity score methods could allow researchers to optimize these competing forces to accurately and precisely estimate the effect of a healthcare policy. While this doubling-down approach could increase the efficiency of an estimator, it conversely has the ability to deliver haphazard results if not applied properly.

There currently exists limited insight in the literature as to how these two techniques work in conjunction. In this dissertation, we have:

- Presented a background on these separate approaches
- Identified the problem with their conjunct use in current research
- Determined the need for a more precise estimation method in this broadly applicable problem
• Developed a framework for precisely estimating a DID treatment effect using propensity score weights while maintaining the strides made in accuracy from existing methods

• Proposed a new AIPW DID estimator

• Applied the proposed methods to a real world example estimating the effect of Medicaid expansion on medication acquisition, medication cost, and total medical charges

• Conducted a motivating simulation study using artificially simulated data which revealed the great potential of proposed techniques

• Conducted a large-scale simulation study which further corroborated the excellent performance of the proposed methods compared to contemporary methods

• Presented insight into the inner mechanisms as to why these proposed techniques are effective

Estimating costs and utilization attributable to a healthcare policy are important for policymakers to enact policies that are feasible, valuable, and affordable. With United States’ healthcare costs surging to prohibitively high levels along with the dramatic overhaul of the healthcare system, techniques herein have never been more crucial.
CITED LITERATURE


Wen, H., Borders, T. F., Druss, B. G. (2016). Number of Medicaid prescriptions grew, drug spending was steady in Medicaid expansion states. *Health Affairs, 35*(9), 1604-1607.


APPENDICES
APPENDIX A

Difference-in-differences estimators: linear model

# program name: MEPS linear.R
# purpose: multivariate simulation of MEPS data, evaluation of 13
# DID estimators using linear models

#library(moments)
library(modmarg)
#library(corrplot)
library(nnet)
library(boot)
library(effsize)
rm(list = ls())

# set working directory
setwd("D:/MEPS")

# load MEPS data
meps <- read.csv(file="data/MEPS_22 Jan 2018.csv")

# Format MEPS dataset
names(meps)[names(meps)%=='trt_NE_S'] <- 't'
names(meps)[names(meps)%=='post']<- 'p'
names(meps)[names(meps)%=='AGE'] <- 'age'
names(meps)[names(meps)%=='POVLEV'] <- 'povlev'
meps$tp <- meps$t * meps$p

# change male to numeric variable
meps$male <- as.numeric(meps$SEX)
# change insurance to numeric variable
meps$ins <- as.numeric(meps$INSCOV)
# format to factor variables
meps$male.f <- factor(meps$male)
# force female to be reference category
meps$male.f <- relevel(meps$male.f, ref = 2)
meps$black.f <- factor(meps$black)
meps$hispanic.f <- factor(meps$hispanic)
meps$edu.f <- factor(meps$edu)
meps$ins.f <- factor(meps$ins)
meps$cc_3.f <- factor(meps$cc_3)
one.96 <- -qnorm(0.025)

# define groups
meps$group[which(meps$t == 1 & meps$p==0)] <- 1
meps$group[which(meps$t == 1 & meps$p==1)] <- 2
meps$group[which(meps$t == 0 & meps$p==0)] <- 3
meps$group[which(meps$t == 0 & meps$p==1)] <- 4

aipw <- function(data, outcome=outcome) {
  treatment.var = "t"
  treatment.vec <- data[, treatment.var]
  treatment.values <- sort(unique(treatment.vec))

  treated.data <- data[treatment.vec == treatment.values[2], ]
  control.data <- data[treatment.vec == treatment.values[1], ]
  n.treated.pre <- nrow(treated.data)
  n.control.pre <- nrow(control.data)

  # propensity score model
  ps <- suppressWarnings(glm(t ~ povlev + age + male.f + black.f + hispanic.f + edu.f +
   ins.f + cc_3.f, family = binomial, data = data))
  pscore.probs <- suppressWarnings(predict(ps, newdata = data, type = "response"))
  pscores.pre <- pscore.probs

  # outcome formula for treated
  outcome.t <- suppressWarnings(glm(y ~ povlev + age + male.f + black.f + hispanic.f + edu.f +
   ins.f + cc_3.f, data = treated.data))
  outcome.c <- suppressWarnings(glm(y ~ povlev + age + male.f + black.f + hispanic.f + edu.f +
   ins.f + cc_3.f, data = control.data))

  outcome.treated.expectation <- suppressWarnings(predict(outcome.t, newdata = data, type =
    "response"))
  outcome.control.expectation <- suppressWarnings(predict(outcome.c, newdata = data, type =
    "response"))
APPENDIX A (continued)

```r
res.c <- suppressWarnings(residuals(outcome.c, type = "response"))
res.t <- suppressWarnings(residuals(outcome.t, type = "response"))

n <- length(treatment.vec)
term1 <- rep(0, n)
term2 <- rep(0, n)
term3 <- rep(0, n)
term4 <- rep(0, n)
control.indic <- treatment.vec == treatment.values[1]
treated.indic <- treatment.vec == treatment.values[2]
outcome.vec <- model.response(model.frame(y ~ povlev + age + male.f + black.f + hispanic.f + edu.f + ins.f + cc_3.f, data = data))

term1 <- outcome.vec/pscore.probs
term1[control.indic] <- 0
term2 <- ((1 - pscore.probs) * outcome.treated.expectation)/pscore.probs
term2[control.indic] <- ((0 - pscore.probs[control.indic]) * outcome.treated.expectation[control.indic]) / pscore.probs[control.indic]
term3 <- outcome.vec/(1 - pscore.probs)
term3[treated.indic] <- 0
term4 <- ((1 - pscore.probs) * outcome.control.expectation)/(1 - pscore.probs)
term4[control.indic] <- ((0 - pscore.probs[control.indic]) * outcome.control.expectation[control.indic]) / (1 - pscore.probs[control.indic])

ATE.AIPW.hat <- (1/n) * (sum(term1) - sum(term2) - sum(term3) - sum(term4))

I.hat <- term1 - term2 - term3 - term4 - ATE.AIPW.hat
ATE.AIPW.sand.var <- (1/n^2) * sum(I.hat^2)
ATE.AIPW.sand.se <- sqrt(ATE.AIPW.sand.var)
ATE.reg.hat <- mean(outcome.treated.expectation) - mean(outcome.control.expectation)
weight.norm.1 <- 1/sum(treated.indic/pscore.probs)
weight.norm.3 <- 1/sum((1 - treated.indic)/(1 - pscore.probs))
ATE.IPW.hat <- weight.norm.1 * sum(term1) - weight.norm.3 * sum(term3)

return(list(ATE.AIPW.hat = ATE.AIPW.hat, ATE.reg.hat = ATE.reg.hat, ATE.IPW.hat = ATE.IPW.hat, ATE.AIPW.sand.SE = ATE.AIPW.sand.se))
```
# seed for bootstrap estimates reproducibility
seed <- 109876
set.seed(seed)

#########################################################################
# Model 1:  Naive FE
# Model 2:  Naive FE with X
# Model 3:  FE weighted separately at each time point
# Model 4a: FE weighted separately at each time point + X
# Model 4b: FE weighted separately at each time point + X. bootstrap basic
# Model 4c: FE weighted separately at each time point + X. bootstrap perc
# Model 5:  FE Stuart's 4-group weighted design
# Model 6a: FE Stuart's 4-group weighted design + X
# Model 6b: FE Stuart's 4-group weighted design + X. bootstrap basic
# Model 6c: FE Stuart's 4-group weighted design + X. bootstrap perc
# Model 7a: AIPW. sandwich
# Model 7b: AIPW. bootstrap basic
# Model 7c: AIPW. bootstrap perc
#########################################################################

meps.est <- function(outcome, n.boot) {
  # set timer
  ptm <- proc.time()

  results <- matrix(NA, nrow=13, ncol=3)
  rownames(results) <- c("1. Naive", "2. Naive with X",
                        "3. Sep time wts", "4a. Sep time wts+X. basic bs",
                        "4c. Sep time wts+X. perc bs",
                        "5. 4-grp wts", "6a. 4-grp wts+X",
                        "6b. 4-grp wts+X. basic bs", "6c. 4-grp wts+X. perc bs",
                        "7a. AIPW. sand", "7b. AIPW. bs basic",
                        "7c. AIPW. bs perc")
  colnames(results) <- c("DID", "cl.l", "cl.u")

  # model formula: y ~ t + p + tp + X
  fmla <- as.formula(paste("y ~ t + p + tp + povlev + age + male.f + black.f + hispanic.f + edu.f +
                           ins.f + cc_3.f"))
  # propensity score model: t ~ X
APPENDIX A (continued)

```r
fmla.ps <- as.formula(paste("t ~ povlev + age + male.f + black.f + hispanic.f + edu.f + ins.f + cc_3.f"))

# outcomes:
# number of prescriptions
if(outcome=="RXTOT") meps$y <- meps$RXTOT

# cost of Medicaid prescriptions
# add small positive value to y for gamma model
if(outcome=="RXMCD") meps$y <- meps$RXMCD + 15

# total charges excluding prescriptions between $1-9,999 for those without private insurance
if(outcome=="TOTTCH") {
  meps <- meps[which(meps$TOTTCH > 0 & meps$TOTTCH < 10000 & meps$ins.f != 1),]
  meps$y <- meps$TOTTCH
}

# prepare data
predata <- meps[which(meps$p==0),]
postdata <- meps[which(meps$p==1),]
wts1 <- rep(1,6830)  # for marginal effect estimates of weighted estimators

# Model 1: Two-way fixed effects
modell <- glm(y ~ t + p + tp, data=meps)
marg.modell <- marg(modell, var_interest = 'tp', type = 'effects')
results[1,1] <- round(marg.modell[[1]][2,2],2)
results[1,2] <- round(marg.modell[[1]][[6]][2,2],2)
results[1,3] <- round(marg.modell[[1]][[7]][2,2],2)

# Model 2: Two-way fixed effects + X
model2 <- glm(fmla, data=meps)
marg.model2 <- marg(model2, var_interest = 'tp', type = 'effects')
results[2,1] <- round(marg.model2[[1]][2,2],2)
results[2,2] <- round(marg.model2[[1]][[6]][2,2],2)
results[2,3] <- round(marg.model2[[1]][[7]][2,2],2)
```
# Models 3, 4: Separate time weights
# propensity score model
prelogit <- glm(fmla.ps, data=predata, family="binomial")
postlogit <- glm(fmla.ps, data=postdata, family="binomial")
pre.pscores <- predict(prelogit, type="response")
post.pscores <- predict(postlogit, type="response")

# generate weights
meps$wts.sep[meps$group==1] <- 1/pre.pscores[predata$t==1]
meps$wts.sep[meps$group==2] <- 1/post.pscores[postdata$t==1]
meps$wts.sep[meps$group==3] <- 1/(1-pre.pscores[predata$t==0])
meps$wts.sep[meps$group==4] <- 1/(1-post.pscores[postdata$t==0])

# Model 3: Separate time weights
model3 <- glm(y ~ t + p + tp, weights=wts.sep, data=meps)
marg.model3 <- suppressWarnings(marg(model3, var_interest = 'tp', type = 'effects', weights = wts1))
results[3,1] <- round(marg.model3[[1]][2,2],2)
results[3,2] <- round(marg.model3[[1]][6][2],2)
results[3,3] <- round(marg.model3[[1]][7][2],2)

# Model 4a: Separate time weights + X
model4 <- glm(fmla, weights=wts.sep, data=meps)
marg.model4 <- suppressWarnings(marg(model4, var_interest = 'tp', type = 'effects', weights = wts1))
results[4,1] <- round(marg.model4[[1]][2,2],2)
results[4,2] <- round(marg.model4[[1]][6][2],2)
results[4,3] <- round(marg.model4[[1]][7][2],2)

# Model 4b, 4c: Separate time weights + X, bootstrap
sep.time.wtd <- function(dat, ind) {
  d <- dat[ind,]  # allows boot to select sample
  model56 <- glm(fmla, weights=wts.sep, data=d)
  suppressWarnings(marg(model56, var_interest = 'tp', type = 'effects', weights = wts1))[[1]][2,2]
}
boot.sep.time.wtd <- boot(meps, sep.time.wtd, R=n.boot)
est <- median(boot.sep.time.wtd$t)
q.025 <- quantile(boot.sep.time.wtd$t, c(0.025, 0.975))[1]
q.975 <- quantile(boot.sep.time.wtd$t, c(0.025, 0.975))[2]

results[5,1] <- results[6,1] <- round(est,2)
# basic bootstrap
results[5,2] <- round(2*est - q.975,2)
results[5,3] <- round(2*est - q.025,2)
# percentile bootstrap
results[6,2] <- round(q.025,2)
results[6,3] <- round(q.975,2)

# Model 5: 4-group weighted
# propensity score model
mlogit <- multinom(group ~ povlev + age + male.f + black.f + hispanic.f + edu.f + ins.f + cc_3.f, data=meps)  
probs <- predict(mlogit, type="probs")

# generate weights
meps$wts.4grp[meps$group==1] <- probs[meps$group==1,1]/probs[meps$group==1,1]
meps$wts.4grp[meps$group==2] <- probs[meps$group==2,1]/probs[meps$group==2,2]
meps$wts.4grp[meps$group==3] <- probs[meps$group==3,1]/probs[meps$group==3,3]
meps$wts.4grp[meps$group==4] <- probs[meps$group==4,1]/probs[meps$group==4,4]

model7 <- glm(y ~ t + p + tp, weights=wts.4grp, data=meps)
marg.model7 <- suppressWarnings(marg(model7, var_interest = 'tp', type = 'effects'))
.results[7,1] <- round(marg.model7[1][2,2],2)
results[7,2] <- round(marg.model7[1][6][2],2)
results[7,3] <- round(marg.model7[1][7][2],2)

# Model 6a: 4-group weighted + X
model8 <- glm(fmla, weights=wts.4grp, data=meps)
marg.model8 <- suppressWarnings(marg(model8, var_interest = 'tp', type = 'effects', weights = wts1))
results[8,1] <- round(marg.model8[1][2,2],2)
results[8,2] <- round(marg.model8[1][6][2],2)
results[8,3] <- round(marg.model8[1][7][2],2)
APPENDIX A (continued)

# Model 6b, c: 4-group weighted + X, bootstrap
four.grp.wtd <- function(dat, ind) {
  d <- dat[ind,] # allows boot to select sample

  model910 <- glm(fmla, weights=wts.4grp, data=d)
  suppressWarnings(marg(model910, var_interest = 'tp', type = 'effects', weights = wts1))[[1]][2,2]
}

boot.four.grp.wtd <- boot(meps, four.grp.wtd, R=n.boot)

est <- median(boot.four.grp.wtd$t)
q.025 <- quantile(boot.four.grp.wtd$t, c(0.025, 0.975))[[1]]
q.975 <- quantile(boot.four.grp.wtd$t, c(0.025, 0.975))[[2]]

results[9,1] <- results[10,1] <- round(est,2)
# basic bootstrap
results[9,2] <- round(2*est - q.975,2)
results[9,3] <- round(2*est - q.025,2)
# percentile bootstrap
results[10,2] <- round(q.025,2)
results[10,3] <- round(q.975,2)

# Model 7a: AIPW, sandwich standard error
aipw.pre <- aipw(predata, outcome=outcome)
aipw.post <- aipw(postdata, outcome=outcome)
est <- aipw.post$ATE.AIPW.hat - aipw.pre$ATE.AIPW.hat
# sandwich standard error
sand.SE <- sqrt((aipw.pre$ATE.AIPW.sand.SE)^2 + (aipw.post$ATE.AIPW.sand.SE)^2)

results[11,1] <- round(est,2)
results[11,2] <- round(est - one.96*sand.SE,2)
results[11,3] <- round(est + one.96*sand.SE,2)

# Model 7b, c: AIPW, bootstrap
aipw.diff <- rep(NA, n.boot)
for (biter in 1:n.boot) {
  boot.inds.pre <- sample(1:dim(predata)[1], 3552, replace = TRUE)
APPENDIX A (continued)

d.pre <- pred data[boot.inds.pre, ]
boot.inds.post <- sample(1:dim(post data)[1], 3278, replace = TRUE)
d.post <- post data[boot.inds.post, ]
# aipw did= aipw.post - aipw_pre
aipw.diff[biter] <- aipw(data = d.post, outcome=outcome)$ATE.AIPW.hat -
                     aipw(data = d.pre, outcome=outcome)$ATE.AIPW.hat
}
est <- median(aipw.diff)
# take 2.5 and 97.5th percentile of bootstrap estimates
q.025 <- quantile(aipw.diff,c(0.025, 0.975))[1]
q.975 <- quantile(aipw.diff,c(0.025, 0.975))[2]

results[12,1] <- results[13,1] <- round(est,2)
# basic bootstrap
results[12,2] <- round(2*est - q.975,2)
results[12,3] <- round(2*est - q.025,2)
# percentile bootstrap
results[13,2] <- round(q.025,2)
results[13,3] <- round(q.975,2)

# end timer
run.time.all <- proc.time() - ptm
run.time <- round(run.time.all[3] / 60,2)

# output system time and simulation settings
params <- data.frame(outcome, n.boot, seed, run.time)
return(list(results, params, meps))

# Output simulation results to Excel
# Number of prescriptions
meps.results <- meps.est(outcome="RXTOT", n.boot=500)
file.name <- paste("output/MEPS_nontrans_TEST_", format(Sys.time(), "%Y-%m-%d"), "_num_rx.csv", sep = "")
write.table(meps.results[[1]], file = file.name,
eol = "\r",sep = ",",row.names = TRUE)
write.table(meps.results[[2]], file = file.name,
APPENDIX A (continued)

eol = "\r", sep = ",", row.names = FALSE, append=TRUE)

# Cost of Medicaid prescriptions
meps.results <- meps.est(outcome="RXMCD", n.boot=500)
file.name <- paste("output/MEPS_nontrans_TEST", format(Sys.time(), "%Y-%m-%d"), "_cost.csv", sep = "")
write.table(meps.results[[1]], file = file.name,
  eol = "\r", sep = ",", row.names = TRUE)
write.table(meps.results[[2]], file = file.name,
  eol = "\r", sep = ",", row.names = FALSE, append=TRUE)

# Total charges excluding prescriptions between $0-10,000 for those without private insurance
meps.results <- meps.est(outcome="TOTTCH", n.boot=500)
file.name <- paste("output/MEPS_nontrans_TEST", format(Sys.time(), "%Y-%m-%d"), "_totch.csv", sep = "")
write.table(meps.results[[1]], file = file.name,
  eol = "\r", sep = ",", row.names = TRUE)
write.table(meps.results[[2]], file = file.name,
  eol = "\r", sep = ",", row.names = FALSE, append=TRUE)


APPENDIX B

Difference-in-differences estimators: generalized linear model

########################################################################
# program name: MEPS GLM.R
# purpose: multivariate simulation of MEPS data, evaluation of 13
#     DID estimators
########################################################################

library(glm2)
library(modmarg)
library(nnet)
library(boot)

rm(list = ls())

# set working directory
setwd("D:/MEPS")

# load MEPS data
meps <- read.csv(file="data/MEPS_22 Jan 2018.csv")

# Format MEPS dataset
names(meps)[names(meps)=='trt_NE_S'] <- 't'
names(meps)[names(meps)=='post'] <- 'p'
names(meps)[names(meps)=='AGE'] <- 'age'
names(meps)[names(meps)=='POVLEV'] <- 'povlev'
meps$tp <- meps$t * meps$p
# change male to numeric variable
meps$male <- as.numeric(meps$SEX)
# change insurance to numeric variable
meps$ins <- as.numeric(meps$INSCOV)
# format to factor variables
meps$male.f <- factor(meps$male)
# force female to be reference category
meps$male.f <- relevel(meps$male.f, ref = 2)
meps$black.f <- factor(meps$black)
meps$hispanic.f <- factor(meps$hispanic)
meps$edu.f <- factor(meps$edu)
meps$ins.f <- factor(meps$ins)
meps$cc_3.f <- factor(meps$cc_3)
one.96 <- -qnorm(0.025)

# define groups
meps$group[which(meps$t == 1 & meps$p==0)] <- 1
meps$group[which(meps$t == 1 & meps$p==1)] <- 2
meps$group[which(meps$t == 0 & meps$p==0)] <- 3
meps$group[which(meps$t == 0 & meps$p==1)] <- 4

aipw <- function(data, outcome=outcome) {
  treatment.var= "t"
  treatment.vec <- data[, treatment.var]
  treatment.values <- sort(unique(treatment.vec))

  treated.data <- data[treatment.vec == treatment.values[2], ]
  control.data <- data[treatment.vec == treatment.values[1], ]
  n.treated.pre <- nrow(treated.data)
  n.control.pre <- nrow(control.data)

  # propensity score model
  ps <- suppressWarnings(glm(t ~ povlev + age + male.f + black.f + hispanic.f + edu.f +
                              ins.f + cc_3.f, family = binomial, data = data))
  pscore.probs <- suppressWarnings(predict(ps, newdata = data, type = "response"))
  pscores.pre <- pscore.probs

  # outcome formula for treated
  outcome.t <- suppressWarnings(glm(y ~ povlev + age + male.f + black.f + hispanic.f + edu.f +
                                    ins.f + cc_3.f, data = treated.data))
  outcome.c <- suppressWarnings(glm(y ~ povlev + age + male.f + black.f + hispanic.f + edu.f +
                                    ins.f + cc_3.f, data = control.data))

  outcome.treated.expectation <- suppressWarnings(predict(outcome.t, newdata = data, type =
                                                 "response")
  outcome.control.expectation <- suppressWarnings(predict(outcome.c, newdata = data, type =
                                                  "response")
  res.c <- suppressWarnings(residuals(outcome.c, type = "response"))
APPENDIX B (continued)

```r
res.t <- suppressWarnings(residuals(outcome.t, type = "response"))

n <- length(treatment.vec)
term1 <- rep(0, n)
term2 <- rep(0, n)
term3 <- rep(0, n)
term4 <- rep(0, n)
control.indic <- treatment.vec == treatment.values[1]
treated.indic <- treatment.vec == treatment.values[2]
outcome.vec <- model.response(model.frame(y ~ povlev + age + male.f + black.f + hispanic.f + edu.f +
                           ins.f + cc_3.f, data = data))

term1 <- outcome.vec/pscore.probs

term1[control.indic] <- 0

term2 <- ((1 - pscore.probs) * outcome.treated.expectation)/pscore.probs

term2[control.indic] <- ((0 - pscore.probs[control.indic]) *
                           outcome.treated.expectation[control.indic])/pscore.probs[control.indic]

term3 <- outcome.vec/(1 - pscore.probs)

term3[treated.indic] <- 0

term4 <- ((1 - pscore.probs) * outcome.control.expectation)/(1 - pscore.probs)

term4[control.indic] <- ((0 - pscore.probs[control.indic]) *
                           outcome.control.expectation[control.indic])/(1 - pscore.probs[control.indic])

ATE.AIPW.hat <- (1/n) * (sum(term1) - sum(term2) - sum(term3) - sum(term4))

I.hat <- term1 - term2 - term3 - term4 - ATE.AIPW.hat

ATE.AIPW.sand.var <- (1/n^2) * sum(I.hat^2)

ATE.AIPW.sand.se <- sqrt(ATE.AIPW.sand.var)

ATE.reg.hat <- mean(outcome.treated.expectation) - mean(outcome.control.expectation)

weight.norm.1 <- 1/sum(treated.indic/pscore.probs)

weight.norm.3 <- 1/sum((1 - treated.indic)/(1 - pscore.probs))

ATE.IPW.hat <- weight.norm.1 * sum(term1) - weight.norm.3 * sum(term3)

return(list(ATE.AIPW.hat = ATE.AIPW.hat, ATE.reg.hat = ATE.reg.hat, ATE.IPW.hat = ATE.IPW.hat,
             ATE.AIPW.sand.SE = ATE.AIPW.sand.se))
```
# seed for bootstrap estimates reproducibility
seed <- 109876
set.seed(seed)

#-------------------------------------------------------------
# Model 1: Naive FE
# Model 2: Naive FE with X
# Model 3: FE weighted separately at each time point
# Model 4a: FE weighted separately at each time point + X
# Model 4b: FE weighted separately at each time point + X. bootstrap basic
# Model 4c: FE weighted separately at each time point + X. bootstrap perc
# Model 5: FE Stuart's 4-group weighted design
# Model 6a: FE Stuart's 4-group weighted design + X
# Model 6b: FE Stuart's 4-group weighted design + X. bootstrap basic
# Model 6c: FE Stuart's 4-group weighted design + X. bootstrap perc
# Model 7a: AIPW. sandwich
# Model 7b: AIPW. bootstrap basic
# Model 7c: AIPW. bootstrap perc
#-------------------------------------------------------------

meps.est <- function(outcome, n.boot) {
  #outcome="TOTTCH"
  #n.boot=50

  ptm <- proc.time()
  results <- matrix(NA, nrow=13, ncol=3)
  rownames(results) <- c("1. Naive", "2. Naive with X",
  "5. 4-grp wts", "6a. 4-grp wts+X", "6b. 4-grp wts+X. basic bs", "6c. 4-grp wts+X. perc bs",
  "7a. AIPW. sand", "7b. AIPW. bs basic", "7c. AIPW. bs perc")

  colnames(results) <- c("DID", "cl.l", "cl.u")
  # model formula: y ~ t + p + tp + X
fmla <- as.formula(paste("y ~ t + p + tp + povlev + age + male.f + black.f + hispanic.f + edu.f + ins.f + cc_3.f"))
# propensity score model: t ~ X
fmla.ps <- as.formula(paste("t ~ povlev + age + male.f + black.f + hispanic.f + edu.f + ins.f + cc_3.f"))
# outcomes:
# number of prescriptions
if(outcome=="RXTOT") meps$y <- meps$RXTOT
# cost of Medicaid prescriptions
# add small positive value to y for gamma model
if(outcome=="RXMCD") meps$y <- meps$RXMCD + 15
# total charges excluding prescriptions between $1-9,999 for those without private insurance
if(outcome=="TOTTCH") {
meps <- meps[which(meps$TOTTCH > 0 & meps$TOTTCH < 10000 & meps$ins.f != 1),]
meps$y <- meps$TOTTCH
}
# prepare data
predata <- meps[which(meps$p==0),]
postdata <- meps[which(meps$p==1),]
wts1 <- rep(1,6830)  # for marginal effect estimates of weighted estimators

# Model 1: Two-way fixed effects
if (outcome=="RXTOT") model1.glm <- glm2(y ~ t + p + tp, data=meps, family=poisson(link="log"))
if (outcome=="RXMCD") model1.glm <- glm2(y ~ t + p + tp, data=meps, family=Gamma(link="log"))
if (outcome=="TOTTCH") model1.glm <- glm2(y ~ t + p + tp, data=meps,
family=inverse.gaussian(link="log"))
marg.model1.glm <- marg(model1.glm, var_interest = 'tp', type = 'effects')
results[1,1] <- round(marg.model1.glm[[1]][2,2],2)
results[1,2] <- round(marg.model1.glm[[1]][[6]][2,2],2)
results[1,3] <- round(marg.model1.glm[[1]][[7]][2,2],2)

# Model 2: Two-way fixed effects + X
if (outcome=="RXTOT") model2.glm <- glm2(fmla, data=meps, family=poisson(link="log"))
APPENDIX B (continued)

```r
if (outcome=="RXMCD") model2.glm <- glm2(fmla, data=meps, family=Gamma(link="log"), maxit = 500)
if (outcome=="TOTTCH") model2.glm <- glm2(fmla, data=meps, family=inverse.gaussian(link="log"))
marg.model2.glm <- marg(model2.glm, var_interest = 'tp', type = 'effects')
results[2,1] <- round(marg.model2.glm[[1]][2,2],2)
results[2,2] <- round(marg.model2.glm[[1]][6][2],2)
results[2,3] <- round(marg.model2.glm[[1]][7][2],2)

# Models 3, 4: Seperate time weights
# propensity score model
prelogit <- glm(fmla.ps, data=predata, family="binomial")
postlogit <- glm(fmla.ps, data=postdata, family="binomial")
pre.pscores <- predict(prelogit, type="response")
post.pscores <- predict(postlogit, type="response")

# generate weights
meps$wts.sep[meps$group==1] <- 1/pre.pscores[predata$t==1]
meps$wts.sep[meps$group==2] <- 1/post.pscores[postdata$t==1]
meps$wts.sep[meps$group==3] <- 1/(1-pre.pscores[predata$t==0])
meps$wts.sep[meps$group==4] <- 1/(1-post.pscores[postdata$t==0])

# Model 3: Separate time weights
if (outcome=="RXTOT") model3.glm <- glm2(y ~ t + p + tp, weights=wts.sep, data=meps, family=poisson(link="log"))
if (outcome=="RXMCD") model3.glm <- glm2(y ~ t + p + tp, weights=wts.sep, data=meps, family=Gamma(link="log"))
if (outcome=="TOTTCH") model3.glm <- glm2(y ~ t + p + tp, weights=wts.sep, data=meps, family=inverse.gaussian(link="log"))
marg.model3.glm <- suppressWarnings(marg(model3.glm, var_interest = 'tp', type = 'effects', weights = wts1))
results[3,1] <- round(marg.model3.glm[[1]][2,2],2)
results[3,2] <- round(marg.model3.glm[[1]][6][2],2)
results[3,3] <- round(marg.model3.glm[[1]][7][2],2)

# Model 4a: Separate time weights + X
if (outcome=="RXTOT") model4.glm <- glm2(fmla, weights=wts.sep, data=meps, family=poisson(link="log"))
```

if (outcome=="RXMCD") model4.glm <- glm2(fmla, weights=wts.sep, data=meps, family=Gamma(link="log"))
if (outcome=="TOTTCH") model4.glm <- glm2(fmla, weights=wts.sep, data=meps,
family=inverse.gaussian(link="log"))
marg.model4.glm <- suppressWarnings(marg(model4.glm, var_interest = 'tp', type = 'effects', weights = wts1))
# an alternate, equivalent method of getting the marginal effect is:
#b.tp <- coef(model4.glm)['tp']
#xb <- predict(model4.glm, type="link")
#est.alt <- mean(ifelse(meps$group==2, exp(xb) - exp(xb-b.tp), exp(xb+b.tp) - exp(xb)))
#est.alt
results[4,1] <- round(marg.model4.glm[[1]][2,2],2)
results[4,2] <- round(marg.model4.glm[[1]][[6]][2],2)
results[4,3] <- round(marg.model4.glm[[1]][[7]][2],2)

# Model 4b, 4c: Separate time weights + X, bootstrap
sep.time.wtd <- function(dat, ind) {
  d <- dat[ind,]  # allows boot to select sample
  if (outcome=="RXTOT") model56.glm <- glm2(fmla, weights=wts.sep, data=d,
family=poisson(link="log"))
  if (outcome=="RXMCD") model56.glm <- glm2(fmla, weights=wts.sep, data=d,
family=Gamma(link="log"))
  if (outcome=="TOTTCH") model56.glm <- glm2(fmla, weights=wts.sep, data=d,
family=inverse.gaussian(link="log"))
  suppressWarnings(marg(model56.glm, var_interest = 'tp', type = 'effects', weights = wts1))[[1]][2,2]
  boot.sep.time.wtd <- boot(meps, sep.time.wtd, R=n.boot)
est <- median(boot.sep.time.wtd$t)
q.025 <- quantile(boot.sep.time.wtd$t, c(0.025, 0.975))[1]
q.975 <- quantile(boot.sep.time.wtd$t, c(0.025, 0.975))[2]
results[5,1] <- results[6,1] <- round(est,2)
# basic bootstrap
results[5,2] <- round(2*est - q.975,2)
results[5,3] <- round(2*est - q.025,2)
# percentile bootstrap

APPENDIX B (continued)
APPENDIX B (continued)

```r
results[6,2] <- round(q.025,2)
results[6,3] <- round(q.975,2)

# Model 5: 4-group weighted
# propensity score model
mlogit <- multinom(group ~ povlev + age + male.f + black.f + hispanic.f + edu.f + ins.f + cc_3.f, data=meps)
probs <- predict(mlogit, type="probs")

# generate weights
meps$wts.4grp[meps$group==1] <- probs[meps$group==1,1]/probs[meps$group==2,1]
meps$wts.4grp[meps$group==2] <- probs[meps$group==2,1]/probs[meps$group==2,2]
meps$wts.4grp[meps$group==3] <- probs[meps$group==3,1]/probs[meps$group==3,3]
meps$wts.4grp[meps$group==4] <- probs[meps$group==4,1]/probs[meps$group==4,4]

if (outcome=="RXTOT") model7.glm <- glm(y ~ t + p + tp, weights=wts.4grp, data=meps, family=poisson(link="log"))
if (outcome=="RXMCD") model7.glm <- glm2(y ~ t + p + tp, weights=wts.4grp, data=meps, family=Gamma(link="log"))
if (outcome=="TOTTCH") model7.glm <- glm2(y ~ t + p + tp, weights=wts.4grp, data=meps, family=inverse.gaussian(link="log"))
  marg.model7.glm <- suppressWarnings(marg(model7.glm, var_interest = 'tp', type = 'effects'))
  results[7,1] <- round(marg.model7.glm[[1]][[2,2]],2)
  results[7,2] <- round(marg.model7.glm[[1]][[6]][[2,2]],2)
  results[7,3] <- round(marg.model7.glm[[1]][[7]][[2,2]],2)

# Model 6a: 4-group weighted + X
if (outcome=="RXTOT") model8.glm <- glm2(fmla, weights=wts.4grp, data=meps, family=poisson(link="log"))
if (outcome=="RXMCD") model8.glm <- glm2(fmla, weights=wts.4grp, data=meps, family=Gamma(link="log"))
if (outcome=="TOTTCH") model8.glm <- glm2(fmla, weights=wts.4grp, data=meps, family=inverse.gaussian(link="log"))
  marg.model8.glm <- suppressWarnings(marg(model8.glm, var_interest = 'tp', type = 'effects', weights = wts1))
  results[8,1] <- round(marg.model8.glm[[1]][[2,2]],2)
  results[8,2] <- round(marg.model8.glm[[1]][[6]][[2,2]],2)
```

APPENDIX B (continued)

```r
results[8,3] <- round(marg.model8.glm[[1]][[7]][2,2], 2)

# Model 6b, c: 4-group weighted + X, bootstrap
four.grp.wtd <- function(dat, ind) {
  d <- dat[ind,] # allows boot to select sample
  if (outcome == "RXTOT")
    model910.glm <- glm2(fmla, weights=wts.4grp, data=d,
                          family=poisson(link="log"))
  if (outcome == "RXMCD")
    model910.glm <- glm2(fmla, weights=wts.4grp, data=d,
                          family=Gamma(link="log"))
  if (outcome == "TOTTCH")
    model910.glm <- glm2(fmla, weights=wts.4grp, data=d,
                          family=inverse.gaussian(link="log"))
  suppressWarnings(marg(model910.glm, var_interest = 'tp', type = 'effects', weights = wts1))[[1]][2,2]
  boot.four.grp.wtd <- boot(meps, four.grp.wtd, R=n.boot)
  est <- median(boot.four.grp.wtd$t)
  q.025 <- quantile(boot.four.grp.wtd$t, c(0.025, 0.975))[1]
  q.975 <- quantile(boot.four.grp.wtd$t, c(0.025, 0.975))[2]

  results[9,1] <- results[10,1] <- round(est, 2)
  # basic bootstrap
  results[9,2] <- round(2*est - q.975, 2)
  results[9,3] <- round(2*est - q.025, 2)
  # percentile bootstrap
  results[10,2] <- round(q.025, 2)
  results[10,3] <- round(q.975, 2)

  # Model 7a: AIPW, sandwich standard error
  aipw.pre <- aipw(predata, outcome=outcome)
  aipw.post <- aipw(postdata, outcome=outcome)
  est <- aipw.post$ATE.AIPW.hat - aipw.pre$ATE.AIPW.hat
  # sandwich standard error
  sand.SE <- sqrt((aipw.pre$ATE.AIPW.sand.SE)^2 + (aipw.post$ATE.AIPW.sand.SE)^2)

  results[11,1] <- round(est, 2)
  results[11,2] <- round(est - one.96*sand.SE, 2)
```
APPENDIX B (continued)

```r
results[11,3] <- round(est + one.96*sand.SE,2)

# Model 7b, c: AIPW, bootstrap
aipw.diff <- rep(NA, n.boot)
for (biter in 1:n.boot) {
  boot.inds.pre <- sample(1:dim(predata)[1], 3552, replace = TRUE)
  d.pre <- predata[boot.inds.pre, ]
  boot.inds.post <- sample(1:dim(postdata)[1], 3278, replace = TRUE)
  d.post <- postdata[boot.inds.post, ]
  # aipw_did = aipw_post - aipw_pre
  aipw.diff[biter] <- aipw(data = d.post, outcome=outcome)$ATE.AIPW.hat -
    aipw(data = d.pre, outcome=outcome)$ATE.AIPW.hat
}
est <- median(aipw.diff)
# take 2.5 and 97.5th percentile of bootstrap estimates
q.025 <- quantile(aipw.diff, c(0.025, 0.975))[[1]]
q.975 <- quantile(aipw.diff, c(0.025, 0.975))[[2]]
results[12,1] <- results[13,1] <- round(est,2)
# basic bootstrap
results[12,2] <- round(2*est - q.975,2)
results[12,3] <- round(2*est - q.025,2)
# percentile bootstrap
results[13,2] <- round(q.025,2)
results[13,3] <- round(q.975,2)

# end timer
run.time.all <- proc.time() - ptm
run.time <- round(run.time.all[3] / 60,2)

# output system time and simulation settings
params <- data.frame(outcome, n.boot, seed, run.time)
return(list(results, params, meps))
```

# Output simulation results to Excel
# number of prescriptions
meps.results <- meps.est(outcome="RXTOT", n.boot=500)
file.name <- paste("output/MEPS_trans_TEST_", format(Sys.time(), "%Y-%m-%d"), "_num_rx.csv", sep = ")
write.table(meps.results[[1]], file = file.name,
    eol = "\r",sep = ",",row.names = TRUE)
write.table(meps.results[[2]], file = file.name,
    eol = "\r",sep = ",",row.names = FALSE, append=TRUE)

# cost of Medicaid prescriptions
meps.results <- meps.est(outcome="RXMCD", n.boot=500)
file.name <- paste("output/MEPS_trans_TEST_", format(Sys.time(), "%Y-%m-%d"), "_cost.csv", sep = ")
write.table(meps.results[[1]], file = file.name,
    eol = "\r",sep = ",",row.names = TRUE)
write.table(meps.results[[2]], file = file.name,
    eol = "\r",sep = ",",row.names = FALSE, append=TRUE)

# total charges excluding prescriptions between 0-10,000 for those without private insurance
meps.results <- meps.est(outcome="TOTTCH", n.boot=500)
file.name <- paste("output/MEPS_trans_TEST_", format(Sys.time(), "%Y-%m-%d"), "_totch.csv", sep = ")
write.table(meps.results[[1]], file = file.name,
    eol = "\r",sep = ",",row.names = TRUE)
write.table(meps.results[[2]], file = file.name,
    eol = "\r",sep = ",",row.names = FALSE, append=TRUE)
Simulation using artificial data

# Simulation using artificial data
#
# Model 1: Naive FE
# Model 2: Naive FE with X
# Model 3: FE weighted separately at each time point
# Model 4a: FE weighted separately at each time point + X
# Model 4b: FE weighted separately at each time point + X, bootstrap basic
# Model 4c: FE weighted separately at each time point + X, bootstrap perc
# Model 5: FE Stuart's 4-group weighted design
# Model 6a: FE Stuart's 4-group weighted design + X
# Model 6b: FE Stuart's 4-group weighted design + X, bootstrap basic
# Model 6c: FE Stuart's 4-group weighted design + X, bootstrap perc
# Model 7a: AIPW, sandwich
# Model 7b: AIPW, bootstrap basic
# Model 7c: AIPW, bootstrap perc

library(modmarg)
library(boot)
library(nnet)
setwd("Z:/simulation/artificial")
rm(list = ls())
set.seed(3543)
nsims <- 2  # number of simulations per setting
n.boot <- 50 # number of bootstrap replications
nset <- 4   # number of settings (at most 4)

# 4 settings in covariate confounding
delta.baseline <- c(0,0.3,0.3,0.3)
delta.t <- c(0,0.2,-0.2)
delta.c <- c(0,0.2,0.1)
bias.final <- rmse.final <- cov.final <- matrix(NA, nrow=nset, ncol=13)
colnames(bias.final) <- colnames(rmse.final) <- colnames(cov.final) <-
c("1. Naive", "2. Naive with X",
"5. 4-grp wts", "6a. 4-grp wts+X", "6b. 4-grp wts+X. basic bs", "6c. 4-grp wts+X. perc bs",
"7a. AIPW. sand", "7b. AIPW. bs", "7c. AIPW. bs perc")
a <- .3
b <- .1
c <- .2
d <- .1
e <- .1
f <- .2
g <- .15
nc.pre <- 500
nt.pre <- 500
nc.post <- 500
nt.post <- 500
n.all <- nc.pre + nt.pre + nc.post + nt.post
aipw <- function(data) {
  treatment.var = "t"
  treatment.vec <- data[, treatment.var]
  treatment.values <- sort(unique(treatment.vec))

  treated.data <- data[treatment.vec == treatment.values[2], ]
  control.data <- data[treatment.vec == treatment.values[1], ]
  n.treated.pre <- nrow(treated.data)
  n.control.pre <- nrow(control.data)

  # propensity score model
  ps.model <- suppressWarnings(glm(t ~ x, family = binomial, data = data))
  pscore.probs <- suppressWarnings(predict(ps.model, newdata = data, type = "response"))
  pscores.pre <- pscore.probs
# outcome formula for treated
glm.t <- suppressWarnings(glm(y ~ x, family = gaussian, data = treated.data))
# outcome formula for control
glm.c <- suppressWarnings(glm(y ~ x, family = gaussian, data = control.data))
outcome.treated.expectation <- suppressWarnings(predict(glm.t, newdata = data, type = "response"))
outcome.control.expectation <- suppressWarnings(predict(glm.c, newdata = data, type = "response"))
res.c <- suppressWarnings(residuals(glm.c, type = "response"))
res.t <- suppressWarnings(residuals(glm.t, type = "response"))

n <- length(treatment.vec)
term1 <- rep(0, n)
term2 <- rep(0, n)
term3 <- rep(0, n)
term4 <- rep(0, n)
control.indic <- treatment.vec == treatment.values[1]
treated.indic <- treatment.vec == treatment.values[2]
outcome.vec <- model.response(model.frame(y ~ x, data = data))

term1 <- outcome.vec/pscore.probs
term1[control.indic] <- 0
term2 <- ((1 - pscore.probs) * outcome.treated.expectation)/pscore.probs
term2[control.indic] <- ((0 - pscore.probs[control.indic]) * outcome.treated.expectation[control.indic])/pscore.probs[control.indic]
term3 <- outcome.vec/(1 - pscore.probs)
term3[treated.indic] <- 0
term4 <- ((1 - pscore.probs) * outcome.control.expectation)/(1 - pscore.probs)
term4[control.indic] <- ((0 - pscore.probs[control.indic]) * outcome.control.expectation[control.indic])/(1 - pscore.probs[control.indic])

ATE.AIPW.hat <- (1/n) * (sum(term1) - sum(term2) - sum(term3) - sum(term4))

I.hat <- term1 - term2 - term3 - term4 - ATE.AIPW.hat
ATE.AIPW.sand.var <- (1/n^2) * sum(I.hat^2)
ATE.AIPW.sand.se <- sqrt(ATE.AIPW.sand.var)
ATE.reg.hat <- mean(outcome.treated.expectation) - mean(outcome.control.expectation)
APPENDIX C (continued)

weight.norm.1 <- 1/sum(treated.indic/pscore.probs)
weight.norm.3 <- 1/sum((1 - treated.indic)/(1 - pscore.probs))
ATE.IPW.hat <- weight.norm.1 * sum(term1) - weight.norm.3 * sum(term3)

return(list(ATE.AIPW.hat = ATE.AIPW.hat, ATE.reg.hat = ATE.reg.hat,
ATE.IPW.hat = ATE.IPW.hat, ATE.AIPW.sand.SE = ATE.AIPW.sand.se))

# Settings (1,2,3,4)
for (s in 1:nset) {
    #s <- 1

    bias <- mse <- coverage <- matrix(NA, nrow=nsims, ncol=13)
    colnames(bias) <- colnames(mse) <- colnames(coverage) <-
        c("1. Naive", "2. Naive with X",
           "5. 4-grp wts", "6a. 4-grp wts+X", "6b. 4-grp wts+X. basic bs", "6c. 4-grp wts+X. perc bs",
           "7a. AIPW. sand", "7b. AIPW. bs", "7c. AIPW. bs perc")

    for (i in 1:nsims) {
        #i <- 1

        # print iteration
        print(paste(s, ".", i, sep=""))

        # Generate covariate X
        xc.pre <- rnorm(nc.pre, 0, 1)
        xt.pre <- rnorm(nt.pre, 0+delta.baseline[s], 1)
        xc.post <- rnorm(nc.pre, 0+delta.c[s], 1)
        xt.post <- rnorm(nt.pre, 0+delta.baseline[s]+delta.t[s], 1)

        # ORDER: T pre, T post, C pre, C post
        x <- c(xt.pre, xt.post, xc.pre, xc.post)
        p <- c(rep(0, nt.pre), rep(1, nt.post), rep(0, nc.pre), rep(1, nc.post))
        t <- c(rep(1, nt.pre), rep(1, nt.post), rep(0, nc.pre), rep(0, nc.post))
# model, $Y = 0 + aT + bP + cX + dT\times P + eT\times X + fP\times X + gT\times P\times X$

\[ y \leftarrow 0 + a\times t + b\times p + c\times x + d\times t\times p + e\times t\times x + f\times p\times x + g\times t\times p\times x + \text{rnorm(length(t), 0, 1)} \]

# Effect in the treatment group:
\[ \text{truth} \leftarrow \text{mean}(d + g\times xt.post) \]

# prepare data
group <- c(rep(1, nt.pre), rep(2, nt.post), rep(3, nc.pre), rep(4, nc.post))
tp <- t * p
sim <- data.frame(y, group, t, p, tp, x)
predata <- sim[which(sim$p==0),] # predata
postdata <- sim[which(sim$p==1),] # postdata
one.96 <- -qnorm(0.025)

# Model 1: Two-way fixed effects
model1 <- glm(y ~ t + p + t*p, data=sim)
est <- summary(model1)$coef["t:p",1]
se <- summary(model1)$coef["t:p", 2]
bias[1,1] <- est-truth
mse[1,1] <- (est-truth)^2
coverage[1,1] <- ifelse((truth > (est-one.96*se)) & (truth < (est+one.96*se)), 1, 0)

# Model 2: Two-way fixed effects + X
model2 <- glm(y ~ t + p + t*p + x, data=sim)
est <- summary(model2)$coef["t:p",1]
se <- summary(model2)$coef["t:p", 2]
bias[1,2] <- est-truth
mse[1,2] <- (est-truth)^2
coverage[1,2] <- ifelse((truth > (est-one.96*se)) & (truth < (est+one.96*se)), 1, 0)

# Models 3, 4: Separate time weights
# propensity score model
prelogit <- glm(t ~ x, data=predata, family="binomial")
postlogit <- glm(t ~ x, data=postdata, family="binomial")
pre.pscores <- predict(prelogit, type="response")
post.pscores <- predict(postlogit, type="response")
# generate weights
sim$wts.sep[sim$group==1] <- 1/pre.pscores[predata$t==1]
sim$wts.sep[sim$group==2] <- 1/post.pscores[postdata$t==1]
sim$wts.sep[sim$group==3] <- 1/(1-pre.pscores[predata$t==0])
sim$wts.sep[sim$group==4] <- 1/(1-post.pscores[postdata$t==0])

# Model 3: Separate time weights
model3 <- glm(y ~ t + p + t*p, weights=wts.sep, data=sim)
est <- summary(model3)$coef["t:p",1]
se <- summary(model3)$coef["t:p", 2]
bias[i,3] <- est-truth
mse[i,3] <- (est-truth)^2
coverage[i,3] <- ifelse((truth > (est-1.96*se)) & (truth < (est+1.96*se)), 1, 0)

# Model 4a: Separate time weights + X
model4 <- glm(y ~ t + p + t*p + x, weights=wts.sep, data=sim)
est <- summary(model4)$coef["t:p",1]
se <- summary(model4)$coef["t:p", 2]
bias[i,4] <- est-truth
mse[i,4] <- (est-truth)^2
coverage[i,4] <- ifelse((truth > (est-1.96*se)) & (truth < (est+1.96*se)), 1, 0)

# Model 4b, 4c: Separate time weights + X, bootstrap
sep.time.wtd <- function(dat, ind) {
d <- dat[ind,]  # allows boot to select sample
    model56 <- glm(y ~ t + p + t*p + x, weights=wts.sep, data=d)
    summary(model56)$coef["t:p",1]
}
boot.sep.time.wtd <- boot(sim, sep.time.wtd, R=n.boot)
est <- median(boot.sep.time.wtd$t)
boot.ci.basic <- boot.ci(boot.out=boot.sep.time.wtd, conf=0.95, type="basic")
boot.ci.perc <- boot.ci(boot.out=boot.sep.time.wtd, conf=0.95, type="perc")
# Save bias and coverage information on the B bootstrap samples
bias[i,5] <- bias[i,6] <- est - truth
mse[i,5] <- mse[i,6] <- (est - truth)^2
coverage[i,5] <- ifelse((truth > boot.ci.basic$basic[4]) & (truth < boot.ci.basic$basic[5]), 1, 0)
coverage[i,6] <- ifelse((truth > boot.ci.perc$perc[4]) & (truth < boot.ci.perc$perc[5]), 1, 0)

# Model 5: 4-group weighted
# propensity score model
mlogit <- multinom(group ~ x, data=sim)
probs <- predict(mlogit, type="probs")

# generate weights
sim$wts.4grp[sim$group==1] <- probs[sim$group==1,1]/probs[sim$group==1,1]
sim$wts.4grp[sim$group==2] <- probs[sim$group==2,1]/probs[sim$group==2,2]
sim$wts.4grp[sim$group==3] <- probs[sim$group==3,1]/probs[sim$group==3,3]
sim$wts.4grp[sim$group==4] <- probs[sim$group==4,1]/probs[sim$group==4,4]
model5 <- glm(y ~ t + p + t*p, weights=wts.4grp, data=sim)
est.4grpwt <- summary(model5)$coef["t:p",1]
se <- summary(model5)$coef["t:p", 2]
bias[i,7] <- est.4grpwt-truth
mse[i,7] <- (est.4grpwt-truth)^2
coverage[i,7] <- ifelse(((truth > (est.4grpwt-1.96*se)) & (truth < (est.4grpwt+1.96*se))), 1, 0)

# Model 6a: 4-group weighted + X
model8 <- glm(y ~ t + p + t*p + x, weights=wts.4grp, data=sim)
est <- summary(model8)$coef["t:p",1]
se <- summary(model8)$coef["t:p", 2]
bias[i,8] <- est-truth
mse[i,8] <- (est-truth)^2
coverage[i,8] <- ifelse(((truth > (est-1.96*se)) & (truth < (est+1.96*se))), 1, 0)

# Model 6b, c: 4-group weighted + X, bootstrap
four.grp.wtd <- function(dat, ind) {
d <- dat[ind,] # allows boot to select sample
model910 <- glm(y ~ t + p + t*p + x, weights=wts.4grp, data=d)
summary(model910)$coef["t:p",1]
}

boot.four.grp.wtd <- boot(sim, four.grp.wtd, R=n.boot)
est <- median(boot.four.grp.wtd$t)

boot.ci.basic <- boot.ci(boot.out=boot.four.grp.wtd, conf=0.95, type="basic")
boot.ci.perc <- boot.ci(boot.out=boot.four.grp.wtd, conf=0.95, type="perc")

# Save bias and coverage information on the B bootstrap samples
bias[i,9] <- bias[i,10] <- est - truth
mse[i,9] <- mse[i,10] <- (est - truth)^2

coverage[i,9] <- ifelse((truth > boot.ci.basic$basic[4]) & (truth < boot.ci.basic$basic[5]), 1, 0)
coverage[i,10] <- ifelse((truth > boot.ci.perc$perc[4]) & (truth < boot.ci.perc$perc[5]), 1, 0)

# Model 7a: AIPW, sandwich standard error
aipw.pre <- aipw(predata)
aipw.post <- aipw(postdata)
est <- aipw.post$ATE.AIPW.hat - aipw.pre$ATE.AIPW.hat

# sandwich standard error
sand.SE <- sqrt((aipw.pre$ATE.AIPW.sand.SE)^2 + (aipw.post$ATE.AIPW.sand.SE)^2)

bias[i,11] <- est-truth
mse[i,11] <- (est-truth)^2
coverage[i,11] <- ifelse((truth > est - one.96*sand.SE) & (truth < est + one.96*sand.SE), 1, 0)

# Model 7b, c: AIPW, bootstrap
aipw.diff <- rep(NA, n.boot)
for (biter in 1:n.boot) {
  boot.inds.pre <- sample(1:dim(predata)[1], dim(predata)[1], replace = TRUE)
d.pre <- predata[boot.inds.pre,]
boot.inds.post <- sample(1:dim(postdata)[1], dim(postdata)[1], replace = TRUE)
d.post <- postdata[boot.inds.post,]
  # aipw_did= aipw_post - aipw_pre
  aipw.diff[biter] <- aipw(data = d.post)$ATE.AIPW.hat - aipw(data = d.pre)$ATE.AIPW.hat}
est <- median(aipw.diff)
# take 2.5 and 97.5th percentile of bootstrap estimates
q.025 <- quantile(aipw.diff, c(0.025, 0.975))[1]
q.975 <- quantile(aipw.diff, c(0.025, 0.975))[2]

bias[i,12] <- bias[i,13] <- est - truth
mse[i,12] <- mse[i,13] <- (est - truth)^2
coverage[i,12] <- ifelse((truth > 2*est - q.975) & (truth < 2*est - q.025), 1, 0)
coverage[i,13] <- ifelse((truth > q.025) & (truth < q.975), 1, 0)

bias.final[s,] <- t(round(apply(bias, 2, mean), 4))
rmse.final[s,] <- t(round(sqrt(apply(mse, 2, mean)), 4))
cov.final[s,] <- t(round(apply(coverage, 2, mean), 3))

bias.final
rmse.final
cov.final

# output system time and simulation settings
sim.settings <- data.frame(nsims, n.boot)
sim.settings

# Output simulation results
file.name <- paste("artificial_sim_TEST_", format(Sys.time(), "%Y-%m-%d"), ".csv", sep = "")
write.table(bias.final, file = file.name,
eol = "\r", sep = "", row.names = FALSE)
write.table(rmse.final, file = file.name,
eol = "\r", sep = "", row.names = FALSE, append=TRUE)
write.table(cov.final, file = file.name,
eol = "\r", sep = "", row.names = FALSE, append=TRUE)
write.table(sim.settings, file = file.name,
eol = "\r", sep = "", col.names = TRUE, row.names = FALSE, append=TRUE)
APPENDIX D

Dataset generation

# program name: meps.sim.dataset.R
#
# purpose: generate a MEPS dataset

meps.sim.dataset <- function(s.seed, meps.n.1=916, meps.n.2=871, meps.n.3=2636, meps.n.4=2407, meps.dat=meps, subset="N") {

  # 6830 observations in MEPS data
  # 2592 observations in subseted MEPS data
  dim(meps.dat)

  # subset the data by groups
  meps.1 <- meps.dat[which(meps.dat$group == 1),]
  meps.2 <- meps.dat[which(meps.dat$group == 2),]
  meps.3 <- meps.dat[which(meps.dat$group == 3),]
  meps.4 <- meps.dat[which(meps.dat$group == 4),]

  # Simulate the 8 covariates using PoisBinOrdNonNor
  meps.sim <- function(dat, dat.n, n.total, seed) {
    set.seed(seed)

    # MEPS moments of continuous variables
    meps.fpl.mean <- mean(dat$povlev)
    meps.fpl.var <- var(dat$povlev)
    meps.fpl.skew <- skewness(dat$povlev)
    meps.fpl.kurt <- kurtosis(dat$povlev)
    fpl.moments <- c(meps.fpl.mean, meps.fpl.var, meps.fpl.skew, meps.fpl.kurt)

    meps.age.mean <- mean(dat$age)
    meps.age.var <- var(dat$age)
    meps.age.skew <- skewness(dat$age)

    ...
APPENDIX D (continued)

meps.age.kurt <- kurtosis(dat$age)
age.moments <- c(meps.age.mean, meps.age.var, meps.age.skew, meps.age.kurt)

# MEPS probabilites of binary variables
male.marg <- table(dat$male)/dat.n
black.marg <- table(dat$black)/dat.n
hispanic.marg <- table(dat$hispanic)/dat.n

# MEPS probabilites of ordinal variables
edu.marg <- table(dat$edu)/dat.n
ins.marg <- table(dat$ins)/dat.n
cc_3.marg <- table(dat$cc_3)/dat.n

d <- data.frame(dat$male, dat$black, dat$hispanic, dat$edu, dat$ins, dat$cc_3, dat$povlev, dat$age)
cor.d <- cor(d)

cmat.star <- find.cor.mat.star(cor.mat = cor.d,
    no.pois = 0, no.bin = 3, no.ord = 3, no.nonn = 2,
    bin.list = list(male.marg, black.marg, hispanic.marg),
    ord.list = list(edu.marg, ins.marg, cc_3.marg),
    nonn.list = list(fpl.moments, age.moments))

sim.data <- genPBONN(n.total, cmat.star = cmat.star,
    no.pois = 0, no.bin = 3, no.ord = 3, no.nonn = 2,
    bin.list = list(male.marg, black.marg, hispanic.marg),
    ord.list = list(edu.marg, ins.marg, cc_3.marg),
    nonn.list = list(fpl.moments, age.moments))

sim.data <- as.data.frame(sim.data)
colnames(sim.data) <- c("male","black","hispanic","edu","ins","cc_3","povlev","age")

# if age is <18 make=18
sim.data$age[which(sim.data$age < 18)] <- 18
# if age is >=65 make=64
sim.data$age[which(sim.data$age >= 65)] <- 64.99
# round age to nearest, lowest integer
sim.data$age <- floor(sim.data$age)
APPENDIX D (continued)

```r
# if povlev is <0 make=0
sim.data$povlev[which(sim.data$povlev < 0)] <- 0
# if povlev is >138 make=138
sim.data$povlev[which(sim.data$povlev >= 138)] <- 137.99
return(sim.data)
```

```r
meps.sim.1 <- meps.sim(dat=meps.1, dat.n=dim(meps.1)[1], n.total=meps.n.1, seed=54321 + s.seed)
meps.sim.2 <- meps.sim(dat=meps.2, dat.n=dim(meps.2)[1], n.total=meps.n.2, seed=4321 + s.seed)
meps.sim.3 <- meps.sim(dat=meps.3, dat.n=dim(meps.3)[1], n.total=meps.n.3, seed=321 + s.seed)
meps.sim.4 <- meps.sim(dat=meps.4, dat.n=dim(meps.4)[1], n.total=meps.n.4, seed=21 + s.seed)

# Append all 4 simulated groups' variables together
meps.sim.1$group <- 1; meps.sim.1$t <- 1; meps.sim.1$p <- 0
meps.sim.2$group <- 2; meps.sim.2$t <- 1; meps.sim.2$p <- 1
meps.sim.3$group <- 3; meps.sim.3$t <- 0; meps.sim.3$p <- 0
meps.sim.4$group <- 4; meps.sim.4$t <- 0; meps.sim.4$p <- 1

final.sim <- rbind(meps.sim.1, meps.sim.2, meps.sim.3, meps.sim.4)
dim(final.sim)
head(final.sim)

# Reformat variables
final.sim$male <- final.sim$male + 1
final.sim$edu <- final.sim$edu + 1

# Covariates: povlev age male black hispanic edu ins cc_3
final.sim$male.f <- factor(final.sim$male)
# force male to be reference category
final.sim$male.f <- relevel(final.sim$male.f, ref = 2)
final.sim$black.f <- factor(final.sim$black)
final.sim$hispanic.f <- factor(final.sim$hispanic)
final.sim$edu.f <- factor(final.sim$edu)
final.sim$cc_3.f <- factor(final.sim$cc_3)
# Create interaction
final.sim$tp <- final.sim$t * final.sim$p
```
# For full data: generate outcomes RXTOT and RXMCD
if (subset=="N") {
    # format insurance variable
    final.sim$ins <- final.sim$ins + 1
    final.sim$ins.f <- factor(final.sim$ins)

    #############
    ## Generate RXTOT ##
    # Estimate a poisson model of RXTOT
    rxtot.reg <- glm2(RXTOT ~ t + p + tp + povlev + age + male.f + black.f + hispanic.f +
                      edu.f + ins.f + cc_3.f, family=poisson(link="log"), data=meps)
    # Predict RXTOT from the model using actual dataset, round estimates down to nearest integer
    final.sim$RXTOT <- floor(predict(rxtot.reg, final.sim, type="response"))

    #############
    ## Generate RXMCD ##
    # Predict RXMCD from the model earlier with the simulated set of covariates
    rxmcd.reg <- suppressWarnings(glm2(RXMCD + 15 ~ t + p + tp + povlev + age + male.f + black.f +
                                       hispanic.f +
                                       edu.f + ins.f + cc_3.f, family=Gamma(link="log"), data=meps))
    # Predict RXTOT from the model using actual dataset, round estimates down to nearest integer
    final.sim$RXMCD <- floor(predict(rxmcd.reg, final.sim, type="response"))
} else if (subset=="Y") {
    # for subgroup: generate outcome TOTTCH

    # format insurance variable, this is different for subgrouped data since now there are only 2 levels
    instead of 3
    final.sim$ins <- final.sim$ins + 2
    final.sim$ins.f <- factor(final.sim$ins)

    ## Generate TOTTCH ##
    # Predict TOTTCH from the model earlier with the simulated set of covariates
APPENDIX D (continued)

tottch.reg <- suppressWarnings(glm2(TOTTCH ~ t + p + tp + povlev + age + male.f + black.f +
    hispanic.f +
    edu.f + ins.f + cc_3.f, family=inverse.gaussian(link="log"), data=meps.subset))
# Predict TOTTCH from the model using actual dataset, round estimates down to nearest integer
final.sim$TOTTCH <- floor(predict(tottch.reg, final.sim, type="response"))
# truncate if outside (0,10000)
# if TOTTCH is <=0 make=1
final.sim$TOTTCH[which(final.sim$TOTTCH <= 0)] <- 1
# if TOTTCH is >10000 make=9999
final.sim$TOTTCH[which(final.sim$TOTTCH >= 10000)] <- 9999

# clean up simulated dataset
final.sim <- final.sim[,c(-1:-6)]

return(final.sim)
APPENDIX E

Truth determination

# program name: get.truth.marg.R
# purpose: to get the truth marginal effect from the big dataset
# confounding in 2 variables (age, povlev), and 3 settings

get.truth.marg <- function(subset="N") {
if (subset=="N") {

# create interactions
big$age.pt <- big$age * big$p * big$t
big$age.s2.sm.pt <- big$age.s2.sm * big$p * big$t
big$age.s2.med.pt <- big$age.s2.med * big$p * big$t
big$age.s2.lar.pt <- big$age.s2.lar * big$p * big$t
big$age.s3.sm.pt <- big$age.s3.sm * big$p * big$t
big$age.s3.med.pt <- big$age.s3.med * big$p * big$t
big$age.s3.lar.pt <- big$age.s3.lar * big$p * big$t
big$age.s4.sm.pt <- big$age.s4.sm * big$p * big$t
big$age.s4.med.pt <- big$age.s4.med * big$p * big$t
big$age.s4.lar.pt <- big$age.s4.lar * big$p * big$t

big$povlev.pt <- big$povlev * big$p * big$t
big$povlev.s2.sm.pt <- big$povlev.s2.sm * big$p * big$t
big$povlev.s2.med.pt <- big$povlev.s2.med * big$p * big$t
big$povlev.s2.lar.pt <- big$povlev.s2.lar * big$p * big$t
big$povlev.s3.sm.pt <- big$povlev.s3.sm * big$p * big$t
big$povlev.s3.med.pt <- big$povlev.s3.med * big$p * big$t
big$povlev.s3.lar.pt <- big$povlev.s3.lar * big$p * big$t
big$povlev.s4.sm.pt <- big$povlev.s4.sm * big$p * big$t
big$povlev.s4.med.pt <- big$povlev.s4.med * big$p * big$t
big$povlev.s4.lar.pt <- big$povlev.s4.lar * big$p * big$t

# get truths
for(v in 1:length(vlist)) {

...
outcome <- "RXTOT"
int.name <- paste(vlist[v],".pt",sep="")

# confounded age
if(v==1 | v==3 | v==5 | v==7 | v==9 | v==11 | v==13 | v==15 | v==17 | v==19) {
fmla <- as.formula(paste("RXTOT ~ t + p + tp + povlev +",
    vlist[v], "+", vlist[v], "+", vlist[v], "+", vlist[v], "+", int.name, "+ male.f + black.f + hispanic.f + edu.f + ins.f + cc_3.f"))
}

# confounded povlev
if(v==2 | v==4 | v==6 | v==8 | v==10 | v==12 | v==14 | v==16 | v==18 | v==20) {
fmla <- as.formula(paste("RXTOT ~ t + p + tp + age +",
    vlist[v], "+", vlist[v], "+", vlist[v], "+", int.name, "+ male.f + black.f + hispanic.f + edu.f + ins.f + cc_3.f"))
}
model <- glm2(fmla, data=big, family=poisson(link="log"))
b.tp <- coef(model)['tp']
b.tpx <- coef(model)[int.name]
xb <- predict(model, type="link")
truth <- mean(ifelse(big$group==2, (exp(xb) - exp(xb-b.tp-big[vlist[v]]*b.tpx))[1],
    (exp(xb+b.tp+big[vlist[v]]*b.tpx) - exp(xb))[1]))
assign(paste("truth." , outcome,"." , vlist[v], sep=""), truth)
}

for(v in 1:length(vlist)) {
outcome <- "RXMCD"
int.name <- paste(vlist[v],".pt",sep="")

# confounded age
if(v==1 | v==3 | v==5 | v==7 | v==9 | v==11 | v==13 | v==15 | v==17 | v==19) {
fmla <- as.formula(paste("RXMCD + 15 ~ t + p + tp + povlev +",
    vlist[v], "+", vlist[v], "+", vlist[v], "+", vlist[v], "+", int.name, "+ male.f + black.f + hispanic.f + edu.f + ins.f + cc_3.f"))
}

# confounded povlev
if(v==2 | v==4 | v==6 | v==8 | v==10 | v==12 | v==14 | v==16 | v==18 | v==20) {
fmla <- as.formula(paste("RXMCD + 15 ~ t + p + tp + age +",

APPENDIX E (continued)

vlist[v], "+", vlist[v], "*t +", vlist[v], "*p +",
int.name, "+ male.f + black.f + hispanic.f + edu.f + ins.f + cc_3.f")
}

model <- glm2(fmla, data=big, family=Gamma(link="log"))
b.tp <- coef(model)['tp']
b.tpx <- coef(model)[int.name]
xb <- predict(model, type="link")
truth <- mean(ifelse(big$group==2, (exp(xb) - exp(xb-b.tp-big[vlist[v]]*b.tpx))[1],
(exp(xb+b.tp+big[vlist[v]]*b.tpx) - exp(xb))[1]))
assign(paste("truth.".,outcome,".".,vlist[v], sep=""), truth)

rm(fmla, outcome, int.name, model, b.tp, b.tpx, xb, truth)

# save truth estimates
my.list <- list(
  truth.RXTOT.age=truth.RXTOT.age,
  truth.RXTOT.povlev=truth.RXTOT.povlev,

  truth.RXTOT.age.s2.sm=truth.RXTOT.age.s2.sm,
  truth.RXTOT.age.s2.med=truth.RXTOT.age.s2.med,
  truth.RXTOT.age.s2.lar=truth.RXTOT.age.s2.lar,

  truth.RXTOT.age.s3.sm=truth.RXTOT.age.s3.sm,
  truth.RXTOT.age.s3.med=truth.RXTOT.age.s3.med,
  truth.RXTOT.age.s3.lar=truth.RXTOT.age.s3.lar,

  truth.RXTOT.age.s4.sm=truth.RXTOT.age.s4.sm,
  truth.RXTOT.age.s4.med=truth.RXTOT.age.s4.med,
  truth.RXTOT.age.s4.lar=truth.RXTOT.age.s4.lar,

  truth.RXTOT.povlev.s2.sm=truth.RXTOT.povlev.s2.sm,
  truth.RXTOT.povlev.s2.med=truth.RXTOT.povlev.s2.med,
  truth.RXTOT.povlev.s2.lar=truth.RXTOT.povlev.s2.lar,

  truth.RXTOT.povlev.s3.sm=truth.RXTOT.povlev.s3.sm,
  truth.RXTOT.povlev.s3.med=truth.RXTOT.povlev.s3.med,
  truth.RXTOT.povlev.s3.lar=truth.RXTOT.povlev.s3.lar,}
APPENDIX E (continued)

truth.RXTOT.povlev.s4.sm=truth.RXTOT.povlev.s4.sm,
truth.RXTOT.povlev.s4.med=truth.RXTOT.povlev.s4.med,
truth.RXTOT.povlev.s4.lar=truth.RXTOT.povlev.s4.lar,
truth.RXMCD.age=truth.RXMCD.age,
truth.RXMCD.povlev=truth.RXMCD.povlev,
truth.RXMCD.age.s2.sm=truth.RXMCD.age.s2.sm,
truth.RXMCD.age.s2.med=truth.RXMCD.age.s2.med,
truth.RXMCD.age.s2.lar=truth.RXMCD.age.s2.lar,
truth.RXMCD.age.s3.sm=truth.RXMCD.age.s3.sm,
truth.RXMCD.age.s3.med=truth.RXMCD.age.s3.med,
truth.RXMCD.age.s3.lar=truth.RXMCD.age.s3.lar,
truth.RXMCD.age.s4.sm=truth.RXMCD.age.s4.sm,
truth.RXMCD.age.s4.med=truth.RXMCD.age.s4.med,
truth.RXMCD.age.s4.lar=truth.RXMCD.age.s4.lar,
truth.RXMCD.povlev.s2.sm=truth.RXMCD.povlev.s2.sm,
truth.RXMCD.povlev.s2.med=truth.RXMCD.povlev.s2.med,
truth.RXMCD.povlev.s2.lar=truth.RXMCD.povlev.s2.lar,
truth.RXMCD.povlev.s3.sm=truth.RXMCD.povlev.s3.sm,
truth.RXMCD.povlev.s3.med=truth.RXMCD.povlev.s3.med,
truth.RXMCD.povlev.s3.lar=truth.RXMCD.povlev.s3.lar,
truth.RXMCD.povlev.s4.sm=truth.RXMCD.povlev.s4.sm,
truth.RXMCD.povlev.s4.med=truth.RXMCD.povlev.s4.med,
truth.RXMCD.povlev.s4.lar=truth.RXMCD.povlev.s4.lar)

} else if (subset=="Y") {

# create interactions
big.subset$age.pt <- big.subset$age * big.subset$p * big.subset$t
big.subset$age.s2.sm.pt <- big.subset$age.s2.sm * big.subset$p * big.subset$t
big.subset$age.s2.med.pt <- big.subset$age.s2.med * big.subset$p * big.subset$t
big.subset$age.s2.lar.pt <- big.subset$age.s2.lar * big.subset$p * big.subset$t
big.subset$age.s3.sm.pt <- big.subset$age.s3.sm * big.subset$p * big.subset$t
big.subset$age.s3.med.pt <- big.subset$age.s3.med * big.subset$p * big.subset$t
big.subset$age.s3.lar.pt <- big.subset$age.s3.lar * big.subset$p * big.subset$t
big.subset$age.s4.sm.pt <- big.subset$age.s4.sm * big.subset$p * big.subset$t
big.subset$age.s4.med.pt <- big.subset$age.s4.med * big.subset$p * big.subset$t
big.subset$age.s4.lar.pt <- big.subset$age.s4.lar * big.subset$p * big.subset$t
big.subset$povlev.pt <- big.subset$povlev * big.subset$p * big.subset$t
big.subset$povlev.s2.sm.pt <- big.subset$povlev.s2.sm * big.subset$p * big.subset$t
big.subset$povlev.s2.med.pt <- big.subset$povlev.s2.med * big.subset$p * big.subset$t
big.subset$povlev.s2.lar.pt <- big.subset$povlev.s2.lar * big.subset$p * big.subset$t
big.subset$povlev.s3.sm.pt <- big.subset$povlev.s3.sm * big.subset$p * big.subset$t
big.subset$povlev.s3.med.pt <- big.subset$povlev.s3.med * big.subset$p * big.subset$t
big.subset$povlev.s3.lar.pt <- big.subset$povlev.s3.lar * big.subset$p * big.subset$t
big.subset$povlev.s4.sm.pt <- big.subset$povlev.s4.sm * big.subset$p * big.subset$t
big.subset$povlev.s4.med.pt <- big.subset$povlev.s4.med * big.subset$p * big.subset$t
big.subset$povlev.s4.lar.pt <- big.subset$povlev.s4.lar * big.subset$p * big.subset$t

for(v in 1:length(vlist)) {
  outcome <- "TOTTCH"
  int.name <- paste(vlist[v],".pt",sep="")

  # confounded age
  if(v==1 | v==3 | v==5 | v==7 | v==9 | v==11 | v==13 | v==15 | v==17 | v==19) {
    fmla <- as.formula(paste("TOTTCH ~ t + p + tp + povlev +",
                              vlist[v], "+", vlist[v], "*t", vlist[v], "*p +",
                              int.name, "+ male.f + black.f + hispanic.f + edu.f + ins.f + cc_3.f"))
  }
  # confounded povlev
  if(v==2 | v==4 | v==6 | v==8 | v==10 | v==12 | v==14 | v==16 | v==18 | v==20) {
    fmla <- as.formula(paste("TOTTCH ~ t + p + tp + age +",
                              vlist[v], "+", vlist[v], "*t", vlist[v], "*p +",
                              int.name, "+ male.f + black.f + hispanic.f + edu.f + ins.f + cc_3.f"))
  }

  model <- glm2(fmla, data=big.subset, family=Gamma(link="log"))
  b.tp <- coef(model)["tp"]
  b.tpx <- coef(model)[int.name]
APPENDIX E (continued)

xb <- predict(model, type="link")
truth <- mean(ifelse(big.subset$group==2, (exp(xb) - exp(xb-b.tp-big[vlist[v]]*b.tpx))[1],
(exp(xb+b.tp+big[vlist[v]]*b.tpx) - exp(xb))[1]))
assign(paste("truth.",outcome,".",vlist[v], sep=""), truth)
rm(fmla, outcome, int.name, model, b.tp, b.tpx, xb, truth)

# save truth estimates
my.list <- list(
  truth.TOTTCH.age=truth.TOTTCH.age,
  truth.TOTTCH.povlev=truth.TOTTCH.povlev,

  truth.TOTTCH.age.s2.sm=truth.TOTTCH.age.s2.sm,
  truth.TOTTCH.age.s2.med=truth.TOTTCH.age.s2.med,
  truth.TOTTCH.age.s2.lar=truth.TOTTCH.age.s2.lar,

  truth.TOTTCH.age.s3.sm=truth.TOTTCH.age.s3.sm,
  truth.TOTTCH.age.s3.med=truth.TOTTCH.age.s3.med,
  truth.TOTTCH.age.s3.lar=truth.TOTTCH.age.s3.lar,

  truth.TOTTCH.age.s4.sm=truth.TOTTCH.age.s4.sm,
  truth.TOTTCH.age.s4.med=truth.TOTTCH.age.s4.med,
  truth.TOTTCH.age.s4.lar=truth.TOTTCH.age.s4.lar,

  truth.TOTTCH.povlev.s2.sm=truth.TOTTCH.povlev.s2.sm,
  truth.TOTTCH.povlev.s2.med=truth.TOTTCH.povlev.s2.med,
  truth.TOTTCH.povlev.s2.lar=truth.TOTTCH.povlev.s2.lar,

  truth.TOTTCH.povlev.s3.sm=truth.TOTTCH.povlev.s3.sm,
  truth.TOTTCH.povlev.s3.med=truth.TOTTCH.povlev.s3.med,
  truth.TOTTCH.povlev.s3.lar=truth.TOTTCH.povlev.s3.lar,

  truth.TOTTCH.povlev.s4.sm=truth.TOTTCH.povlev.s4.sm,
  truth.TOTTCH.povlev.s4.med=truth.TOTTCH.povlev.s4.med,
  truth.TOTTCH.povlev.s4.lar=truth.TOTTCH.povlev.s4.lar)
}
return(my.list)
}

APPENDIX E (continued)
APPENDIX F

Augmented inverse probability weighted estimator

# program name: aipw.R
# purpose: augmented inverse probability weighted DID estimator in
# Poisson, Gamma, and inverse Gaussian models
####################################################################
aipw <- function(data, outcome=outcome) {
  treatment.var = "t"
  treatment.vec <- data[, treatment.var]
  treatment.values <- sort(unique(treatment.vec))

  treated.data <- data[treatment.vec == treatment.values[2], ]
  control.data <- data[treatment.vec == treatment.values[1], ]
  n.treated.pre <- nrow(treated.data)
  n.control.pre <- nrow(control.data)

  # propensity score model
  glm.ps <- suppressWarnings(glm2(t ~ povlev + age + male.f + black.f + hispanic.f + edu.f +
                                ins.f + cc_3.f, family = binomial, data = data))
  pscore.probs <- suppressWarnings(predict(glm.ps, newdata = data, type = "response"))
  pscores.pre <- pscore.probs

  # outcome formula for treated
  if (outcome=="RXTOT") {
    glm.t <- suppressWarnings(glm2(y ~ povlev + age + male.f + black.f + hispanic.f + edu.f +
                                  ins.f + cc_3.f, family = poisson(link="log"), data = treated.data))
    glm.c <- suppressWarnings(glm2(y ~ povlev + age + male.f + black.f + hispanic.f + edu.f +
                                  ins.f + cc_3.f, family = poisson(link="log"), data = control.data))
  }
  if (outcome=="RXMCD") {
    glm.t <- suppressWarnings(glm2(y ~ povlev + age + male.f + black.f + hispanic.f + edu.f +
                                  ins.f + cc_3.f, family = Gamma(link="log"), data = treated.data))
    glm.c <- suppressWarnings(glm2(y ~ povlev + age + male.f + black.f + hispanic.f + edu.f +
                                  ins.f + cc_3.f, family = Gamma(link="log"), data = control.data))
  }
}


if (outcome=='TOTTCH') {
    glm.t <- suppressWarnings(glm2(y ~ povlev + age + male.f + black.f + hispanic.f + edu.f + ins.f + cc_3.f, family= inverse.gaussian(link="log"), data = treated.data))
    glm.c <- suppressWarnings(glm2(y ~ povlev + age + male.f + black.f + hispanic.f + edu.f + ins.f + cc_3.f, family= inverse.gaussian(link="log"), data = control.data))
    outcome.treated.expectation <- suppressWarnings(predict(glm.t, newdata = data, type = "response"))
    outcome.control.expectation <- suppressWarnings(predict(glm.c, newdata = data, type = "response"))
    res.c <- suppressWarnings(residuals(glm.c, type = "response"))
    res.t <- suppressWarnings(residuals(glm.t, type = "response"))

    n <- length(treatment.vec)
    term1 <- rep(0, n)
    term2 <- rep(0, n)
    term3 <- rep(0, n)
    term4 <- rep(0, n)
    control.indic <- treatment.vec == treatment.values[1]
    treated.indic <- treatment.vec == treatment.values[2]
    outcome.vec <- model.response(model.frame(y ~ povlev + age + male.f + black.f + hispanic.f + edu.f + ins.f + cc_3.f, data = data))
    term1 <- outcome.vec/pscore.probs
    term1[control.indic] <- 0
    term2 <- ((1 - pscore.probs) * outcome.treated.expectation)/pscore.probs
    term2[control.indic] <- ((0 - pscore.probs[control.indic]) * outcome.treated.expectation[control.indic])/pscore.probs[control.indic]
    term3 <- outcome.vec/(1 - pscore.probs)
    term3[treated.indic] <- 0
    term4 <- ((1 - pscore.probs) * outcome.control.expectation)/(1 - pscore.probs)
    term4[control.indic] <- ((0 - pscore.probs[control.indic]) * outcome.control.expectation[control.indic])/(1 - pscore.probs[control.indic])
    ATE.AIPW.hat <- (1/n) * (sum(term1) - sum(term2) - sum(term3) - sum(term4))
    I.hat <- term1 - term2 - term3 - term4 - ATE.AIPW.hat
    ATE.AIPW.sand.var <- (1/n^2) * sum(I.hat^2)
    ATE.AIPW.sand.se <- sqrt(ATE.AIPW.sand.var)
ATE.reg.hat <- mean(outcome.treated.expectation) - mean(outcome.control.expectation)
weight.norm.1 <- 1/sum(treated.indic/pscore.probs)
weight.norm.3 <- 1/sum((1 - treated.indic)/(1 - pscore.probs))
ATE.IPW.hat <- weight.norm.1 * sum(term1) - weight.norm.3 * sum(term3)

return(list(ATE.AIPW.hat = ATE.AIPW.hat, ATE.reg.hat = ATE.reg.hat, ATE.IPW.hat = ATE.IPW.hat,
ATE.AIPW.sand.SE = ATE.AIPW.sand.se))
}
APPENDIX G

Difference-in-differences estimators

#########################################################
# program name: DID.sim.R
# purpose: Multivariate simulation of MEPS data, evaluation of 13
# DID estimators
#########################################################

library(PoisBinOrdNonNor)
library(glm2)
library(modmarg)
library(nnet)
library(boot)
library(effsize)

rm(list = ls())

# set working directory
setwd("D:/")

# load MEPS data
meps <- read.csv(file="D:/MEPSdata/MEPS_22 Jan 2018.csv")

# Format MEPS dataset
names(meps)[names(meps)=='trt_NE_S'] <- 't'
names(meps)[names(meps)=='post'] <- 'p'
names(meps)[names(meps)=='AGE'] <- 'age'
names(meps)[names(meps)=='POVLEV'] <- 'povlev'
meps$tp <- meps$t * meps$p

# change male to numeric variable
meps$male <- as.numeric(meps$SEX)

# change insurance to numeric variable
meps$ins <- as.numeric(meps$INSCOV)

# format to factor variables
meps$male.f <- factor(meps$male)

# force female to be reference category
meps$male.f <- relevel(meps$male.f, ref = 2)
meps$black.f <- factor(meps$black)
meps$hispanic.f <- factor(meps$hispanic)
meps$edu.f <- factor(meps$edu)
meps$ins.f <- factor(meps$ins)
meps$cc_3.f <- factor(meps$cc_3)
one.96 <- -qnorm(0.025)

# define groups
meps$group[which(meps$t == 1 & meps$p==0)] <- 1
meps$group[which(meps$t == 1 & meps$p==1)] <- 2
meps$group[which(meps$t == 0 & meps$p==0)] <- 3
meps$group[which(meps$t == 0 & meps$p==1)] <- 4

# subset: total charges excluding prescriptions between $0-10,000 for those without private insurance
meps.subset <- meps[which(meps$TOTTCH > 0 & meps$TOTTCH < 10000 & meps$ins.f != 1),]
# adjust the number of factors of insurance variable from 3 to 2
meps.subset$ins.f <- factor(meps.subset$ins.f)
#dim(meps.subset)  # 2592

# Call programs
source("simulation/meps.sim.dataset.R")
source("simulation/get.truth.marg.R")
source("simulation/aipw.R")

# Large simulated dataset
# generate big dataset
#big <- meps.sim.dataset(s.seed=69300, meps.n.1=9160, meps.n.2=8710, meps.n.3=26360, meps.n.4=24070)
#save(big, file = "simulation/datasets/2018-04-22/big_69300.RData")
load(file = "simulation/datasets/2018-04-22/big_69300.RData")

# generate big dataset of subted group
#big.subset <- meps.sim.dataset(s.seed=69300, meps.n.1=4010, meps.n.2=3540, meps.n.3=9820, meps.n.4=8550,
#meps.dat=meps.subset, subset="Y")
#save(big.subset, file = "simulation/datasets/2018-04-22/big_subset_25920.RData")
load(file = "simulation/datasets/2018-04-22/big_subset_25920.RData")

# effect sizes of counfounding for each group
# 0.2= small, 0.5= medium, 0.8= large
APPENDIX G (continued)

factors <- list(
    age=c(0,0,0,0),
    povlev=c(0,0,0,0),
    age.s2.sm=c(0.26,0.26,0,0),
    povlev.s2.sm=c(0.21,0.21,0,0),
    age.s2.med=c(0.55,0.55,0,0),
    povlev.s2.med=c(0.52,0.52,0,0),
    age.s2.lar=c(0.87,0.86,0,0),
    povlev.s2.lar=c(0.83,0.83,0,0),
    age.s3.sm=c(0.26,0.26+0.17,0,0.18),
    povlev.s3.sm=c(0.21,0.21+0.16,0,0.17),
    age.s3.med=c(0.55,0.55+0.40,0,0.44),
    povlev.s3.med=c(0.52,0.52+0.38,0,0.38),
    age.s3.lar=c(0.87,0.86+0.59,0,0.61),
    povlev.s3.lar=c(0.83,0.83+0.57,0,0.58),
    age.s4.sm=c(0.24,0.24-0.15,0,0.18),
    povlev.s4.sm=c(0.21,0.21-0.15,0,0.1),
    age.s4.med=c(0.55,0.55-0.35,0,0.18),
    povlev.s4.med=c(0.52,0.52-0.35,0,0.16),
    age.s4.lar=c(0.87,0.86-0.55,0,0.28),
    povlev.s4.lar=c(0.83,0.83-0.55,0,0.27))

vlist <- names(factors)

# create confounded age variable
confound.age <- function(dat, var, age.sd, d=0) {
    c <- unlist(factors[var])
    var <- rep(NA, dim(dat)[1])
    var[which(dat$group==1)] <- (c[1] * age.sd[1]) + dat$age[which(dat$group==1)]
    var[which(dat$group==2)] <- (c[2] * age.sd[2]) + dat$age[which(dat$group==2)]
    var[which(dat$group==3)] <- (c[3] * age.sd[3]) + dat$age[which(dat$group==3)]
    var[which(dat$group==4)] <- (c[4] * age.sd[4]) + dat$age[which(dat$group==4)]
    # truncate and round new age variable
    # if age is <18 make=18
    var[which(var < 18)] <- 18
    # if age is >=65 make=64
    var[which(var >= 65)] <- 64.99
    # round down age to nearest integer
    var <- floor(var)
APPENDIX G (continued)

```r
if(d==1) {
    print(cohen.d(d=var[which(dat$group==1)], f=dat$age[which(dat$group==1)]))
    print(cohen.d(d=var[which(dat$group==2)], f=dat$age[which(dat$group==2)]))
    print(cohen.d(d=var[which(dat$group==3)], f=dat$age[which(dat$group==3)]))
    print(cohen.d(d=var[which(dat$group==4)], f=dat$age[which(dat$group==4)]))
}
}

# create confounded poverty level variable
confound.povlev <- function(dat, var, povlev.sd, d=0) {
    c <- unlist(factors[var])
    var <- rep(NA, dim(dat)[1])
    var[which(dat$group==1)] <- (c[1] * povlev.sd[1]) + dat$povlev[which(dat$group==1)]
    var[which(dat$group==2)] <- (c[2] * povlev.sd[2]) + dat$povlev[which(dat$group==2)]
    var[which(dat$group==3)] <- (c[3] * povlev.sd[3]) + dat$povlev[which(dat$group==3)]
    var[which(dat$group==4)] <- (c[4] * povlev.sd[4]) + dat$povlev[which(dat$group==4)]
    # truncate and round povlev variable
    # if povlev is <0 make=0
    var[which(var < 0)] <- 0
    # if povlev is >138 make=138
    var[which(var >= 138)] <- 137.99
    if(d==1) {
        print(cohen.d(d=var[which(dat$group==1)], f=dat$povlev[which(dat$group==1)]))
        print(cohen.d(d=var[which(dat$group==2)], f=dat$povlev[which(dat$group==2)]))
        print(cohen.d(d=var[which(dat$group==3)], f=dat$povlev[which(dat$group==3)]))
        print(cohen.d(d=var[which(dat$group==4)], f=dat$povlev[which(dat$group==4)]))
    }
    var
}

########################################
# create confounded age and povlev in big datasets and obtain true DID treatment effect

## decriptive of age and poverty level in original data
age.1 <- mean(big$age[which(big$group==1)]); age.1.sd <- sd(big$age[which(big$group==1)]); n1 <-
table(big$group)[1]
age.2 <- mean(big$age[which(big$group==2)]); age.2.sd <- sd(big$age[which(big$group==2)]); n2 <-
table(big$group)[2]
```
APPENDIX G (continued)

age.3 <- mean(big$age[which(big$group==3)]); age.3.sd <- sd(big$age[which(big$group==3)]); n3 <- table(big$group)[3]
age.4 <- mean(big$age[which(big$group==4)]); age.4.sd <- sd(big$age[which(big$group==4)]); n4 <- table(big$group)[4]
age.sd.big <- c(age.1.sd, age.2.sd, age.3.sd, age.4.sd)
povlev.1 <- mean(big$povlev[which(big$group==1)]); povlev.1.sd <- sd(big$povlev[which(big$group==1)]); n1 <- table(big$group)[1]
povlev.2 <- mean(big$povlev[which(big$group==2)]); povlev.2.sd <- sd(big$povlev[which(big$group==2)]); n2 <- table(big$group)[2]
povlev.3 <- mean(big$povlev[which(big$group==3)]); povlev.3.sd <- sd(big$povlev[which(big$group==3)]); n3 <- table(big$group)[3]
povlev.4 <- mean(big$povlev[which(big$group==4)]); povlev.4.sd <- sd(big$povlev[which(big$group==4)]); n4 <- table(big$group)[4]
povlev.sd.big <- c(povlev.1.sd, povlev.2.sd, povlev.3.sd, povlev.4.sd)

## create confounded variables for age
# for(i in c(1,3,5,7,9,11,13,15,17,19)) {
#   big[vlist[i]] <- confound.age(dat=big, var=vlist[i], d=1)
# }
## create confounded variables for povlev
# for(i in c(2,4,6,8,10,12,14,16,18,20)) {
#   big[vlist[i]] <- confound.povlev(dat=big, var=vlist[i], d=1)
# }

# decriptives of age and poverty level in the subgroup
age.1 <- mean(big.subset$age[which(big.subset$group==1)]); age.1.sd <- sd(big.subset$age[which(big.subset$group==1)]); n1 <- table(big.subset$group)[1]
age.2 <- mean(big.subset$age[which(big.subset$group==2)]); age.2.sd <- sd(big.subset$age[which(big.subset$group==2)]); n2 <- table(big.subset$group)[2]
age.3 <- mean(big.subset$age[which(big.subset$group==3)]); age.3.sd <- sd(big.subset$age[which(big.subset$group==3)]); n3 <- table(big.subset$group)[3]
age.4 <- mean(big.subset$age[which(big.subset$group==4)]); age.4.sd <- sd(big.subset$age[which(big.subset$group==4)]); n4 <- table(big.subset$group)[4]
age.sd.subset <- c(age.1.sd, age.2.sd, age.3.sd, age.4.sd)
povlev.1 <- mean(big.subset$povlev[which(big.subset$group==1)]); povlev.1.sd <- sd(big.subset$povlev[which(big.subset$group==1)]); n1 <- table(big.subset$group)[1]
povlev.2 <- mean(big.subset$povlev[which(big.subset$group==2)]); povlev.2.sd <- sd(big.subset$povlev[which(big.subset$group==2)]); n2 <- table(big.subset$group)[2]
povlev.3 <- mean(big.subset$povlev[which(big.subset$group==3)]); povlev.3.sd <- sd(big.subset$povlev[which(big.subset$group==3)]); n3 <- table(big.subset$group)[3]
povlev.4 <- mean(big.subset$povlev[which(big.subset$group==4)]); povlev.4.sd <- sd(big.subset$povlev[which(big.subset$group==4)]); n4 <- table(big.subset$group)[4]
povlev.sd.subset <- c(povlev.1.sd, povlev.2.sd, povlev.3.sd, povlev.4.sd)

## create confounded variables for age
# for(i in c(1,3,5,7,9,11,13,15,17,19)) {
#   big.subset[vlist[i]] <- confound.age(dat=big.subset, var=vlist[i], d=0)
#}
## create confounded variables for povlev
# for(i in c(2,4,6,8,10,12,14,16,18,20)) {
#   big.subset[vlist[i]] <- confound.povlev(dat=big.subset, var=vlist[i], d=0)
#}

# get truth marginal effect (discrete-difference) estimates from big dataset
# get truth for outcomes RXTOT and RXMCD on big dataset
# all.truth <- get.truth.marg()
# save(all.truth, file = "simulation/datasets/2018-04-22/_truth_marg.RData")
load(file = "simulation/datasets/2018-04-22/_truth_marg.RData")

# get truth for outcome TOTTCH on big subsetted dataset
# all.truth.subset <- get.truth.marg(subset="Y")
# save(all.truth.subset, file = "simulation/datasets/2018-04-22/_truth_marg_subset.RData")
load(file = "simulation/datasets/2018-04-22/_truth_marg_subset.RData")

#############################

# generate and save simulated datasets
# for (i in 1:1000) {
#   # simulate full dataset (n=6830)
#   meps.sim <- meps.sim.dataset(s.seed=i)
#   save(meps.sim, file = paste("simulation/datasets/2018-04-22/meps.sim_", s.seed=i, ".RData", sep=""))
#   # simulate subseted dataset (n=2592)
#   meps.sim <- meps.sim.dataset(s.seed=i, meps.n.1=401, meps.n.2=354, meps.n.3=982, meps.n.4=855,
APPENDIX G (continued)

# meps.dat=meps.subset, subset="Y")
# save(meps.sim, file = paste("simulation/datasets/2018-04-22/meps.sim.subset_", s.seed=i, ".RData", sep=""))
#
##########################################################
# Model 1:  Naive FE
# Model 2:  Naive FE with X
# Model 3:  FE weighted separately at each time point
# Model 4a:  FE weighted separately at each time point + X
# Model 4b:  FE weighted separately at each time point + X, bootstrap basic
# Model 4c:  FE weighted separately at each time point + X, bootstrap perc
# Model 5:  FE Stuart's 4-group weighted design
# Model 6a:  FE Stuart's 4-group weighted design + X
# Model 6b:  FE Stuart's 4-group weighted design + X, bootstrap basic
# Model 6c:  FE Stuart's 4-group weighted design + X, bootstrap perc
# Model 7a:  AIPW, sandwich
# Model 7b:  AIPW, bootstrap basic
# Model 7c:  AIPW, bootstrap perc
##########################################################

sim <- function(outcome, v, nsims, n.boot) {

  # seed for bootstrap estimates reproducibility
  set.seed(109876)

  bias <- mse <- coverage <- matrix(NA, nrow=nsims, ncol=13)
  colnames(bias) <- colnames(mse) <- colnames(coverage) <-
    c("1. Naive", "2. Naive with X",
      "5. 4-grp wts", "6a. 4-grp wts+X", "6b. 4-grp wts+X. basic bs", "6c. 4-grp wts+X. perc bs",
      "7a. AIPW. sand", "7b. AIPW. bs basic", "7c. AIPW. bs perc")

  all.estimates <- matrix(NA, nrow=nsims, ncol=13*3)
  colnames(all.estimates) <-
    c("DID.1", "cll.1", "clu.1",
      "DID.2", "cll.2", "clu.2",
      "DID.3", "cll.3", "clu.3",
      "DID.4", "cll.4", "clu.4",
      "DID.5", "cll.5", "clu.5",
      "DID.6", "cll.6", "clu.6",
      "DID.7", "cll.7", "clu.7",
      "DID.8", "cll.8", "clu.8",
      "DID.9", "cll.9", "clu.9",
      "DID.10", "cll.10", "clu.10",
      "DID.11", "cll.11", "clu.11",
      "DID.12", "cll.12", "clu.12")

APPENDIX G (continued)

"DID.3", "cll.3", "clu.3",
"DID.4", "cll.4", "clu.4",
"DID.5", "cll.5", "clu.5",
"DID.6", "cll.6", "clu.6",
"DID.7", "cll.7", "clu.7",
"DID.8", "cll.8", "clu.8",
"DID.9", "cll.9", "clu.9",
"DID.10", "cll.10", "clu.10",
"DID.11", "cll.11", "clu.11",
"DID.12", "cll.12", "clu.12",
"DID.13", "cll.13", "clu.13")

##### Poisson and Gamma Models #####

ptm <- proc.time()
my.var <- vlist[v]

# Model formulas with covariates, confounded for age or povlev
fmla.age <- as.formula(paste("y ~ t + p + tp + povlev +",my.var,"+ male.f + black.f + hispanic.f +
edu.f + ins.f + cc_3.f"))
fmla.povlev <- as.formula(paste("y ~ t + p + tp +",my.var,"+ age + male.f + black.f + hispanic.f +
edu.f + ins.f + cc_3.f"))

# define formulas for v
if(v==1 | v==3 | v==5 | v==7 | v==9 | v==11 | v==13 | v==15 | v==17) {
  fmla <- fmla.age
  fmla.group <- as.formula(paste("group ~ povlev +", my.var, "+ male.f + black.f + hispanic.f +
edu.f + ins.f + cc_3.f"))
  fmla.ps <- as.formula(paste("t~ povlev +", my.var, "+ male.f + black.f + hispanic.f + edu.f +
  ins.f + cc_3.f"))
} else {
  fmla <- fmla.povlev
  fmla.group <- as.formula(paste("group ~", my.var, "+ age + male.f + black.f + hispanic.f +
edu.f + ins.f + cc_3.f"))
  fmla.ps <- as.formula(paste("t~", my.var, "+ age + male.f + black.f + hispanic.f + edu.f +
  ins.f + cc_3.f"))
}

# truth and update desc stats if using the full dataset (for outcomes RXTOT and RXMCD)
APPENDIX G (continued)

```r
if(outcome=="RXTOT") truth <- unlist(all.truth[paste("truth.",outcome,".",vlist[v], sep=""\)])
if(outcome=="RXMCD") truth <- unlist(all.truth[paste("truth.",outcome,".",vlist[v], sep=""\)])
if(outcome=="TOTTCH") truth <- unlist(all.truth.subset[paste("truth.",outcome,".",vlist[v], sep=""\)])
for (i in 1:nsims) {
  # load simulated dataset
  if(outcome=="RXTOT" | outcome=="RXMCD") load(file = paste("simulation/datasets/2018-04-22/meps.sim_", s.seed=i, ",",RData", sep=""\)])
  if(outcome=="TOTTCH") load(file = paste("simulation/datasets/2018-04-22/meps.sim.subset ", s.seed=i, ",",RData", sep=""\)])
  # print iteration
  print(i)
  # outcomes and truth
  # number of prescriptions
  if(outcome=="RXTOT") meps.sim$y <- meps.sim$RXTOT
  # cost of Medicaid prescriptions. add small positive value to y for gamma model
  if(outcome=="RXMCD") meps.sim$y <- meps.sim$RXMCD + 15
  # total charges
  if(outcome=="TOTTCH") meps.sim$y <- meps.sim$TOTTCH
  # create confounded variables
  if(v==1 | v==3 | v==7 | v==9 | v==11 | v==13 | v==15 | v==17) {
    #age.1.sd <- sd(meps.sim$age[which(meps.sim$group==1)])
    #age.2.sd <- sd(meps.sim$age[which(meps.sim$group==2)])
    #age.3.sd <- sd(meps.sim$age[which(meps.sim$group==3)])
    #age.4.sd <- sd(meps.sim$age[which(meps.sim$group==4)])
    #age.sd <- c(age.1.sd, age.2.sd, age.3.sd, age.4.sd)
    if(outcome=="RXTOT" | outcome=="RXMCD") age.sd <- age.sd.big
    if(outcome=="TOTTCH") age.sd <- age.sd.subset
    meps.sim[vlist[v]] <- confound.age(dat=meps.sim, var=vlist[v], age.sd=age.sd)
  } else {
    #povlev.1.sd <- sd(meps.sim$povlev[which(meps.sim$group==1)])
    #povlev.2.sd <- sd(meps.sim$povlev[which(meps.sim$group==2)])
    #povlev.3.sd <- sd(meps.sim$povlev[which(meps.sim$group==3)])
  }
}
```
APPENDIX G (continued)

```r
# povlev.4.sd <- sd(meps.sim$povlev[which(meps.sim$group==4)])
# povlev.sd <- c(povlev.1.sd, povlev.2.sd, povlev.3.sd, povlev.4.sd)
if(outcome=="RXTOT" | outcome=="RXMCD") povlev.sd <- povlev.sd.big
if(outcome=="TOTTCH") povlev.sd <- povlev.sd.subset

meps.sim[vlist[v]] <- confound.povlev(dat=meps.sim, var=vlist[v], povlev.sd)
```

```r
# prepare data
# group <- c( rep(1, length(which(meps.sim$t == 1 & meps.sim$p==0))),
#             rep(2, length(which(meps.sim$t == 1 & meps.sim$p==1))),
#             rep(3, length(which(meps.sim$t == 0 & meps.sim$p==0))),
#             rep(4, length(which(meps.sim$t == 0 & meps.sim$p==1)))) )
# group <- c(rep(1, 916), rep(2, 871), rep(3, 2636), rep(4, 2407))
predata <- meps.sim[which(meps.sim$p==0),]
postdata <- meps.sim[which(meps.sim$p==1),]
wts1 <- rep(1,6830)  # for marginal effect estimates of weighted estimators

# Model 1: Two-way fixed effects
if (outcome=="RXTOT") model1.glm <- glm2(y ~ t + p + tp, data=meps.sim, family=poisson(link="log"))
if (outcome=="RXMCD") model1.glm <- glm2(y ~ t + p + tp, data=meps.sim, family=Gamma(link="log"))
if (outcome=="TOTTCH") model1.glm <- glm2(y ~ t + p + tp, data=meps.sim,
                          family=inverse.gaussian(link="log"))
    marg.modell.glm <- marg(model1.glm, var_interest = 'tp', type = 'effects')
est <- marg.modell.glm[[1]][2,2]

```
APPENDIX G (continued)

```r
marg.model2.glm <- marg(model2.glm, var_interest = 'tp', type = 'effects')
est <- marg.model2.glm[1][2,2]

bias[i,2] <- est-truth
mse[i,2] <- (est-truth)^2
coverage[i,2] <- ifelse((truth > marg.model2.glm[1][6][2]) & (truth <
marg.model2.glm[1][7][2]), 1, 0)
all.estimates[i,4] <- est
all.estimates[i,5] <- marg.model2.glm[1][6][2]
all.estimates[i,6] <- marg.model2.glm[1][7][2]

# Models 3, 4: Separate time weights
# propensity score model
prelogit <- glm(fmla.ps, data=predata, family="binomial")
postlogit <- glm(fmla.ps, data=postdata, family="binomial")
pre.pscores <- predict(prelogit, type="response")
post.pscores <- predict(postlogit, type="response")

# generate weights
meps.sim$wts.sep[meps.sim$group==1] <- 1/pre.pscores[predata$t==1]
meps.sim$wts.sep[meps.sim$group==2] <- 1/post.pscores[postdata$t==1]
meps.sim$wts.sep[meps.sim$group==3] <- 1/(1-pre.pscores[predata$t==0])
meps.sim$wts.sep[meps.sim$group==4] <- 1/(1-post.pscores[postdata$t==0])

# Model 3: Separate time weights
if (outcome=="RXTOT") model3.glm <- glm2(y ~ t + p + tp, weights=wts.sep, data=meps.sim,
family=poisson(link="log"))
if (outcome=="RXMCD") model3.glm <- glm2(y ~ t + p + tp, weights=wts.sep, data=meps.sim,
family=Gamma(link="log"))
if (outcome=="TOTTCH") model3.glm <- glm2(y ~ t + p + tp, weights=wts.sep, data=meps.sim,
family=inverse.gaussian(link="log"))
marg.model3.glm <- suppressWarnings(marg(model3.glm, var_interest = 'tp', type = 'effects', weights =
wts1))
est <- marg.model3.glm[1][2,2]

bias[i,3] <- est-truth
mse[i,3] <- (est-truth)^2
```

APPENDIX G (continued)

coverage[i,3] <- ifelse((truth > marg.model3.glm[[1]][[6]][2]) & (truth <
marg.model3.glm[[1]][[7]][2]), 1, 0)
all.estimates[i,7] <- est
all.estimates[i,8] <- marg.model3.glm[[1]][[6]][2]
all.estimates[i,9] <- marg.model3.glm[[1]][[7]][2]

# Model 4a: Separate time weights + X
if (outcome == "RXTOT") model4.glm <- glm2(fmla, weights=wts.sep, data=meps.sim,
family=poisson(link="log"))
if (outcome == "RXMCD") model4.glm <- glm2(fmla, weights=wts.sep, data=meps.sim,
family=Gamma(link="log"))
if (outcome == "TOTTCH") model4.glm <- glm2(fmla, weights=wts.sep, data=meps.sim,
family=inverse.gaussian(link="log"))
marg.model4.glm <- suppressWarnings(marg(model4.glm, var_interest = 'tp', type = 'effects', weights = wts1))
est <- marg.model4.glm[[1]][2,2]
# an alternate, equivalent method of getting the marginal effect is:
#b.tp <- coef(model4.glm)['tp']
#xb <- predict(model4.glm, type="link")
#est.alt <- mean(ifelse(meps.sim$group==2, exp(xb) - exp(xb-b.tp), exp(xb+b.tp) - exp(xb)))
#est.alt

bias[i,4] <- est-truth
mse[i,4] <- (est-truth)^2
coverage[i,4] <- ifelse((truth > marg.model4.glm[[1]][[6]][2]) & (truth <
marg.model4.glm[[1]][[7]][2]), 1, 0)
all.estimates[i,10] <- est
all.estimates[i,11] <- marg.model4.glm[[1]][[6]][2]
all.estimates[i,12] <- marg.model4.glm[[1]][[7]][2]

# Model 4b, 4c: Separate time weights + X, bootstrap
sep.time.wtd <- function(dat, ind) {
d <- dat[ind,]  # allows boot to select sample
    if (outcome == "RXTOT") model56.glm <- glm2(fmla, weights=wts.sep, data=d,
family=poisson(link="log"))
    if (outcome == "RXMCD") model56.glm <- glm2(fmla, weights=wts.sep, data=d,
family=Gamma(link="log"))
APPENDIX G (continued)

if (outcome=="TOTTCH")
  model56.glm <- glm2(fmla, weights=wts.sep, data=d,
                   family=inverse.gaussian(link="log"))
  suppressWarnings(marg(model56.glm, var_interest = 'tp', type = 'effects', weights = wts1))[[1]][2,2]
}

boot.sep.time.wtd <- boot(meps.sim, sep.time.wtd, R=n.boot)
est <- median(boot.sep.time.wtd$t)

bias[i,5] <- bias[i,6] <- est-truth
mse[i,5] <- mse[i,6] <- (est-truth)^2
q.025 <- quantile(boot.sep.time.wtd$t, c(0.025, 0.975))[[1]]
q.975 <- quantile(boot.sep.time.wtd$t, c(0.025, 0.975))[[2]]
coverage[i,5] <- ifelse((truth > 2*est - q.975) & (truth < 2*est - q.025), 1, 0)
coverage[i,6] <- ifelse((truth > q.025) & (truth < q.975), 1, 0)
all.estimates[i,13] <- est
all.estimates[i,14] <- 2*est - q.975
all.estimates[i,15] <- 2*est - q.025
all.estimates[i,16] <- est
all.estimates[i,17] <- q.025
all.estimates[i,18] <- q.975

# Model 5: 4-group weighted
# propensity score model
mlogit <- multinom(fmla.group, data=meps.sim)
probs <- predict(mlogit, type="probs")

# generate weights
meps.sim$wts.4grp[meps.sim$group==1] <- probs[meps.sim$group==1,1]/probs[meps.sim$group==1,1]
meps.sim$wts.4grp[meps.sim$group==2] <- probs[meps.sim$group==2,1]/probs[meps.sim$group==2,2]
meps.sim$wts.4grp[meps.sim$group==3] <- probs[meps.sim$group==3,1]/probs[meps.sim$group==3,3]
meps.sim$wts.4grp[meps.sim$group==4] <- probs[meps.sim$group==4,1]/probs[meps.sim$group==4,4]

if (outcome=="RXTOT")
  model7.glm <- glm(y ~ t + p + tp, weights=wts.4grp, data=meps.sim,
                   family=poisson(link="log"))
if (outcome=="RXMCD")
  model7.glm <- glm2(y ~ t + p + tp, weights=wts.4grp, data=meps.sim,
                    family=Gamma(link="log"))
if (outcome=="TOTTCH")
  model7.glm <- glm2(y ~ t + p + tp, weights=wts.4grp, data=meps.sim,
                    family=inverse.gaussian(link="log"))
APPENDIX G (continued)

marg.model7.glm <- suppressWarnings(marg(model7.glm, var_interest = 'tp', type = 'effects'))
est <- marg.model7.glm[[1]][2,2]

bias[i,7] <- est - truth
mse[i,7] <- (est - truth)^2
coverage[i,7] <- ifelse((truth > marg.model7.glm[[1]][[6]][2]) & (truth < marg.model7.glm[[1]][[7]][2]), 1, 0)
all.estimates[i,19] <- est
all.estimates[i,20] <- marg.model7.glm[[1]][[6]][2] # ci.model7[1]

# Model 6a: 4-group weighted + X
if (outcome == "RXTOT")
  model8.glm <- glm2(fmla, weights=wts.4grp, data=meps.sim, family=poisson(link="log"))
else
  model8.glm <- glm2(fmla, weights=wts.4grp, data=meps.sim, family=Gamma(link="log"))

marg.model8.glm <- suppressWarnings(marg(model8.glm, var_interest = 'tp', type = 'effects', weights = wts1))
est <- marg.model8.glm[[1]][2,2]

bias[i,8] <- est - truth
mse[i,8] <- (est - truth)^2
coverage[i,8] <- ifelse((truth > marg.model8.glm[[1]][[6]][2]) & (truth < marg.model8.glm[[1]][[7]][2]), 1, 0)
all.estimates[i,22] <- est
all.estimates[i,23] <- marg.model8.glm[[1]][[6]][2] # ci.model8[1]

# Model 6b, c: 4-group weighted + X, bootstrap
four.grp.wtd <- function(dat, ind) {
d <- dat[ind] # allows boot to select sample
  if (outcome == "RXTOT")
    model910.glm <- glm2(fmla, weights=wts.4grp, data=d, family=poisson(link="log"))
  else
    model910.glm <- glm2(fmla, weights=wts.4grp, data=d, family=Gamma(link="log"))
}
APPENDIX G (continued)

    if (outcome=="TOTTCH") model910.glm <- glm2(fmla, weights=wts.4grp, data=d, 
            family=inverse.gaussian(link="log")
        suppressWarnings(marg(model910.glm, var_interest = 'tp', type = 'effects', weights =
                          wts1))[[1]][2,2] 
    boot.four.grp.wtd <- boot(meps.sim, four.grp.wtd, R=n.boot) 
    est <- median(boot.four.grp.wtd$t)
    bias[i,9] <- bias[i,10] <- est - truth 
    mse[i,9] <- mse[i,10] <- (est - truth)^2 
    q.025 <- quantile(boot.four.grp.wtd$t, c(0.025, 0.975))[1] 
    q.975 <- quantile(boot.four.grp.wtd$t, c(0.025, 0.975))[2] 
    coverage[i,9] <- ifelse((truth > 2*est - q.975) & (truth < 2*est - q.025), 1, 0) 
    coverage[i,10] <- ifelse((truth > q.025) & (truth < q.975), 1, 0) 
    all.estimates[i,25] <- est 
    all.estimates[i,26] <- 2*est - q.975 
    all.estimates[i,27] <- 2*est - q.025 
    all.estimates[i,28] <- est 
    all.estimates[i,29] <- q.025 
    all.estimates[i,30] <- q.975 

    # Model 7a: AIPW, sandwich standard error 
    aipw.pre <- aipw(predata, outcome=outcome) 
    aipw.post <- aipw(postdata, outcome=outcome) 
    est <- aipw.post$ATE.AIPW.hat - aipw.pre$ATE.AIPW.hat 
    # sandwich standard error 
    sand.SE <- sqrt((aipw.pre$ATE.AIPW.sand.SE)^2 + (aipw.post$ATE.AIPW.sand.SE)^2) 
    bias[i,11] <- est - truth 
    mse[i,11] <- (est-truth)^2 
    coverage[i,11] <- ifelse((truth > est - one.96*sand.SE) & (truth < est + one.96*sand.SE), 1, 0) 
    all.estimates[i,31] <- est 
    all.estimates[i,32] <- est - one.96*sand.SE 
    all.estimates[i,33] <- est + one.96*sand.SE 

    # Model 7b, c: AIPW, bootstrap 
    aipw.diff <- rep(NA, n.boot) 
    est <- rep(NA, n.boot) 
    for (biter in 1:n.boot) {
        
    }

    boot.four.grp.wtd <- boot(meps.sim, four.grp.wtd, R=n.boot) 
    est <- median(boot.four.grp.wtd$t) 
    bias[i,9] <- bias[i,10] <- est - truth 
    mse[i,9] <- mse[i,10] <- (est - truth)^2 
    q.025 <- quantile(boot.four.grp.wtd$t, c(0.025, 0.975))[1] 
    q.975 <- quantile(boot.four.grp.wtd$t, c(0.025, 0.975))[2] 
    coverage[i,9] <- ifelse((truth > 2*est - q.975) & (truth < 2*est - q.025), 1, 0) 
    coverage[i,10] <- ifelse((truth > q.025) & (truth < q.975), 1, 0) 
    all.estimates[i,25] <- est 
    all.estimates[i,26] <- 2*est - q.975 
    all.estimates[i,27] <- 2*est - q.025 
    all.estimates[i,28] <- est 
    all.estimates[i,29] <- q.025 
    all.estimates[i,30] <- q.975 

    # Model 7a: AIPW, sandwich standard error 
    aipw.pre <- aipw(predata, outcome=outcome) 
    aipw.post <- aipw(postdata, outcome=outcome) 
    est <- aipw.post$ATE.AIPW.hat - aipw.pre$ATE.AIPW.hat 
    # sandwich standard error 
    sand.SE <- sqrt((aipw.pre$ATE.AIPW.sand.SE)^2 + (aipw.post$ATE.AIPW.sand.SE)^2) 
    bias[i,11] <- est - truth 
    mse[i,11] <- (est-truth)^2 
    coverage[i,11] <- ifelse((truth > est - one.96*sand.SE) & (truth < est + one.96*sand.SE), 1, 0) 
    all.estimates[i,31] <- est 
    all.estimates[i,32] <- est - one.96*sand.SE 
    all.estimates[i,33] <- est + one.96*sand.SE 

    # Model 7b, c: AIPW, bootstrap 
    aipw.diff <- rep(NA, n.boot) 
    est <- rep(NA, n.boot) 
    for (biter in 1:n.boot) {
APPENDIX G (continued)

```r
boot.inds.pre <- sample(1:dim(predata)[1], dim(predata)[1], replace = TRUE)
d.pre <- predata[boot.inds.pre, ]
boot.inds.post <- sample(1:dim(postdata)[1], dim(postdata)[1], replace = TRUE)
d.post <- postdata[boot.inds.post, ]
#
# aipw_did= aipw_post - aipw_pre
aipw.diff[biter] <- aipw(data = d.post, outcome=outcome)$ATE.AIPW.hat -
aipw(data = d.pre, outcome=outcome)$ATE.AIPW.hat
}
est <- median(aipw.diff)
# take 2.5 and 97.5th percentile of bootstrap estimates
q.025 <- quantile(aipw.diff,c(0.025, 0.975))[1]
q.975 <- quantile(aipw.diff,c(0.025, 0.975))[2]

bias[i,12] <- bias[i,13] <- est-truth
mse[i,12] <- mse[i,13] <- (est-truth)^2
coverage[i,12] <- ifelse((truth > 2*est - q.975) & (truth < 2*est - q.025), 1, 0)
coverage[i,13] <- ifelse((truth > q.025) & (truth < q.975), 1, 0)
all.estimates[i,34] <- all.estimates[i,37] <- est
all.estimates[i,35] <- 2*est - q.975
all.estimates[i,36] <- 2*est - q.025
all.estimates[i,38] <- q.025
all.estimates[i,39] <- q.975

run.time.all <- proc.time() - ptm
# run.time in minutes
run.time <- round(run.time.all[3] / 60,2)

bias
mse
coverage
all.estimates

bias.final <- rmse.final <- cov.final <- matrix(NA, nrow=1, ncol=13)
colnames(bias.final) <- colnames(rmse.final) <- colnames(cov.final) <-
```
"5. 4-grp wts", "6a. 4-grp wts+X", "6b. 4-grp wts+X. basic bs", "6c. 4-grp wts+X. perc bs", "7a. AIPW. sand", "7b. AIPW. bs basic", "7b. AIPW. bs perc")

all.estimates.final <- matrix(NA, nrow=1, ncol=13*3)
colnames(all.estimates.final) <-
c("DID.1", "cll.1", "clu.1",
  "DID.2", "cll.2", "clu.2",
  "DID.3", "cll.3", "clu.3",
  "DID.4", "cll.4", "clu.4",
  "DID.5", "cll.5", "clu.5",
  "DID.6", "cll.6", "clu.6",
  "DID.7", "cll.7", "clu.7",
  "DID.8", "cll.8", "clu.8",
  "DID.9", "cll.9", "clu.9",
  "DID.10", "cll.10", "clu.10",
  "DID.11", "cll.11", "clu.11",
  "DID.12", "cll.12", "clu.12",
  "DID.13", "cll.13", "clu.13")
bias.final[1,] <- t(round(apply(bias, 2, mean),4))
rmse.final[1,] <- t(round(sqrt(apply(mse, 2, mean)),4))
cov.final[1,] <- t(round(apply(coverage, 2, mean),3))
all.estimates.final[1,] <- t(round(apply(all.estimates, 2, mean),3))

# output system time and simulation settings
sim.settings <- data.frame(outcome, vlist[v], truth, nsims, n.boot, run.time)

# Output simulation results to Excel
setwd("simulation/output/")
file.name <- paste(outcome,"_TEST_",vlist[v],"_DID.sim.csv", sep = "")
write.table(bias.final, file = file.name,
eol = "\r", sep = "", row.names = FALSE)
write.table(rmse.final, file = file.name,
eol = "\r", sep = "", row.names = FALSE, append=TRUE)
write.table(cov.final, file = file.name,
eol = "\r", sep = "", row.names = FALSE, append=TRUE)
write.table(all.estimates.final, file = file.name,
eol = "\r", sep = "", row.names = FALSE, append=TRUE)
write.table(sim.settings, file = file.name,
eol = "\r", sep = "", col.names = TRUE, row.names = FALSE, append=TRUE)
APPENDIX G (continued)

# call simulation
# sim(outcome="TOTTCH", v=6, nsims=10, n.boot=200)

# v = confounded variable
# 1 "age"       # 11 "age.s3.med"
# 2 "povlev"    # 12 "povlev.s3.med"
# 3 "age.s2.sm"  # 13 "age.s3.lar"
# 4 "povlev.s2.sm" # 14 "povlev.s3.lar"
# 5 "age.s2.med"  # 15 "age.s4.sm"
# 6 "povlev.s2.med" # 16 "povlev.s4.sm"
# 7 "age.s2.lar"   # 17 "age.s4.med"
# 8 "povlev.s2.lar" # 18 "povlev.s4.med"
# 9 "age.s3.sm"   # 19 "age.s4.lar"
# 10 "povlev.s3.sm" # 20 "povlev.s4.lar"

# nsims= number of simulations
# n.boot= number of bootstrap replications
LAUREN AMELIA ABDERHALDEN

EDUCATION

<table>
<thead>
<tr>
<th>Date Range</th>
<th>Degree/Title</th>
<th>Institution</th>
</tr>
</thead>
<tbody>
<tr>
<td>08/2011 – 12/2018</td>
<td>Ph.D. in Biostatistics, School of Public Health, University of Illinois at Chicago</td>
<td></td>
</tr>
<tr>
<td>09/2002 – 06/2007</td>
<td>B.A. in Mathematics, University of California at San Diego</td>
<td></td>
</tr>
</tbody>
</table>

EXPERIENCE

<table>
<thead>
<tr>
<th>Date Range</th>
<th>Position/Responsibilities</th>
</tr>
</thead>
<tbody>
<tr>
<td>04/2018 – present</td>
<td>Biostatistician, Centre on Aging and Mobility, University Hospital Zurich, Switzerland</td>
</tr>
<tr>
<td></td>
<td>- Statistical analysis of clinical trials</td>
</tr>
<tr>
<td></td>
<td>- Report, manuscript, and grant writing</td>
</tr>
<tr>
<td></td>
<td>- Research and methodological support to collaborators and students</td>
</tr>
<tr>
<td></td>
<td>- SAS and R programming</td>
</tr>
<tr>
<td>01/2016 – 05/2016</td>
<td>Teaching Assistant for BSTT 506 Clinical Trials, School of Public Health, University of Illinois at Chicago, United States</td>
</tr>
<tr>
<td></td>
<td>- Provide guidance to students in a graduate class for clinical trials</td>
</tr>
<tr>
<td></td>
<td>- Grade homework assignments</td>
</tr>
<tr>
<td></td>
<td>- Power and sample size calculations for clinical trials using SAS</td>
</tr>
<tr>
<td>01/2013 – 01/2016</td>
<td>Mathematical Statistician, Center of Innovation for Complex Chronic Healthcare (CINCCH), Edward Hines Jr. VA Hospital, Department of Veterans Affairs, United States</td>
</tr>
<tr>
<td></td>
<td>- Provide statistical expertise in study design, statistical analysis, grant and manuscript writing, and power and sample size analysis</td>
</tr>
<tr>
<td></td>
<td>- Develop innovative methods to provide study-specific solutions</td>
</tr>
<tr>
<td></td>
<td>- SAS, Stata, and R programming</td>
</tr>
<tr>
<td>08/2011 – 12/2012</td>
<td>Graduate Research Assistant, Institute for Health Research and Policy (IHRP), University of Illinois at Chicago, United States</td>
</tr>
<tr>
<td></td>
<td>- Statistical modeling for PhD Researchers and Physicians</td>
</tr>
<tr>
<td></td>
<td>- SAS and R programming</td>
</tr>
<tr>
<td>10/2009 – 07/2011</td>
<td>Research Statistician, Health Policy Research Institute (HPRI), University of California at Irvine, United States</td>
</tr>
<tr>
<td></td>
<td>- Provide statistical expertise to Health Economists of potential implications of various health policy strategies</td>
</tr>
<tr>
<td></td>
<td>- Statistical modeling and programming in SAS and Stata</td>
</tr>
<tr>
<td></td>
<td>- One-on-one instruction for subjects: Statistics, Calculus, Algebra, Trigonometry, Geometry, and Linear Algebra</td>
</tr>
</tbody>
</table>
LANGUAGES AND PROGRAMMING SKILLS

Languages  English (Native), German (A1)
Programming  R, SAS, Stata (all advanced)

PUBLICATIONS

Google Scholar: https://tinyurl.com/y8bb7hlg


