

Summer 8-2019

NOVEL PROPENSITY SCORE METHODS FOR MULTIPLE AND CONTINUOUS TREATMENTS: APPLICATIONS TO EHR DATA

DEREK W. BROWN

UTHealth School of Public Health

Follow this and additional works at: https://digitalcommons.library.tmc.edu/uthsph_dissertsopen



Part of the [Community Psychology Commons](#), [Health Psychology Commons](#), and the [Public Health Commons](#)

Recommended Citation

BROWN, DEREK W., "NOVEL PROPENSITY SCORE METHODS FOR MULTIPLE AND CONTINUOUS TREATMENTS: APPLICATIONS TO EHR DATA" (2019). *UT School of Public Health Dissertations (Open Access)*. 91.

https://digitalcommons.library.tmc.edu/uthsph_dissertsopen/91

This is brought to you for free and open access by the School of Public Health at DigitalCommons@TMC. It has been accepted for inclusion in UT School of Public Health Dissertations (Open Access) by an authorized administrator of DigitalCommons@TMC. For more information, please contact digitalcommons@library.tmc.edu.

NOVEL PROPENSITY SCORE METHODS FOR MULTIPLE AND CONTINUOUS
TREATMENTS: APPLICATIONS TO EHR DATA

by

DEREK W BROWN, BS, MS

APPROVED:

STACIA DESANTIS, PHD

MICHAEL SWARTZ, PHD

ANNA WILKINSON, PHD

DEAN, THE UNIVERSITY OF TEXAS
SCHOOL OF PUBLIC HEALTH

Copyright
by
Derek W Brown, BS, MS, PhD
2019

DEDICATION

To My Mother, Diane Lea Brown

NOVEL PROPENSITY SCORE METHODS FOR MULTIPLE AND CONTINUOUS
TREATMENTS: APPLICATIONS TO EHR DATA

by

DEREK W BROWN
BS, The University of Kansas, 2012
MS, The George Washington University, 2014

Presented to the Faculty of The University of Texas

School of Public Health

in Partial Fulfillment

of the Requirements

for the Degree of

DOCTOR OF PHILOSOPHY

THE UNIVERSITY OF TEXAS
SCHOOL OF PUBLIC HEALTH
Houston, Texas
August, 2019

ACKNOWLEDGEMENTS

I would like to thank my dissertation supervisor and academic advisor Dr. Stacia DeSantis, for her motivation and encouragement. This work would not have been accomplished without her constant guidance and support. I would also like to thank Dr. Michael Swartz and Dr. Anna Wilkinson for their feedback and guidance throughout this research process. Finally, I would like to acknowledge the entire EHR collaboration working group for their help with cleaning and supplying the necessary data to complete this work.

This work would not have been possible without the support given to me by my family and friends. Thank you for listening to me talk/vent about upcoming tests, coding issues, work, and most importantly, Kansas basketball ('Rock Chalk'). I would especially like to thank my wife, Abby Brown, for moving to Houston and letting me pursue this goal. I want to also thank her for letting me take over the dining room (and guest room) as I finished my work. I would also like to give a special welcome my twin daughters, Parker Sue and Eleanor Lea, who decided to come into our lives a bit early, but we could not be happier! Finally, I would like to give a special thank you to my dog, Penny, who never left my side even during late night coding/writing sessions.

This research was graciously supported by the NIGMS Predoctoral Training grant (T32 GM074902).

NOVEL PROPENSITY SCORE METHODS FOR MULTIPLE AND CONTINUOUS TREATMENTS: APPLICATIONS TO EHR DATA

Derek W Brown, BS, MS, PhD
The University of Texas
School of Public Health, 2019

Dissertation Chair: Stacia M DeSantis, PhD

Propensity scoring is often utilized to overcome the challenges posed by covariate imbalance to make causal inferences within observational studies. While methods for utilizing propensity scoring in a binary treatment case are well studied and established, generalizations to multiple unordered (multinomial) and continuous treatments are more complicated. In Aim 1, we developed and tested a novel multinomial treatment propensity score method, the GPS-CDF method, which derives a single scalar balancing score that can match and stratify subjects. Simulation results showed superior performance of the new methodology compared to standard multinomial propensity score methods. The proposed GPS-CDF method was also applied to an electronic health records study to determine the causal relationship between vasopressor choice and mortality in patients with non-traumatic aneurysmal subarachnoid hemorrhage (SAH). The GPS-CDF method indicated that phenylephrine may be the superior vasopressor choice for patients that present with non-traumatic SAH. We further applied the GPS-CDF method to the Emergency Truncal Hemorrhage Control Study to determine whether emerging hemorrhage control interventions influence patient mortality. Based on the GPS-CDF method, patients receiving resuscitative

endovascular balloon occlusion of the aorta (REBOA) had similar morality as patients who received Laparotomy. In Aim 2, we extended the GPS-CDF method to the continuous treatment setting and further introduced the npGPS-CDF method. Both novel methods use empirical cumulative distribution functions (CDF) in order to stratify subjects based on pretreatment confounders to produce causal estimates. A detailed simulation study showed superiority of the novel methods based on the empirical CDF when compared to standard weighting techniques. The proposed methods were applied to the “Mexican American Tobacco use in Children” (MATCh) study and found a significant association between exposure to smoking imagery in movies and smoking initiation among Mexican-American adolescents. Finally in Aim 3, we developed an R package for researchers to implement the proposed GPS-CDF method in practice. Overall this research provides investigators with new options for implementing multinomial and continuous treatment propensity scoring.

TABLE OF CONTENTS

| | |
|--|-----|
| List of Tables | i |
| List of Figures | ii |
| List of Appendices | iii |
| Background | 1 |
| 1.1 Introduction | 1 |
| Current Methods for Multinomial Treatments | 4 |
| 1.1.1 Generalized Propensity Score | 4 |
| 1.1.2 Distance Metrics | 5 |
| 1.1.3 Clustering Techniques | 6 |
| 1.1.4 Machine Learning Methods | 7 |
| 1.1.5 Stratification, Matching, Adjustment | 9 |
| Current Methods for Continuous Treatments | 10 |
| 1.1.6 Maximum Likelihood Estimation | 11 |
| 1.1.7 Generalized Boosted Model | 13 |
| 1.1.8 Covariate Balancing Generalized Propensity Score | 14 |
| 1.2 Public Health Significance | 16 |
| 1.3 Specific Aims | 17 |
| 1.3.1 To develop a novel method of propensity score analysis for multinomial treatments | 17 |
| 1.3.2 To develop a novel method of propensity score analysis for continuous treatments | 18 |
| 1.3.3 To develop an R package to implement multiple treatments propensity scoring methods | 18 |
| A Novel Approach for Propensity Score Matching and Stratification in the Presence of Multiple Treatments: Application to an EHR-Derived Study of Subarachnoid Hemorrhage | 25 |
| 2.1 Introduction | 27 |
| 2.2 Methods | 32 |
| 2.2.1 GPS-CDF Matching | 35 |
| 2.2.2 GPS-CDF Stratification | 37 |

| | |
|--|-----|
| 2.3 Simulation Study | 38 |
| 2.4 Results..... | 40 |
| 2.5 Data Applications | 43 |
| 2.5.1 Cerner Health Facts database..... | 44 |
| 2.5.2 Emergency Truncal Hemorrhage Control Study | 47 |
| 2.6 Discussion | 48 |
| 2.7 Conclusion..... | 51 |
| Sampling Based Propensity Score Stratification for Continuous Treatments | 67 |
| 3.1 Introduction | 69 |
| 3.2 Methods..... | 75 |
| 3.2.1 GPS-CDF - parametric approach | 75 |
| 3.2.2 npGPS-CDF - nonparametric approach | 78 |
| 3.3 Simulation Study | 81 |
| 3.4 Results..... | 82 |
| 3.5 Data Application: Effect of exposure to smoking imagery on smoking initiation in youth..... | 85 |
| 3.6 Discussion | 88 |
| 3.7 Conclusion..... | 91 |
| Generalized Propensity Score Cumulative Distribution Function (GPS-CDF) R Package | 104 |
| 4.1 Introduction | 105 |
| 4.2 R Documentation | 105 |
| 4.2.1 User Defined Inputs | 106 |
| 4.2.2 GPS-CDF Package Outputs..... | 106 |
| 4.2.3 GPS-CDF Package Examples | 107 |
| 4.3 Illustrative Data Example..... | 110 |
| 4.3.1 Obtaining the Scalar Balancing Score..... | 112 |
| 4.3.2 Outcome Analyses | 114 |
| 4.4 Conclusion..... | 116 |
| Conclusions and Future Work | 119 |

| | |
|---------------------|-----|
| 5.1 Conclusion..... | 119 |
| Appendices..... | 123 |

List of Tables

| | |
|---|----|
| Table 2.1. True association between baseline covariates with treatment and outcome. Note, x_1 , x_2 , x_4 , and x_5 are simulated to be pretreatment confounders. | 53 |
| Table 2.2. Model estimates after applying each analytical method to SAH patients to determine the association between vasopressor and mortality within the EHR dataset. Outcome models (with the exception of the unadjusted model) adjusted for age at baseline, gender, race, and marital status. | 54 |
| Table 3.1. Association of covariates with treatment and outcome. From the table, it may be noted that x_1 , x_2 , x_4 , and x_5 are simulated to be pretreatment confounders. | 93 |
| Table 3.2. Model estimates after GPS-CDF and npGPS-CDF stratification from the MATCH study. Note: HS = High School, POE = Positive outcome expectation, TAS = Thrill and adventure seeking, DAA = Drug and alcohol, SD = Social disinhibition score, SSS = Subjective Social Status. | 94 |

List of Figures

| | |
|--|-----|
| Figure 2.1. Graphical representation of the convex and concave modeling produced by the power function..... | 55 |
| Figure 2.2. Graphical representation of the covariate balance achieved by each method under the correctly specified and incorrectly specified treatment assignment models. SMD was calculated for all baseline covariates within each treatment pair, and the maximum SMD across treatment pairs was retained. | 56 |
| Figure 2.3. Distribution of the ATE for each method under each scenario between treatment 1 and treatment 3. The true ATE value of 0.4 is included as the dotted horizontal line..... | 57 |
| Figure 2.4. Graphical representation of the covariate balance achieved by each method for SAH patients within the Cerner Health Facts EHR database. The left plot presents the maximum pairwise SMD across treatment groups for each potential confounder. The right plot presents the average pairwise SMD across treatment groups for each potential confounder. | 58 |
| Figure 2.5. Graphical representation of the covariate balance achieved by each method for hemorrhage patients within the Emergency Truncal Hemorrhage Control Study. The plot presents the average pairwise SMD across treatment groups for each potential confounder..... | 59 |
| Figure 3.2. Distribution of the ATE for each method under each scenario. The true ATE value is included as the dotted horizontal line. | 96 |
| Figure 3.3. Graphical representation of the covariate balance achieved by each propensity score method within the MATCH study. The left plot presents the absolute Pearson correlation between treatment and each potential confounder (including square terms). The right plot presents F -statistics obtained from regressing T on each potential confounder one at a time. | 97 |
| Supplemental Figure 1. Distribution of the ATE for each method under each scenario between treatment 1 and treatment 2. The true ATE value of 0.7 is included as the dotted horizontal line..... | 123 |
| Supplemental Figure 2. Distribution of the ATE for each method under each scenario between treatment 2 and treatment 3. The true ATE value of -0.3 is included as the dotted horizontal line..... | 124 |
| Supplemental Figure 3. Graphical representation of the CDF mapping produced by the GPS-CDF method under 3, 4, 6, and 10 multinomial treatment group scenarios, respectively..... | 125 |

List of Appendices

| | |
|--|-----|
| Appendix A. Supplemental Figures | 123 |
| Appendix B. Code for implementation of GPS-CDF method using R software | 126 |

CHAPTER I

Background

1.1 Introduction

While randomized experiments are considered the gold standard when evaluating causal treatment effects, it is sometimes difficult to implement this design due to potential logistical and ethical issues, high cost, and low generalizability to a larger population. Instead, researchers often use observational studies to measure treatment effects. Currently, data from large observational studies including national surveys, electronic health records (EHR), and genome wide association studies (GWAS) are becoming publicly available. Although there is an influx in the amount of data available, treatment assignment in observational studies is not randomly assigned. Thus subjects receiving a treatment may differ from subjects not receiving a treatment based on one or more covariates. This covariate imbalance between treatment groups makes causal inference more challenging.

Propensity scoring is often utilized to overcome the challenges posed by covariate imbalances to make causal inference. In a binary treatment case, which includes one treatment and one control group, the propensity score is the probability of receiving the treatment conditional on a given set of observed covariates (Rosenbaum and Rubin, 1983). This probability (commonly called the propensity score) can be calculated using standard regression techniques (typically logistic regression) with the treatment (Z) being considered the outcome and the covariates (\mathbf{X}) the predictors. Treatment and control subjects with

similar estimated values for their propensity scores will have, on average, similar sets of covariate vectors (Greene, 2017). Since it can be used to remove covariate imbalance between treatment groups, the propensity score is known as a balancing score (Greene, 2017).

In order for the propensity score to be used to conduct causal inference in observational studies, a few assumptions need to be met, namely: consistency, exchangeability, positivity, and no misspecification of the propensity score model (Austin and Stuart, 2015). Consistency implies that a subject's potential outcome under the treatment they received is equal to the subject's observed outcome. Exchangeability assumes all true confounders for treatment assignment and the relationship with the outcome are observed and measured. Although not testable, conducting propensity scoring without measuring all possible variables that influence treatment assignment and the outcome can result in biased estimates of the treatment effect. The positivity assumption states that all subjects have a non-zero (positive) probability of receiving each treatment. This assumption can be tested by confirming overlap of histograms or boxplots of each subject's propensity score stratified by treatment group. Finally, misspecification of the propensity score model, although formally untestable, seeks to find the 'true' propensity score to balance covariates between treatment groups. In practice, there are many balancing scores that can remove covariate imbalance between groups (Rosenbaum and Rubin, 1983), so this assumption is considered met if covariate balance is achieved between treatment groups (Austin and Stuart, 2015).

When conducting causal inference using propensity scoring, researchers are interested in two possible summary measures of the treatment effect, the average treatment effect (ATE) or the average treatment effect among the treated (ATT). ATE is of interest for

comparisons of the mean outcome when the entire population is eligible for all treatments (McCaffrey et al., 2013). In a binary treatment setting with one treatment and one control group, ATE is the effect of giving the treatment to the entire population instead of giving the control to the entire population. ATE is calculated by taking the expectation across the *entire* population and is given by:

$$ATE_{k,k'} = E[Y(k) - Y(k')] = E[Y(k)] - E[Y(k')] \quad (1.1)$$

where Y is the outcome for the comparison of treatment k and treatment k' . ATT is of interest when comparing the effectiveness of a particular treatment relative to the alternatives available to the population of interest (McCaffrey et al., 2013). Thus in the binary treatment setting, ATT finds the effect of the treatment only on those *who actually received* the treatment. The formal definition of ATT is given by:

$$ATT_{k,k'} = E[Y(k') | Z = k] - E[Y(k) | Z = k] \quad (1.2)$$

where Y is the outcome and Z is the treatment of interest.

It has been shown that the difference between treatment and control subjects at each value of a balancing score is an unbiased unit-level estimate of the ATE or ATT if treatment assignment between subjects is independent given a set of covariates (Rosenbaum and Rubin, 1983). That is to say, within matched pairs or strata of a balancing score, treatment assignment is independent of observed covariates. By using a balancing score for matching, stratification, or adjustment, the outcome analysis will produce unbiased estimates of ATE or ATT. Thus by using the propensity score in the final outcome analysis, researchers are able to make causal inferences with observational data.

Current Methods for Multinomial Treatments

While methods for utilizing propensity scoring in a binary treatment case are well studied and established, generalizations to multiple unordered (multinomial) treatments are more complicated. Intuitively in this treatment setting, analyses among multinomial treatments can be broken down into a series of binary comparisons (Lechner, 2001; 2002). For example, in an experimental setting with 3 treatment arms (A, B, and C), analyses can be conducted within treatment pairs (A,B), (A,C), and (B,C). This decomposition facilitates the utilization of binary propensity score methods on each of the pairwise treatment comparisons. Alternatively, using common reference matching, propensity scores can be used to match subjects, for example, within the pairs (A,B) and (A,C). Then using these two matched samples, a final cohort of 1:1:1 matched triplets can be created (Rassen et al., 2011). This final cohort would include subjects who received treatment A that were matched to both a subject receiving treatment B and a subject receiving treatment C (Rassen et al., 2011). Although these two approaches enable the implementation of standard binary propensity scoring, the results are compromised by limited external validity. Furthermore, these approaches are only able to estimate ATT, which is not always the estimate of interest. This makes it difficult to identify a superior treatment for the general population which is a common goal in many analyses (Lopez and Gutman, 2017).

1.1.1 Generalized Propensity Score

Instead of decomposing the multiple treatment setting into binary treatment comparisons, the generalized propensity score (GPS) can be used to extend the theory of

causal inference from a binary treatment setting to a multiple treatment setting (Joffe and Rosenbaum, 1999; Imbens, 2000; Imai and Van Dyk, 2004). The GPS is defined as the probability of receiving one of K treatments conditional on a given set of observed covariates (Imbens, 2000). Unlike the binary treatment case where the propensity score is a single value representing the probability a subject was treated, the GPS is a vector, of length K , representing the probabilities of a subject being treated under each of the K conditions.

Commonly used methods for propensity scoring in the presence of multinomial treatments have relied on the GPS vector that is produced from some type of multinomial regression model, e.g.,

$$\log \left[\frac{Pr(Z_i = k)}{Pr(Z_i = K)} \right] = \theta_k + x_i' \beta_k \quad (1.3)$$

where θ_k is a constant, β_k is a vector of regression coefficients, Z is the treatment received, and K is the total number of treatments, for $k = \{1, 2, \dots, K - 1\}$. In this nominal case, treatments do not follow any set order, so there is no defined relationship between the various treatment assignments.

1.1.2 Distance Metrics

Distance metrics can be used to match subjects with similar GPS vector distributions. Aitchison distance, a compositional data analysis tool, has been proposed as one way to match subjects based on the relative distance of their GPS vectors (Seya and Yoshida, 2017). Additionally, Rassen et al. developed the ‘within-trio’ matching algorithm that finds a triplet of patients, with different treatments, while minimizing the within-trio distance (Rassen et

al., 2013). Although these procedures successfully match subjects, there are potential drawbacks. First, they cannot be used to derive ATE, in the case of Aitchison distance, and second, they cannot be used for more than three treatments, in the case of ‘within-trio’ matching. Mahalanobi’s distance, a commonly used multivariate distance metric, can also be used to match subjects (Rubin, 1979; Zhao, 2004). Instead of using the GPS vector, Mahalanobi’s distance matches subjects with similar covariate distributions. While Mahalanobi’s distance has been shown to be effective for matching in analyses involving a limited number of covariates (Rubin, 1979; Zhao, 2004), it does not perform well when there are more than 8 covariates or when the covariates are not normally distributed (for example if they are non-continuous (Gu and Rosenbaum, 1993; Stuart, 2010)). Additionally, distance metrics need to be modified to enable matching across different treatments since subjects within a treatment group will more likely have similar distances.

1.1.3 Clustering Techniques

Extensions to the above include clustering techniques as a possible method to group subjects with similar GPS vectors. Tu et al. demonstrated how four popular clustering techniques could be used to group subjects with similar observed covariate distributions based on a transformation of the GPS vector (Tu et al., 2013). Ultimately, Tu et al. showed that k-Means clustering (KMC), which minimizes the sum of squares between a subject in a cluster and the centroid of that cluster, provides the best covariate similarity between subjects in different treatment groups (Tu et al., 2013; Lopez and Gutman, 2017). Lopez et al. further extended these methods by combining KMC and 1:1 matching to create a matched analysis cohort (Lopez and Gutman, 2017). Lopez et al. first utilized KMC to place subjects into

clusters with similar values for one or more components of the GPS vector, and subsequently matched subjects (within each sub-cluster) using standard propensity score techniques (Lopez and Gutman, 2017). Matching within sub-clusters ensures that subjects will have matched values for one component of the GPS vector and similar values for the other components. However, with clustering, there exists the possibility of obtaining clusters without representation from every possible treatment group (Lopez and Gutman, 2017). Moreover, after running KMC, Lopez et al. limits matching to within each cluster which may lead to some possible matches not being considered by the method (Lopez and Gutman, 2017). Furthermore, as clustering techniques utilize distance metrics in their algorithms, they are subject to the same limitations as Mahalanobi's distance when used to balance covariates.

1.1.4 Machine Learning Methods

Although multinomial regression is the most commonly used method for estimating the GPS vector, other techniques may be considered. Notably, non-parametric machine learning methods of estimating the GPS vector, such as the generalized boosted model (GBM), recursive partitioning, and neural nets have been proposed (McCaffrey et al., 2004; Setoguchi et al., 2008; McCaffrey et al., 2013; Burgette et al., 2017). Although not well-studied, GBM and other tree-based methods could provide a few notable benefits over parametric regression. For example, variable selection including the decision to include higher order or interaction terms in the model occurs automatically, unlike with parametric models. This is of particular importance when working with 'big data' (i.e. EHR, GWAS, etc.) as there are a large number of covariates available to be selected for the GPS (McCaffrey et al., 2013). Further, the iterative estimation procedure used by GBM, which fits

regression trees that maximize log likelihood in order to produce a piecewise constant model that perfectly fits the data, can easily be fine-tuned to provide the propensity score model with the best balance between treatment groups (McCaffrey et al., 2013). After estimating the GPS vector, inverse probability weighting (where the weight is the inverse propensity of the treatment an individual actually received (Imbens, 2000; Feng et al., 2012; McCaffrey et al., 2013; Burgette et al., 2017)) can be applied directly to the outcome (Feng et al., 2012) or be utilized within weighted regression models (McCaffrey et al., 2013; Burgette et al., 2017) to estimate ATEs. Additionally, when the weights derived by multiplying the inverse probability weight by the probability an individual received the target treatment, weighted regression models can estimate ATTs (McCaffrey et al., 2013; Burgette et al., 2017).

However, while it might initially appear that non-parametric methods of estimating the GPS vector could be immediately adapted to the multiple treatment setting in conjunction with inverse probability weighting (IPW), this approach is limiting for several notable reasons. First, because non-parametric methods only estimate a GPS vector, matching and stratification in the outcome analysis cannot be performed since no obvious scalar balancing score can currently be produced from the resultant vector. Thus machine learning algorithms are limited to IPW to derive treatment effects; however, IPW may produce unreliable outcome estimates with large sample variances due to extreme weights (Busso et al., 2014; Lopez and Gutman, 2017; Li et al., 2018). Alternative weighting methods have been proposed that are not as susceptible to these extremes (Hirano and Imbens, 2001; Imai and Ratkovic, 2014; Li et al., 2018), but since weighting directly uses the scalar estimated propensity score in determining the effect of treatment (Austin et al., 2007; Rubin, 2004), as Rubin (Rubin, 2004) suggests, this results in the greatest sensitivity to misspecification of the

propensity score. Furthermore, the use of GBM may suffer from issues of power when some (but not complete) structure in the data can be assumed. Finally, the black box nature of nonparametric approaches prevent the user from being able to discern how the model specifically utilizes each variable in producing the propensity scores (Ridgeway et al., 2016). These limitations have impeded the use of nonparametric methods in propensity scoring over the last decade, despite their promise.

1.1.5 Stratification, Matching, Adjustment

Even though a function that maps the GPS vector to a scalar balancing score does not currently exist, methods for covariate balancing using stratification, matching, and adjustment based on the GPS vector have been studied in the multiple treatment setting. Stratification techniques using the GPS vector were first described by Zanutto et al. (Zanutto et al., 2005) and further extended by Huang et al (Huang et al., 2005). Huang et al. showed that stratification of subjects by propensity scores at each treatment level, in combination with weighted averages, can produce estimates of the average potential outcome (Huang et al., 2005). Yang et al. utilized a similar stratification and weighting method for covariate balancing, but further extended the method to match subjects in order to produce causal treatment effects (Yang et al., 2016). A similar matching procedure was also used by Lechner (Lechner, 2001). Additionally, Feng et al. was able to estimate ATEs by using generalized linear models to assess the relationship between outcomes and the GPS vector; then by using this model, estimate the expected outcome of a subject under a certain treatment given the GPS (Feng et al., 2012).

Though these procedures are more relatable to the methods available in standard binary propensity scoring, a cornerstone of these approaches is the estimation of average potential outcomes that are performed separately for each treatment level. This does simplify the inherent problems that arise when creating balance among multiple treatment groups, but since these approaches are only concentrating on one element of the GPS vector instead of the full GPS vector (Greene, 2017), they might suffer from a loss of information and thus not create the best possible covariate balance. Additionally, the methods proposed by Huang et al. (Huang et al., 2005) and Yang et al. (Yang et al., 2016) suffer from their inability to adjust for covariates in the outcome model and further lack the flexibility to be applied to complex analyses.

Current Methods for Continuous Treatments

Although methods for propensity scoring in both binary and multiple treatment settings have been well studied (Rosenbaum and Rubin, 1983, 1984, 1985; Joffe and Rosenbaum, 1999; Imbens, 2000; Imai and Van Dyk, 2004), there has been less research devoted to propensity score methods for continuous treatments. In practice when presented with continuous treatments, researchers often dichotomize or categorize the treatment in order to utilize standard and well established propensity score techniques (e.g. Chertow et al., 2004; Davidson et al., 2006; Donohue and Ho, 2007; Flores-Lagunes, Gonzalez, and Neumann, 2007; Harder et al., 2008; Boyd et al., 2010; Nielsen et al., 2011; Greene, 2017). Although propensity score methods are correctly applied in these settings, categorization of a continuous treatment may lead to loss of information and power during the outcome analysis (Royston et al., 2006; Zhu et al., 2015; Fong et al., 2018).

1.1.6 Maximum Likelihood Estimation

Instead of decomposing the continuous treatment setting into binary or categorical treatment comparisons, maximum likelihood estimates (MLE) derived from linear models have been proposed to estimate the generalized propensity score (GPS) (Robins et al., 2000; Imai and Van Dyk, 2004; Hirano and Imbens, 2004). In the continuous treatment framework, $r(t, x)$, the conditional density of the treatment given the covariates, is defined as

$$r(t, x) = f_{T|X}(t|x) \quad (1.4)$$

where T represents the continuous treatment and \mathbf{X} the covariates of interest (Hirano and Imbens, 2004). Thus the GPS for continuous treatments is defined as $R = r(T, X)$ (Hirano and Imbens, 2004). In practice, this GPS can be estimated by fitting a linear regression model of the form

$$T = \boldsymbol{\beta}\mathbf{X} + \varepsilon \quad (1.5)$$

where $\varepsilon \sim N(0, \sigma^2)$. Then the GPS is estimated as

$$\hat{R}_i = \frac{1}{\sqrt{2\pi\hat{\sigma}^2}} \exp\left(-\frac{1}{2\hat{\sigma}^2}(T_i - \hat{T}_i)^2\right) \quad (1.6)$$

for the i^{th} individual (Hirano and Imbens, 2004). Although \hat{R}_i , applied directly in regression adjustment (Hirano and Imbens, 2004), or the scalar value, $\hat{\boldsymbol{\beta}}\mathbf{X}_i$, utilized for matching or stratification (Imai and Van Dyk, 2004), can produce causal estimates, using the GPS in weighted outcome analyses has been given more attention recently (Robins et al., 2000; Zhu et al., 2015; Schuler et al., 2016; Fong et al., 2018; Austin, 2018a; Austin, 2018b).

Robins et al. proposed using the GPS to produce causal estimates by implementing inverse probability weighting (IPW) (Robins et al., 2000). Briefly, IPW seeks to weight subjects with dissimilar covariate distributions higher than subjects with similar covariate profiles within the same treatment. That is to say, IPW gives higher weight to subjects, under a certain treatment, who have similar covariate distributions as subjects who received a different treatment. In the calculation of \widehat{R}_i , subjects with unexpected covariate distributions will have large estimates for $T_i - \widehat{T}_i$ and conversely have small values for \widehat{R}_i . Thus the IPW, given by

$$w_i = \frac{1}{\widehat{R}_i} \quad (1.7)$$

will be higher, effectively giving more weight to subjects that have unexpected covariate distributions based on their treatment assignment.

Although weights using the GPS can be applied directly in the form given above, Robins et al. warn that w_i has infinite variance and a stabilizing factor should be applied to w_i to be used in practice (Robins et al., 2000). This stabilizing factor, $W(T_i)$, is given by the marginal density of T and can be estimated by first fitting an intercept only model of the form

$$T = \beta_0 + \varepsilon \quad (1.8)$$

where $\varepsilon \sim N(0, \sigma_{sample}^2)$. Then the stabilizing factor is estimated as

$$\widehat{W}(T_i) = \frac{1}{\sqrt{2\pi\widehat{\sigma}_{sample}^2}} \exp\left(-\frac{1}{2\widehat{\sigma}_{sample}^2}(T_i - \widehat{\mu})^2\right) \quad (1.9)$$

where $\hat{\mu}$ is the mean treatment value of the sample (Austin, 2018b). Thus the final estimated stabilized IPW is given by

$$sw_i = \frac{\widehat{W}(T_i)}{\widehat{R}_i}. \quad (1.10)$$

Although the calculation of these weights is straightforward, the MLE method detailed above relies heavily on correctly specifying the linear treatment model. If the model is not correctly specified or the model assumptions are not met (i.e. normality assumptions), the MLE method can produce extreme weights that can lead to severely biased outcome estimates (Fong et al., 2018). Therefore, methods that operate outside of this MLE framework may produce better weights, leading to more covariate balance, and less biased estimates of the outcome.

1.1.7 Generalized Boosted Model

While Flores et al. estimates the GPS through generalized linear models (Flores et al., 2007), non-parametric methods of estimating the GPS vector, such as kernel methods, penalized spline models, and the generalized boosted model (GBM), can provide more accurate estimates of the GPS compared to parametric regression (Bia et al., 2014; Zhu et al., 2015). Instead of estimating a linear model for treatment on \mathbf{X} , like in a parametric setting, GBM fits a more general model of the form

$$T = m(\mathbf{X}) + \varepsilon \quad (1.11)$$

where $\varepsilon \sim N(0, \sigma^2)$ and $m(\mathbf{X})$ is the mean function of T given \mathbf{X} (Zhu et al., 2015). This mean function can be estimated using a machine learning algorithm, boosting, that additively fits regression trees until the model is sufficiently flexible to fit the data (McCaffrey et al., 2013; Zhu et al., 2015). Boosting automatically selects important covariates, nonlinear terms, and interaction terms to accurately estimate the mean function thus providing better estimates of the GPS (McCaffrey et al., 2013; Zhu et al., 2015). With the mean function derived, stabilized IPW can be calculated and implemented just like in the MLE weighting procedure.

While it seems that GBM provides better estimates of the GPS, thereby providing less biased outcome estimates, there are still drawbacks that limit this method. First, the method does not give users the ability to force variables into the final treatment model. The black box nature of GBM prevents the user from knowing how the model specifically utilizes each variable in producing the final propensity scores (Ridgeway et al., 2016). Additionally, although GBM has been shown to outperform MLE in simulation, covariate balance after GBM weighting can still remain poor leading to unstable estimates, worse than if no weights had been applied at all (Fong et al., 2018). Finally, while GBM does attempt to optimize balance, the only way to improve balance is by increasing the number of regression trees used by the method which may still not provide adequate control over sample imbalance (Fong et al., 2018).

1.1.8 Covariate Balancing Generalized Propensity Score

Recently, work has been done to extend the Covariate Balancing Propensity Score, which models treatment assignment while optimizing covariate balance, to the continuous treatment setting (Imai and Ratkovic, 2014; Fong et al., 2018). This new method, termed

Covariate Balancing Generalized Propensity Score (CBGPS), uses moment conditions to derive IPW such that the weighted correlation between \mathbf{X} and T is minimized (Fong et al., 2018). While the CBGPS method has a parametric approach that follows closely to the MLE method, it also has a nonparametric extension that places no parametric restrictions on the GPS nor on the marginal distribution of the treatment (Fong et al., 2018). This nonparametric Covariate Balancing Generalized Propensity Score (npCBGPS) gives researchers a method to directly derive weights without giving a functional form to the propensity scores (Fong et al., 2018).

Even though the CBGPS offers a method to optimize covariate balance while estimating the GPS, it is still not without limitations. Specifically, in simulation studies, it has been shown that GBM produces less biased outcome estimates compared to CBGPS and npCBGPS when sample sizes are large ($\sim 1,000$) (Fong et al., 2018). Additionally, since the nonparametric extension, npCBGPS, is based on an empirical likelihood approach, there is no guarantee that the optimization procedures find the global optimum (Fong et al., 2018). Furthermore, when the number of covariates is large or if \mathbf{X} strongly predicts T , the npCBGPS can fail to find a solution leaving the researcher to sacrifice covariate balance to derive weights (Fong et al., 2018).

Although there are many methods that seek to create balanced data for multiple and continuous treatments within observational studies, these methods are complex, possibly difficult to implement in practice, and do not have the same flexibility as a scalar balancing score used in typical binary propensity score methods. The goal of this work will be to develop and test new methodology to better conduct propensity scoring for more complicated treatment settings.

1.2 Public Health Significance

Randomized control trials are considered the gold standard when conducting research. The reason for this is due to the randomization that can be performed by researchers at the start of the study. Before the initiation of the study, researchers have the ability to balance covariates among the treatment groups by randomly assigning subjects to the groups. This randomization effectively creates treatment groups, with no systematic differences, that, on average, are identical in terms of their covariates. At the end of the study, when conducting the outcome analysis, any differences in the outcome can be attributed directly to the treatment, and not due to any covariate bias, as randomization successfully produces covariate balance between the treatment groups. Thus due to the randomization performed at the start of the study, causal inference in randomized control trials is made possible as there is covariate balance between the treatment groups. Although randomized experiments allow for direct causal inference between a treatment and an outcome, their utilization is not practical under many scenarios.

Instead, researchers can implement observation studies when randomized experiments are not feasible. That is, when the disease of interest is rare, subjects cannot be randomized to exposure groups due to ethical issues, the study will be too costly, etc. Thus, observation studies are highly utilized in public health analyses. While these study designs facilitate research that otherwise may not be feasible to conduct, investigators using observational study designs relinquish the ability to randomize subjects into treatment groups as their data is observational. Thus at the end of an observational study, there is no longer

covariate balance between the treatment groups. As a result, observational studies restrict an investigator's capacity to make causal inferences without utilizing complex statistical tools.

To create covariate balance between the treatment groups within observational studies, researchers often use propensity scoring. The underlying theory is that subjects with the same values for their propensity scores will have, on average, similar covariate profiles. Thus by using the propensity score in the outcome analysis, researchers are able to create covariate balance and produce causal estimates with observational studies.

Existing research in the area of propensity score methodology is limited outside binary treatment comparisons. Additionally, as data from large observational studies becomes more readily available, new propensity score methods are needed for these data types. Continuation of research into novel methods of propensity scoring is needed to ensure causal estimates can be made under all treatment scenarios and data types.

1.3 Specific Aims

1.3.1 To develop a novel method of propensity score analysis for multinomial treatments

As detailed above, current multinomial methods do not utilize the full GPS vector to match, stratify, or weight subjects. Although GBM, the most commonly used machine learning method for propensity scoring in multiple treatment scenarios, does have the ability to accurately derive the GPS vector, outcome analyses are limited by implementation of the GPS vector through IPW. Using a single value of the GPS may result in inaccurate outcome estimates. Therefore, developing a method that has the same flexibility as the scalar value used in binary treatment settings that can encapsulate the full GPS vector would be a useful addition to current propensity score literature. The proposed method can be used with all

types of propensity score models (parametric and non-parametric) and will adapt machine learning methods for the use in regular as well as complex observational data types, such that both matching and stratification are supported.

1.3.2 To develop a novel method of propensity score analysis for continuous treatments

In the continuous treatment setting, the GPS is not estimable through logistic regression models, as is common with binary and multiple treatments. Instead, methods for creating balanced data for the continuous treatment setting within observational studies have relied heavily on weighting procedures. Although methods have been proposed that operate outside of a parametric setting to derive weights, all weighting methods may produce unreliable outcome estimates due to extreme weights. Thus methods that do not utilize weighting nor rely on parametric assumptions may produce more reliable outcome estimates. Specifically, a method that can accurately stratify subjects based on a desired set of covariates would be a valuable tool for researchers.

1.3.3 To develop an R package to implement multiple treatments propensity scoring methods

The utility of new a propensity score method is directly related to the ease by which it can be used by researchers. Methods often require significant data manipulation, nuanced selection procedures, and complex algorithms to be used in practice. Thus having a standard R package that can be downloaded and easily adapted for various research projects will exponentially increase the notoriety of new propensity score methods and help facilitate more robust propensity score analyses by researchers.

References

- Austin, P. C. (2018a). Assessing covariate balance when using the generalized propensity score with quantitative or continuous exposures. *Statistical methods in medical research*, 0962280218756159.
- Austin, P. C. (2018b). Assessing the performance of the generalized propensity score for estimating the effect of quantitative or continuous exposures on binary outcomes. *Stat Med* **37**, 1874-1894.
- Austin, P. C., Grootendorst, P., and Anderson, G. M. (2007). A comparison of the ability of different propensity score models to balance measured variables between treated and untreated subjects: a Monte Carlo study. *Stat Med* **26**, 734-753.
- Austin, P. C., and Stuart, E. A. (2015). Moving towards best practice when using inverse probability of treatment weighting (IPTW) using the propensity score to estimate causal treatment effects in observational studies. *Stat Med* **34**, 3661-3679.
- Bia, M., Flores, C. A., Flores-Lagunes, A., and Mattei, A. (2014). A Stata package for the application of semiparametric estimators of dose–response functions. *Stata Journal* **14**, 580-604.
- Boyd, C. L., Epstein, L., and Martin, A. D. (2010). Untangling the causal effects of sex on judging. *American journal of political science* **54**, 389-411.
- Burgette, L., Griffin, B. A., and McCaffrey, D. (2017). Propensity scores for multiple treatments: A tutorial for the mnps function in the twang package. *R package. Rand Corporation*.

- Busso, M., DiNardo, J., and McCrary, J. (2014). New evidence on the finite sample properties of propensity score reweighting and matching estimators. *Review of Economics and Statistics* **96**, 885-897.
- Chertow, G. M., Normand, S.-L. T., and McNeil, B. J. (2004). “Renalism”: inappropriately low rates of coronary angiography in elderly individuals with renal insufficiency. *Journal of the American Society of Nephrology* **15**, 2462-2468.
- Davidson, M. B., Hix, J. K., Vidt, D. G., and Brotman, D. J. (2006). Association of impaired diurnal blood pressure variation with a subsequent decline in glomerular filtration rate. *Archives of internal medicine* **166**, 846-852.
- Donohue, J. J., and Ho, D. E. (2007). The Impact of Damage Caps on Malpractice Claims: Randomization Inference with Difference-in-Differences. *Journal of Empirical Legal Studies* **4**, 69-102.
- Feng, P., Zhou, X. H., Zou, Q. M., Fan, M. Y., and Li, X. S. (2012). Generalized propensity score for estimating the average treatment effect of multiple treatments. *Stat Med* **31**, 681-697.
- Flores-Lagunes, A., Gonzalez, A., and Neumann, T. (2007). Estimating the effects of length of exposure to a training program: the case of Job Corps.
- Fong, C., Hazlett, C., and Imai, K. (2018). Covariate balancing propensity score for a continuous treatment: Application to the efficacy of political advertisements. *The Annals of Applied Statistics* **12**, 156-177.
- Greene, T.J. (2017). Utilizing Propensity Score Methods for Ordinal Treatments and Prehospital Trauma Studies. *Texas Medical Center Dissertations (via ProQuest)*.

- Gu, X. S., and Rosenbaum, P. R. (1993). Comparison of multivariate matching methods: Structures, distances, and algorithms. *Journal of Computational and Graphical Statistics* **2**, 405-420.
- Harder, V. S., Stuart, E. A., and Anthony, J. C. (2008). Adolescent cannabis problems and young adult depression: male-female stratified propensity score analyses. *American journal of epidemiology* **168**, 592-601.
- Hirano, K., and Imbens, G. W. (2001). Estimation of causal effects using propensity score weighting: An application to data on right heart catheterization. *Health Services and Outcomes research methodology* **2**, 259-278.
- Huang, I.-C., Frangakis, C., Dominici, F., Diette, G. B., and Wu, A. W. (2005). Application of a propensity score approach for risk adjustment in profiling multiple physician groups on asthma care. *Health services research* **40**, 253-278.
- Imai, K., and Ratkovic, M. (2014). Covariate balancing propensity score. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **76**, 243-263.
- Imai, K., and Van Dyk, D. A. (2004). Causal inference with general treatment regimes: Generalizing the propensity score. *Journal of the American Statistical Association* **99**, 854-866.
- Imbens, G. W. (2000). The role of the propensity score in estimating dose-response functions. *Biometrika* **87**, 706-710.
- Joffe, M. M., and Rosenbaum, P. R. (1999). Invited commentary: propensity scores. *American journal of epidemiology* **150**, 327-333.

- Lechner, M. (2001). Identification and estimation of causal effects of multiple treatments under the conditional independence assumption. In *Econometric evaluation of labour market policies*, 43-58: Springer.
- Lechner, M. (2002). Program heterogeneity and propensity score matching: An application to the evaluation of active labor market policies. *Review of Economics and Statistics* **84**, 205-220.
- Li, F., Morgan, K. L., and Zaslavsky, A. M. (2018). Balancing covariates via propensity score weighting. *Journal of the American Statistical Association* **113**, 390-400.
- Lopez, M. J., and Gutman, R. (2017). Estimation of causal effects with multiple treatments: a review and new ideas. *Statistical Science* **32**, 432-454.
- McCaffrey, D. F., Griffin, B. A., Almirall, D., Slaughter, M. E., Ramchand, R., and Burgette, L. F. (2013). A tutorial on propensity score estimation for multiple treatments using generalized boosted models. *Stat Med* **32**, 3388-3414.
- McCaffrey, D. F., Ridgeway, G., and Morral, A. R. (2004). Propensity score estimation with boosted regression for evaluating causal effects in observational studies. *Psychological methods* **9**, 403.
- Nielsen, R. A., Findley, M. G., Davis, Z. S., Candland, T., and Nielson, D. L. (2011). Foreign aid shocks as a cause of violent armed conflict. *American journal of political science* **55**, 219-232.
- Rassen, J. A., Solomon, D. H., Glynn, R. J., and Schneeweiss, S. (2011). Simultaneously assessing intended and unintended treatment effects of multiple treatment options: a pragmatic “matrix design”. *Pharmacoepidemiology and drug safety* **20**, 675-683.

- Ridgeway, G., McCaffrey, D., Morral, A., Griffin, B., and Burgette, L. (2016). Toolkit for Weighting and Analysis of Nonequivalent Groups (Version 1.4-9.5).
- Robins, J. M., Hernan, M. A., and Brumback, B. (2000). Marginal structural models and causal inference in epidemiology. LWW.
- Rosenbaum, P. R., and Rubin, D. B. (1983). The Central Role of the Propensity Score in Observational Studies for Causal Effects. *Biometrika* **70**, 41-55.
- Rosenbaum, P. R., and Rubin, D. B. (1984). Reducing bias in observational studies using subclassification on the propensity score. *Journal of the American Statistical Association* **79**, 516-524.
- Rosenbaum, P. R., and Rubin, D. B. (1985). Constructing a control group using multivariate matched sampling methods that incorporate the propensity score. *The American Statistician* **39**, 33-38.
- Royston, P., Altman, D. G., and Sauerbrei, W. (2006). Dichotomizing continuous predictors in multiple regression: a bad idea. *Stat Med* **25**, 127-141.
- Rubin, D. B. (1979). Using multivariate matched sampling and regression adjustment to control bias in observational studies. *Journal of the American Statistical Association* **74**, 318-328.
- Rubin, D. B. (2004). On principles for modeling propensity scores in medical research. *Pharmacoepidemiology and drug safety* **13**, 855-857.
- Schuler, M. S., Chu, W., and Coffman, D. (2016). Propensity score weighting for a continuous exposure with multilevel data. *Health Services and Outcomes research methodology* **16**, 271-292.

- Setoguchi, S., Schneeweiss, S., Brookhart, M. A., Glynn, R. J., and Cook, E. F. (2008). Evaluating uses of data mining techniques in propensity score estimation: a simulation study. *Pharmacoepidemiology and drug safety* **17**, 546-555.
- Seya, H., and Yoshida, T. (2017). Propensity score matching for multiple treatment levels: A CODA-based contribution. *arXiv preprint arXiv:1710.08558*.
- Stuart, E. A. (2010). Matching methods for causal inference: A review and a look forward. *Statistical science: a review journal of the Institute of Mathematical Statistics* **25**, 1.
- Tu, C., Jiao, S., and Koh, W. Y. (2013). Comparison of clustering algorithms on generalized propensity score in observational studies: a simulation study. *Journal of Statistical Computation and Simulation* **83**, 2206-2218.
- Yang, S., Imbens, G. W., Cui, Z., Faries, D. E., and Kadziola, Z. (2016). Propensity score matching and subclassification in observational studies with multi-level treatments. *Biometrics* **72**, 1055-1065.
- Zanutto, E., Lu, B., and Hornik, R. (2005). Using propensity score subclassification for multiple treatment doses to evaluate a national antidrug media campaign. *Journal of Educational and Behavioral Statistics* **30**, 59-73.
- Zhao, Z. (2004). Using matching to estimate treatment effects: Data requirements, matching metrics, and Monte Carlo evidence. *Review of Economics and Statistics* **86**, 91-107.
- Zhu, Y., Coffman, D. L., and Ghosh, D. (2015). A boosting algorithm for estimating generalized propensity scores with continuous treatments. *Journal of causal inference* **3**, 25-40.

CHAPTER II

A Novel Approach for Propensity Score Matching and Stratification in the Presence of Multiple Treatments: Application to an EHR-Derived Study of Subarachnoid Hemorrhage

Proposed Journal: Computational Statistics and Data Analysis

Derek W. Brown^a, Stacia M. DeSantis^a, Thomas J. Greene^b, Vahed Marouf^a, Hulin Wu^{a,c},
George Williams^d, Michael D. Swartz^a

Author Affiliations

- a) Department of Biostatistics and Data Science, University of Texas Health School of Public Health, 1200 Pressler St, Houston, TX 770303, USA
- b) GlaxoSmithKline, Division of Biostatistics, 5 Crescent Dr, Philadelphia, PA 19112, USA
- c) University of Texas School of Biomedical Informatics, 7000 Fannin St Suite 600, Houston, TX 77030, USA
- d) McGovern Medical School at UTHealth, 6431 Fannin St, Houston, TX 77030, USA

Corresponding Author

Derek Brown, PhD(c), MS
Department of Biostatistics and Data Science
University of Texas Health Science Center at Houston
1200 Pressler Street, Houston, TX 77030
[derek.brown@uth.tmc.edu]

Abstract

The generalized propensity score (GPS), a vector whose elements represent the probabilities a subject was assigned each treatment, is used to extend work in binary treatment propensity scoring to the multiple treatment setting. Currently, methods for conducting multiple treatment propensity scoring in the presence of high-dimensional covariate spaces that result from ‘big data’ are lacking – the most prominent method relies on inverse probability treatment weighting (IPTW). However, IPTW only utilizes one element of the GPS vector and can lead to a loss of information and inadequate covariate balance in the presence of multiple treatment groups. The above limitations motivate the development of a novel propensity score method that uses the entire GPS vector to establish a scalar balancing score that when adjusted for, achieves covariate balance in the presence of potentially high-dimensional covariates. Specifically, the generalized propensity score cumulative distribution function (GPS-CDF) method is introduced. A one-parameter power function fits the CDF of the GPS vector and a resulting scalar balancing score is used for matching and/or stratification. Simulation results show superior performance of the new method compared to IPTW both in achieving covariate balance and estimating average treatment effects in the presence of multiple treatments. The proposed approach is applied to a study derived from electronic medical records to determine the causal relationship between three different vasopressors and mortality in patients with non-traumatic aneurysmal subarachnoid hemorrhage. Our results suggest that the GPS-CDF method performs well when applied to large observational studies with multiple treatments that have large covariate spaces.

Keywords: Causal Inference, Multinomial Treatments, Observational Study, Propensity Score

2.1 Introduction

Propensity scoring is utilized to overcome the covariate imbalance prevalent in observational studies, enabling causal estimates of treatment outcome relationships, when propensity scoring assumptions are met. Increasingly, large data sources including national surveys, electronic health records (EHRs), and genome wide association studies (GWAS) with phenotypic and covariate data are becoming publicly available. These observational data sources are indexed by large covariate spaces for example, patient demographics, vital signs, laboratory findings, medications/prescriptions, comorbidities, etc. (Patorno et al., 2014; Chen and Moskowitz, 2016). It is of interest to use these potential pretreatment confounders in a propensity model to ultimately assess the causal relationship between treatments and an outcome. While these data types have gained rapid traction in the literature, propensity scoring methods for assessing the effects of multiple (non-binary) treatments in the presence of high-dimensional covariate spaces that result from these data sources are lacking (Schneeweiss et al., 2009; Schuemie et al., 2012; Stuart et al., 2013; Low, Gallego and Shah, 2016; Ju et al., 2019).

Methods for utilizing propensity scoring in a binary treatment case are well studied and established (e.g., Rosenbaum and Rubin 1983, 1984, 1985; Stuart, 2010; Gutman and Rubin, 2015). However, generalizations to multiple unordered (multinomial) treatments are more complicated. The generalized propensity score (GPS) is often used to extend the theory of causal inference from a binary treatment setting to a multiple treatment setting (Joffe and Rosenbaum, 1999; Imbens, 2000; Imai and Van Dyk, 2004). The GPS is defined as the probability of receiving one of K treatments conditional on a given set of observed covariates (Imbens, 2000). Unlike the binary treatment case where the propensity score is a single value representing the probability a subject was treated, the GPS is a vector, of length K ,

representing the conditional probabilities of a subject being treated under each of the K treatments.

An important distinction for causal inference using propensity scoring for multinomial treatments are the two different summary measures of the treatment effect: the average treatment effect (ATE) and the average treatment effect among the treated (ATT). ATE is of interest for comparisons of the mean outcome when the entire population is eligible for all treatments (McCaffrey et al., 2013). ATE is calculated by taking the expectation across the *entire* population and is given by:

$$ATE_{k,k'} = E[Y(k) - Y(k')] = E[Y(k)] - E[Y(k')] \quad (2.1)$$

where Y is the outcome for the comparison of treatment k and treatment k' . ATT is of interest when comparing the effectiveness of a particular treatment relative to the alternatives available to the population of interest (McCaffrey et al., 2013). ATT finds the effect of the treatment of interest among only those subjects *who actually received* the treatment. ATT is formally defined as:

$$ATT_{k,k'} = E[Y(k') | Z = k] - E[Y(k) | Z = k] \quad (2.2)$$

where Y is the outcome and Z is the treatment of interest.

Traditionally, methods for conducting propensity scoring in the presence of multinomial treatments have relied on the GPS vector that is produced from some type of multinomial regression model, e.g.,

$$\log \left[\frac{Pr(Z_i = k)}{Pr(Z_i = K)} \right] = \theta_k + x_i' \beta_k \quad (2.3)$$

where θ_k is a constant, β_k is a vector of regression coefficients, Z is the treatment received, and K is the total number of treatments, for $k = \{1, 2, \dots, K - 1\}$. Commonly used methods of conducting multinomial propensity scoring based on a parametrically derived GPS can be classified into distance metrics (Seya and Yoshida, 2017; Rassen et al., 2013; Rubin, 1979; Zhao, 2004), clustering techniques (Tu, Jiao and Koh, 2013; Lopez and Gutman, 2017), and stratification, matching, and adjustment methods (Zanutto, Lu and Hornik, 2005; Huang et al., 2005; Yang et al., 2016; Lechner, 2001; Feng et al., 2012).

Although many methods have been proposed to conduct multinomial propensity scoring, there is no unified method, and current methods have drawbacks that diminish their utility, especially in the context of big data. For example, as most of the aforementioned methods exclusively estimate either ATE or ATT, their practical utility is limited.

Additionally, the distance based matching approach proposed by Rassen et. al cannot be extended past three treatments (Rassen et al., 2013). Likewise, matching based on Mahalanobi's distance (Rubin, 1979; Zhao, 2004), does not perform well with more than 8 covariates or when covariates are not normally distributed (for example if they are non-continuous (Gu and Rosenbaum, 1993; Stuart, 2010)). These are major limitations for big data applications, which as previously stated, often have multiple treatment groups and a large number of pretreatment confounders. Furthermore, although methods that produce covariate balance using stratification, matching, and adjustment based on the GPS vector have been studied in the multiple treatment setting (Zanutto, Lu and Hornik, 2005; Huang et al., 2005; Yang et al., 2016; Lechner, 2001; Feng et al., 2012), a function that maps the GPS vector to a *scalar* balancing score has not, to our knowledge, been proposed. A cornerstone

of these current approaches is the estimation of treatment effects that are performed separately for each treatment level. This simplifies the inherent problems that arise when assessing balance among multiple treatment groups. However, since these approaches utilize one element of the GPS vector instead of the full GPS vector (Greene, 2017), in some cases they may suffer from a loss of information, resulting in suboptimal covariate balance.

To address the aforementioned issues without placing any parametric restrictions on the relationship between treatment groups and pretreatment confounders, non-parametric machine learning methods of estimating the GPS vector, such as generalized boosted models (GBM), recursive partitioning, neural nets, and super learners have been proposed (McCaffrey, Ridgeway and Morral, 2004; Setoguchi et al., 2008; McCaffrey et al., 2013; Burgette, Griffin and McCaffrey, 2017; Ju et al., 2019). The most popular method, due to the availability of a comprehensive R package, appears to be GBM (Burgette, Griffin and McCaffrey, 2017) – this and other tree-based methods provide notable benefits over parametric regression. For example, variable selection including the decision to accommodate higher order or interaction terms in the model occurs automatically. This is of particular importance when working with EHRs since there are a large number of potential confounders available (McCaffrey et al., 2013). Further, the iterative estimation procedure used by GBM, which fits regression trees that maximize the log likelihood in order to produce a piecewise constant model, can easily be refined to provide the propensity score model with the best balance between treatment groups (McCaffrey et al., 2013). After estimating the GPS vector, inverse probability treatment weighting (IPTW), where the weight is the inverse propensity of the treatment an individual actually received (Imbens, 2000; Feng et al., 2012; McCaffrey et al., 2013; Burgette, Griffin and McCaffrey, 2017), may be either applied directly to the outcome (Feng et al., 2012) or utilized within weighted

regression models (McCaffrey, 2013; Burgette, Griffin and McCaffrey, 2017) to estimate ATEs. Additionally, weights derived from multiplying the inverse probability weight by the probability of the target treatment may be used to estimate ATTs (McCaffrey et al., 2013; Burgette, Griffin and McCaffrey, 2017).

One issue with utilizing machine learning methods to estimate the GPS vector is that they are only compatible with IPTW. Matching and stratification cannot be utilized in the outcome model since no obvious scalar balancing score is produced. Although this simplifies many of the inherent issues that arise with multiple treatments, IPTW may produce unreliable outcome estimates, with large sample variances, due to extreme weights (Busso, DiNardo, McCrary, 2014; Lopez and Gutman, 2017; Li, Morgan, Zaslavsky, 2018). Alternative weighting methods have been proposed that are less susceptible to these extremes (Hirano and Imbens, 2001; Imai and Ratkovic, 2014; Li, Morgan, Zaslavsky, 2018). However, since weighting directly uses the scalar estimated propensity score in determining the effect of treatment (Austin, Grootendorst and Anderson, 2007; Rubin, 2004), as Rubin (Rubin, 2004) suggests, this results in the greatest sensitivity to misspecification of the propensity score. Furthermore, the utility of IPTW may diminish as the number of treatment groups increases past $K = 3$; Yang et al. (2016) showed that IPTW performs well in the presence of three treatments but has inferior performance with six treatments. These limitations have precluded the use of machine learning methods in propensity scoring over the last decade, despite their promise.

In sum, while multinomial propensity score methods exist, they are somewhat ad hoc, potentially difficult to implement, and do not have the flexibility of the scalar balancing score obtained from binary propensity scoring. To address these limitations, this paper presents a novel approach, the generalized propensity score cumulative distribution function (GPS-

CDF), which maps a GPS vector to a single scalar value that can be used for propensity score matching and stratification in order to produce causal inferences with multinomial treatments. The methodology is a natural extension of the binary setting. It is tested via simulation and applied to an EHR-derived study to evaluate the effect of vasopressor choice on mortality in patients with non-traumatic aneurysmal subarachnoid hemorrhage. The proposed GPS-CDF methodology is publicly available through the *GPSCDF* R package (Brown et al., 2019).

2.2 Methods

Since the GPS vector represents the probability a subject received each of the K treatments conditional on a given set of observed covariates (Imbens, 2000), the GPS vector can be thought of as a discrete probability distribution that can be used to create a probability mass function (PMF) (Greene, 2017). In this way, subjects with similar shapes for their PMFs will have similar values for their GPS vectors, which in turn means, on average, they will have similar covariate distributions. A single scalar parameter function that can accurately describe the shape of the PMF could be used as a balancing score to easily match or stratify subjects.

The PMF is not a monotonically increasing or decreasing function. The shape of the PMF will vary for each subject depending on their treatment probabilities; therefore, estimating a one parameter function that describes the shape of the PMF is difficult. Instead, a cumulative distribution function (CDF) can be created for each subject by summing across values of the GPS vector. By definition, the CDF is a strictly increasing function bounded by zero and one, so fitting a one parameter function to the shape of the CDF is possible. As the CDF is a 1-to-1 function of the GPS vector, subjects with similar values for a one parameter

function that maps the shape of the CDF will have similar GPS vectors. As such, a one parameter power function can be used to model the CDF. The equation of this proposed power function that maps the CDF of the GPS vector is given by:

$$P(Z_i \leq d) = F_k(Z) \approx f(\tilde{a}) = d_k^{\exp(\tilde{a})} \text{ for } k = 1, \dots, K - 1 \quad (2.4)$$

where the left side represents the CDF for the GPS vector, d_k is a standardized treatment dose which lies between 0 and 1, \tilde{a} is the scalar that dictates the shape of the power function fitting the CDF, and k is the indicator of the treatment group. In a three treatment setting, for example, the standardized dose d_k is taken to be 0.33, 0.66 for the first two treatment groups and equal to 1 for the final treatment group. The chosen power function in equation (2.4) allows both convex and concave CDFs to be accurately modeled (Storer, 1989; O’Quigley, Pepe and Fisher, 1990) as shown by Figure 2.1. Other one parameter functions (e.g. exponential, logarithmic, sigmoid) might initially seem obvious to apply but do not have this advantage (O’Quigley, Pepe and Fisher, 1990). Once the CDF has been calculated for each subject, a non-linear least squares (NLS) algorithm (Marquardt, 1963) is used to fit the power function. This NLS algorithm iteratively fits values for \tilde{a} , the shape parameter, until the residual distance between the CDF and fitted power function is minimized. Formally, NLS estimation is given by:

$$\min_{\tilde{a}} \sum_{k=1}^{K-1} \left(d_k^{\exp(\tilde{a})} - F_k(Z) \right)^2 \text{ for } k = 1, \dots, K - 1. \quad (2.5)$$

An important feature of the proposed method is its compatibility with any parametric or machine learning model that produces a GPS vector. Currently, there are no methods for multinomial treatment propensity scoring that are both “propensity model free” and produce

a scalar balancing score that facilitates matching and stratification. While it may initially appear that other methods (e.g. multivariate distances, Kolmogorov-Smirnov test statistic, isotonic regression, Kullback–Leibler divergence, etc.) could be used to differentiate CDFs derived from the GPS vector, these methods do not result in a scalar balancing score. Although these alternative options may be used to match subjects, they do not allow for adjustment through stratification. The proposed approach in equation (2.4) will both accurately describe the curvature of the CDF and also produce a single scalar balancing score that can easily be used for both matching and stratification of subjects.

Unlike ordinal treatment settings where there is a natural ordering to the treatments, multinomial treatments can be aligned in any order within the GPS vector. In a setting with three multinomial treatments (A, B, and C), for example, the GPS vector can be ordered as A-B-C, A-C-B, B-A-C, B-C-A, C-A-B, and C-B-A. Each ordering is intuitive and will produce a different CDF, and subsequently a new shape parameter, $\tilde{\alpha}$, for each subject. Since a balancing score is just a function of covariates such that the conditional distribution of the covariates given the balancing score is the same for all treatment groups (Rosenbaum and Rubin, 1983), for three multinomial treatments, there are 6 different balancing scores produced by the proposed GPS-CDF method. By rearranging the GPS vector for all possible orderings, $K!$ balancing scores can be created in a K treatment setting. As with standard propensity score methods, covariate balance can be assessed after matching or stratification based on each of the $K!$ orderings of the GPS vector to choose the ordering and method that creates the best covariate balance in the data. While this may at first appear ad hoc, it has substantial precedent in the literature - several recently proposed propensity score methods (e.g., Fong, Hazlett and Imai, 2018; Imai and Ratkovic, 2014; Papadogeorgou, Choirat and Zigler, 2018) seek to optimize covariate balance before implementing propensity scoring in

the outcome analysis. For example, the spatial propensity score method proposed by Papadogeorgou et al. (2018) incorporates an automated data-driven process of selecting matched pairs over a possible range of weights, and further selects the weight that achieves the best covariate balance. In this way, the proposed iterative nature of finding the \tilde{a} that optimizes covariate balance is analogous to the aforementioned method. Covariate balance should be assessed using each resultant $K!$ balancing score produced by the GPS-CDF method, and the ordering that achieves the best covariate balance, among all subjects, should be retained for the outcome analysis.

2.2.1 GPS-CDF Matching

As the estimated power parameter, \tilde{a} , is a scalar value, it can be used in either greedy or optimal matching algorithms to pair subjects, with similar \tilde{a} values, who received different treatments. The proposed metric for matching is the absolute difference between the power parameters for two subjects, \tilde{a}_i and \tilde{a}_j who received different treatments. Minimizing this difference will jointly pair subjects with similar values of \tilde{a} while ensuring the subjects received different treatments. This metric for two subjects, i and j , is given by equation (2.6).

$$\Delta_p = \Delta(\mathbf{x}_i, \mathbf{x}_j) = |\tilde{a}_i - \tilde{a}_j| \quad (2.6)$$

After the matching procedure is performed for each of the $K!$ orderings of the GPS vector, the order that creates matches with the best covariate balance should be retained for the outcome analysis. As is standard with propensity score methods, we propose selecting the ordering that minimizes the standardized mean difference (SMD) within matches (Austin, 2011; Burgette, Griffin and McCaffrey, 2017; Fong, Hazlett and Imai, 2018; Imai and

Ratkovic, 2014; Lopez and Gutman, 2017; McCaffrey et al., 2013; Papadogeorgou, Choirat and Zigler, 2018; Yang et al., 2016). Formally, this selection can be written as:

$$\operatorname{argmin}_{\tilde{a}_k} \left[\sum_{p=1}^P \sum_{m=1}^{M_{\tilde{a}_k}} \sum_{t_i \neq t_j} \left(\frac{\bar{x}_{p,m,t_i} - \bar{x}_{p,m,t_j}}{\sqrt{\frac{s_{p,m,t_i}^2 + s_{p,m,t_j}^2}{2}}} \right) \right] \quad (2.7)$$

where \tilde{a}_k is the \tilde{a} derived from each $K!$ ordering of the GPS vector, P is the number of covariates, $M_{\tilde{a}_k}$ is the index for the number of matched pair treatment groups created from each ordering of the GPS vector, and $t_i = 1, 2, \dots, K$ represents the multinomial treatment groups. The following steps detail the matching procedure:

1. Choose variables related to the treatments to include in the propensity model (see Brookhart et al. (2006) for a detailed discussion of variable selection for propensity models).
2. Estimate the GPS vector for each subject using any desired parametric or machine learning model (i.e. multinomial logistic regression, GBM, etc.).
3. Choose an ordering of the K treatments within the GPS vector.
4. Calculate the CDF of the ordered GPS vector for each subject.
5. Fit a one parameter power function to the CDF of each subject to obtain \tilde{a} .
6. Calculate the $\Delta_{(i,j)}$ matrix between all pairs of subjects.
7. Create matched pairs using a matching algorithm.
8. Repeat steps 3-7 for each additional $(K! - 1)$ ordering of the GPS vector.
9. Assess covariate balance after matching separately for each of the $K!$ balancing scores via SMD.
10. Retain the balancing score that creates matches with the best covariate balance.
11. Conduct a matched outcome analysis to estimate ATE or ATT (e.g. conditional logistic regression).

2.2.2 GPS-CDF Stratification

The method for stratification follows closely to the method proposed for matching. Using the estimated power parameter, \tilde{a} , strata can be created to group subjects with similar values of \tilde{a} who received different treatments. Thus within strata, subjects have similar covariate distributions and received different treatments. Although any number of strata can be created, it has been shown in previous studies that stratifying the data into quintiles removes approximately 90% of the initial observed covariate imbalance (Cochran, 1968; Rosenbaum and Rubin, 1984; Zanutto, Lu and Hornik, 2005; Austin, 2011).

As with the matching procedure, stratification can be performed for each of the $K!$ orderings of the GPS vector. Again, the ordering that creates the strata with the best covariate balance should be retained for the outcome analysis. We propose selecting the ordering that minimizes the SMD within strata (Austin, 2011; Burgette, Griffin and McCaffrey, 2017; Fong, Hazlett and Imai, 2018; Imai and Ratkovic, 2014; Lopez and Gutman, 2017; McCaffrey et al., 2013; Papadogeorgou, Choirat and Zigler, 2018; Yang et al., 2016). Formally, this selection can be written as:

$$\underset{\tilde{a}_k}{\operatorname{argmin}} \left[\sum_{p=1}^P \sum_{s=1}^{S_{\tilde{a}_k}} \sum_{t_i \neq t_j} \left(\frac{\bar{x}_{p,s,t_i} - \bar{x}_{p,s,t_j}}{\sqrt{\frac{S_{p,s,t_i}^2 + S_{p,s,t_j}^2}{2}}} \right) \right] \quad (2.8)$$

where \tilde{a}_k is the \tilde{a} derived from each $K!$ ordering of the GPS vector, P is the number of covariates, $S_{\tilde{a}_k}$ is the number of strata created for each ordering of the GPS vector, and $t_i = 1, 2, \dots, K$ represents the multinomial treatment groups. The following steps detail the stratification procedure:

1. Repeat steps 1 and 2 from the CDF matching procedure.
2. Choose an ordering of the K treatments within the GPS vector.
3. Calculate the CDF of the ordered GPS vector for each subject.
4. Fit a one parameter power function to the CDF of each subject to obtain \tilde{a} .
5. Rank observations based on their value for the power parameter \tilde{a} and separate the data into quintiles.
6. Repeat steps 3-5 for each additional $(K! - 1)$ ordering of the GPS vector.
7. Assess covariate balance after stratification separately for each of the $K!$ orderings via SMD.
8. Retain the GPS vector ordering that creates strata with the best covariate balance.
9. Conduct a stratified outcome analysis to estimate ATE or ATT (e.g. conditional logistic regression).

2.3 Simulation Study

A simulation study is conducted to determine how the GPS-CDF matching and stratification methods perform under different data scenarios with varying degrees of model misspecification. The design of the current simulation follows closely to several recently published simulations that seek to be representative of real data (Austin, Grootendorst, and Anderson, 2007; Fong, Hazlett and Imai, 2018; Greene, 2017). Four data scenarios are considered with three treatment categories, one binary outcome, and nine baseline covariates. Six covariates are associated with treatment assignment probability, and six covariates are associated with outcome assignment probability, producing various levels of treatment and outcome confounding. A table describing the associations of the baseline covariates with the treatment and outcome variables is shown in Table 2.1. From the table, it can be observed that x_1, x_2, x_4 , and x_5 are generated to be pretreatment confounders.

The four data scenarios considered within this simulation study are similar to those of Fong et al. (Fong, Hazlett and Imai, 2018) and Greene (2017) and vary whether treatment

and outcome assignment models are correctly specified. Incorrect specification is created through inclusion of a non-linear term. The nine baseline covariates are multivariate normally distributed with mean 0, variance 1, and covariances of 0.2.

Scenario 1 assumes both the treatment and outcome models are correct through inclusion of only linear terms. The true treatment and outcome models are given by equations (2.9) and (2.10), respectively.

$$\log \left[\frac{Pr(Z_i = k)}{Pr(Z_i = 3)} \right] = \theta_k + \beta_{1,k}x_{i,1} + \beta_{2,k}x_{i,2} + \beta_{4,k}x_{i,4} + \beta_{5,k}x_{i,5} + \beta_7x_{i,7} + \beta_8x_{i,8} \quad (2.9)$$

$$\text{for } k = 1, 2, \theta = (0.25, 0.3), \beta_1 = \beta_4 = (0.7, 0.4), \\ \beta_2 = \beta_5 = (0.2, 0.3), \beta_7 = 0.6, \text{ and } \beta_8 = 0.2$$

$$\log \left[\frac{Pr(Y_i = 1)}{1 - Pr(Y_i = 1)} \right] = \alpha + \beta_Z Z_i + \beta_1 x_{i,1} + \beta_2 x_{i,2} + \beta_3 x_{i,3} + \beta_4 x_{i,4} + \beta_5 x_{i,5} + \beta_6 x_{i,6} \quad (2.10)$$

$$\text{for } \alpha = -0.2, \beta_Z = (-0.1, 0.6, 0.3), \beta_1 = \beta_2 = \beta_3 = 0.6, \\ \text{and } \beta_4 = \beta_5 = \beta_6 = 0.4$$

The three level multinomial treatment and binary outcome variables are simulated by sampling one value from a multinomial distribution and Bernoulli distribution using the probabilities calculated from equation (2.9) and (2.10), respectively, as the probability sampling weights. From the data generation procedure, the true treatment effects are 0.7, 0.4, and -0.3 for treatment pairs (1, 2), (1, 3), and (2, 3), respectively.

Scenario 2 introduces a non-linear term based on a mis-measured variable, $(x_{i,1} + 0.5)^2$, into the treatment assignment model, while the outcome model remains the same as equation (2.10). The misspecified treatment model is given by:

$$\log \left[\frac{Pr(Z_i = k)}{Pr(Z_i = 3)} \right] = \theta_k + \beta_{1,k}(x_{i,1} + 0.5)^2 + \beta_{1,k}x_{i,1} + \beta_{2,k}x_{i,2} + \beta_{4,k}x_{i,4} + \beta_{5,k}x_{i,5} + \beta_7x_{i,7} + \beta_8x_{i,8} \quad (2.11)$$

$$\begin{aligned} \text{for } k = 1, 2, \theta = (-0.5, 0), \beta_1 = \beta_4 = (0.7, 0.4), \\ \beta_2 = \beta_5 = (0.2, 0.3), \beta_7 = 0.6, \text{ and } \beta_8 = 0.2. \end{aligned}$$

Scenario 3 introduces a non-linear term based on a mis-measured variable, $(x_{i,1} + 0.5)^2$, into the outcome assignment model, while the treatment model remains the same as equation (2.9). The misspecified outcome model is given by:

$$\log \left[\frac{Pr(Y_i = 1)}{1 - Pr(Y_i = 1)} \right] = \alpha + \beta_Z Z_i + 0.5(x_{i,1} + 0.5)^2 + \beta_1x_{i,1} + \beta_2x_{i,2} + \beta_3x_{i,3} + \beta_4x_{i,4} + \beta_5x_{i,5} + \beta_6x_{i,6} \quad (2.12)$$

$$\begin{aligned} \text{for } \alpha = -0.8, \beta_Z = (-0.1, 0.6, 0.3), \beta_1 = \beta_2 = \beta_3 = 0.6, \\ \text{and } \beta_4 = \beta_5 = \beta_6 = 0.4. \end{aligned}$$

Finally in Scenario 4, both the treatment and outcome models are misspecified using the treatment and outcome assignment models detailed in equations (2.11) and (2.12).

2.4 Results

For each data scenario considered, 1000 datasets each containing 1000 observations are generated. Five analytic tools are used to estimate and compare ATEs: unadjusted (crude odds ratio) model, adjusted (adjusted odds ratio) model, GBM with IPTW, GPS-CDF greedy matching, and GPS-CDF stratification. The GBM propensity model adjusts for all nine baseline covariates. Additionally, the GPS vector generated through GBM is used for GPS-CDF matching and stratification, and a caliper of 0.25 standard deviations of \tilde{a} is used for GPS-CDF greedy matching (Cochran and Rubin, 1973; Lunt, 2014). Outcome models to obtain ATE estimates utilize logistic regression for the unadjusted and adjusted models,

survey-weighted generalized linear models for GBM weighting, and conditional logistic regression for GPS-CDF matching and stratification. Furthermore, outcome models (with the exception of the unadjusted model) adjusted for all first order covariates associated with outcome assignment (Rosenbaum and Rubin, 1984; Hirano and Imbens, 2001; Imai and Van Dyk, 2004).

Figure 2.2 is a graphical depiction of the amount of covariate balance achieved by each analytical tool under both the correctly and incorrectly specified treatment model. The adjusted model was not included in this balance assessment plot as it has the same covariate balance as the unadjusted model. The plot depicts the maximum pairwise SMD for each of the nine baseline covariates within each of the simulated datasets (Lopez and Gutman, 2017). For each treatment pair, the SMD is calculated and the maximum value across treatment pairs is retained. Methods that achieve covariate balance have smaller maximum SMD values.

The three propensity based methods achieve better covariate balance, on average, compared to the original unweighted data. Within the correctly specified treatment model (left plot), GBM weighting produces better balance than both GPS-CDF matching and stratification. GPS-CDF matching and stratification produce similar balance in the correctly specified treatment model, but it appears that GPS-CDF stratification is less prone to outliers. Within the incorrectly specified treatment model (right plot), GBM weighting and GPS-CDF matching produce similar balance results. GPS-CDF stratification produces slightly worse balance than GBM weighting and GPS-CDF matching, but still produces better balance than the original data.

For each of the data scenarios considered, the five analysis methods are compared using average bias, mean squared error (MSE), and coverage probability of the estimated

ATEs. As there are three treatment groups, the performance of each method is assessed for each of the three treatment group comparisons. Figure 2.3 shows the distribution of the ATE estimates from each analytical method under each data scenario between treatment 1 and treatment 3. The true pairwise treatment effect of 0.4 is included as the dotted horizontal line.

Within Scenario 1, all methods produce estimates with minimal bias and high coverage probabilities, with the exception of the unadjusted model. The adjusted model, which does not include any propensity scoring, actually has the smallest MSE compared to the methods that include propensity models. This result was anticipated as there is no misspecification in either the treatment or outcome model under this data scenario. Additionally, GBM weighting performs better in terms of bias and MSE compared to GPS-CDF matching. However, even though GBM weighting produces better balance than GPS-CDF stratification, as indicated in Figure 2.2, GPS-CDF stratification has lower bias and higher coverage probability compared to GBM weighting. In Scenario 2, where the treatment model is misspecified but the outcome model is correct (Figure 2.3, Scenario 2), results are consistent with Scenario 1. The adjusted model performs better than GBM in terms of bias, MSE, and coverage probability. Again, GPS-CDF stratification produces lower bias than all other methods and obtained the highest coverage probability.

For Scenario 3, which includes a correctly specified treatment model but misspecified outcome model (Figure 2.3, Scenario 3), GBM weighting has lower bias but higher MSE compared to the adjusted model. GPS-CDF matching and stratification have lower bias compared to all other methods, with both GPS-CDF methods outperforming GBM weighting in terms of coverage probability. Finally in Scenario 4 (Figure 2.3 Scenario 4), the adjusted model, GBM weighting, and GPS-CDF matching all fail to obtain accurate ATE estimates. GPS-CDF stratification is still able to produce minimally biased ATE estimates while

maintaining low MSE and coverage probability close to 0.9 when both the treatment and outcome models are misspecified.

The ATE estimate results from the comparisons between treatments 1 and 2 (Appendix A: Supplemental Figure 1) and between treatments 2 and 3 (Appendix A: Supplemental Figure 2) are consistent with those detailed above. GPS-CDF stratification produces ATE estimates with minimal bias and MSE, while maintaining high coverage probability across all data scenarios.

Finally, Supplemental Figure 3 (Appendix A) is a graphical representation of the CDF mapping produced by the GPS-CDF method under 4 different multinomial treatment group scenarios: 3, 4, 6, and 10 treatments. For each multinomial treatment group, 1000 subjects are simulated in a manner similar to the above simulation study, and GBM is used to estimate the GPS vector for each simulated subject. The subsequent CDF vector for each subject is found by summing across the subject specific GPS vectors. The GPS-CDF method is then applied to derive subject specific \tilde{a} values. Each panel of Supplemental Figure 3 (Appendix A) shows the resultant CDF vector and estimated power function for 5 simulated subjects across each treatment group scenario. Supplemental Figure 3 (Appendix A) indicates that as the number of treatment groups increase, the power function, based on \tilde{a} , still accurately maps the CDF of the GPS vector. The average of the absolute difference between the CDF of the GPS vector and the produced power function is 0.034, 0.047, 0.049, and 0.055 for 3, 4, 6, and 10 treatments, respectively, for all simulated subjects.

2.5 Data Applications

To illustrate the utility of the above proposed methods, two data applications are conducted. First, electronic health records (EHR) data from the Cerner Health Facts database

are used to analyze the relationship between vasopressors and mortality in patients with non-traumatic aneurysmal subarachnoid hemorrhage (SAH). Additionally, the GPS-CDF methods are further applied to the Emergency Truncal Hemorrhage Control Study (ETHCS), a prospective observational study, to determine whether emerging hemorrhage control interventions influence patient mortality.

2.5.1 Cerner Health Facts database

The utility of the novel approach is demonstrated on EHR data from the Cerner Health Facts database. The database was used to analyze the relationship between vasopressor choice and mortality in patients with non-traumatic aneurysmal subarachnoid hemorrhage (SAH). SAH is defined as a blood vessel that bursts in the brain, and is a devastating cerebrovascular condition not only due to the effect of the hemorrhage but also the complicated treatment regimens required to manage such patients. A major complication resulting from SAH includes delayed cerebral ischemia (DCI), which is a main source of morbidity following SAH (Roy et al., 2017). Although current guidelines suggest maintaining an elevated blood pressure after management of an aneurysm may reduce the incidence of DCI, there is little data to suggest which vasopressor is the most efficacious to achieving this end with regards to mortality. The effectiveness of the three most commonly accepted drugs used to achieve an increase in blood pressure (dopamine, phenylephrine, and norepinephrine) are studied in relation to mortality in patients with non-traumatic SAH.

The study population included in the current analysis has been previously described (Williams et al., submitted). Briefly, the Cerner Health Facts EHR database was queried from years 2000 to 2015 to select adult patients (over age 17) with a new diagnosis of aneurysmal SAH based on ICD-9 code 430. Only patients who received infusions of dopamine,

phenylephrine, or norepinephrine were included in the study population (Williams et al., submitted). Among the 4,850 patients that met the above study inclusion criteria, 40 patients presented with multiple first vasopressor treatments; these patients were excluded from the cohort. Furthermore, patients whose diagnosis included a traumatic cause of SAH (based on ICD 9 codes: 800.2x, 800.7x, 801.2x, 801.7x, 803.2x, 803.7x, 804.2x, 804.7x, 852.x) or with unknown mortality status were excluded from the study population leaving 2,634 patients in the final cohort.

The propensity score analysis presented here includes 2,417 patients with complete data for demographic variables (age, gender, race, and marital status) as well as pretreatment medication variables. Of the patients included in the analysis, 492, 1,253, and 672 were administered dopamine, phenylephrine, and norepinephrine, respectively. In total, 170 pretreatment variables are entered into GBM in order to produce patient specific GPS vectors. More details on variable selection as well as variables included in the propensity model can be found in previous work (Williams et al., submitted).

The five analytical methods investigated within the simulation study are applied to this EHR dataset to determine the causal relationship between vasopressor choice and mortality. A visual representation of the covariate balance achieved by each method is depicted in Figure 2.4. The left plot shows maximum pairwise SMD for each potential pretreatment confounder. Similarly, the right plot shows the average pairwise SMD, which is calculated by averaging the SMD for each potential pretreatment confounder across treatment pairs. It has been suggested that values of SMD less than 0.2 indicate small levels of covariate imbalance (Cohen, 1988; McCaffrey et al., 2013). Based on this cutoff, both GBM weighting and GPS-CDF matching produce better covariate balance compared to the

original data, while GPS-CDF stratification produces less satisfactory levels of covariate balance.

As all subjects within this EHR dataset were eligible for all three treatments, ATE is the estimand of interest. Results from applying each of the five analytical methods, assessed within the simulation study, are shown in Table 2.2. GPS-CDF matching and stratification show that the odds of mortality are significantly higher in patients who received dopamine versus patients who received phenylephrine ($OR_{GPS-CDF \text{ Matching}} = 1.53$, 95% CI [1.11, 2.10], $p = 0.008$; $OR_{GPS-CDF \text{ Stratification}} = 2.59$, 95% CI [2.03, 3.31], $p = <0.001$). Patients receiving norepinephrine are found to have a higher odds of mortality versus patients receiving phenylephrine when analyses are conducted using the GPS-CDF stratification method ($OR_{GPS-CDF \text{ Stratification}} = 3.21$, 95% CI [2.55, 3.31], $p = <0.001$), but this association is not significant with GPS-CDF matching ($OR_{GPS-CDF \text{ Matching}} = 1.41$, 95% CI [1.00, 1.99], $p = 0.051$). Furthermore, GPS-CDF matching and stratification do not show any significant differences in the odds of mortality between patients who received dopamine and patients who received norepinephrine.

Importantly, all three propensity scoring approaches applied attenuate the unadjusted and covariate adjusted association between vasopressor choice and mortality. Overall, it does appear that phenylephrine is superior to dopamine in relation to mortality in patients with non-traumatic SAH, but the comparison between phenylephrine and norepinephrine remains unclear. Although results from GBM weighting indicate nearly a 50% reduction in mortality in patients given phenylephrine, the effects of vasopressor choice on patient mortality are not as strong when the analyses are conducted using GPS-CDF matching. GPS-CDF matching creates satisfactory levels of covariate balance within the data and further attenuates the association between vasopressor choice and mortality. Again, outcome models to obtain

ATEs utilize logistic regression, survey-weighted generalized linear models, and conditional logistic regression for the unadjusted and adjusted model, GBM weighting, and GPS-CDF matching and stratification, respectively. Additionally, the GPS-CDF approach is computationally quick; results were available in 8 minutes using a dual-core Intel Core i3-3110M with 4 GB RAM.

2.5.2 *Emergency Truncal Hemorrhage Control Study*

Severe hemorrhage of the non-compressible torso is the leading cause of potentially survivable deaths in trauma cases. Non-compressible torso hemorrhage (NCTH) is defined as blood loss due to trauma of the torso (chest, abdomen, and pelvis), pulmonary parenchyma, solid abdominal organs, and disruption of the bony pelvis causing hypotension or shock (Stannard, Eliason and Rasmussen, 2011; Eastridge et al., 2012; Kisat et al., 2013). A new treatment, namely, resuscitative endovascular balloon occlusion of the aorta (REBOA), is a technique that could temporarily mitigate hemorrhage from the abdomen and pelvis. ETHCS aims to compare various hemorrhage control techniques (laparotomy, thoracotomy, and REBOA) in relation to patient mortality.

As ETHCS is an observational study, and patients undergoing REBOA or other procedures (laparotomy or thoracotomy) have different covariate distributions, it is an excellent example for the utility of the GPS-CDF multinomial propensity score method. The current analysis contains 409 subjects, of which 264 (64.5%), 67 (16.4%), and 78 (19.1%) were treated with laparotomy, thoracotomy, and REBOA, respectively. Again, the five analytical methods investigated within the simulation study are applied to this ETHCS dataset to determine the causal relationship between hemorrhage control techniques and mortality. A visual representation of the covariate balance achieved by each method is depicted in Figure 2.5. The plot shows the average pairwise SMD for all pre-treatment

confounders: age, race, gender, injury mechanism, and study site. Based on Figure 2.5, GPS-CDF matching produces better covariate balance compared to all other methods investigated.

Again as all subjects were eligible for all three hemorrhage control techniques, ATE is the estimand of interest. Mortality results were not computed for comparisons with the thoracotomy group, as 94% of subjects within this group died. GPS-CDF matching shows that the odds of mortality are not significantly different for patients treated with REBOA compared to patients treated with laparotomy ($OR_{GPS-CDF \text{ Matching}} = 5.75$, 95% CI [0.98, 33.83], $p = 0.053$). Conversely, results within the unadjusted and GBM weighted models do indicate a significant difference in mortality between these two techniques ($OR_{Unadjusted} = 6.54$, 95% CI [3.70, 11.56], $p < 0.001$; $OR_{GBM \text{ Weighted}} = 8.31$, 95% CI [3.46, 19.94], $p < 0.001$). As the balance achieved via GPS-CDF matching is by far superior to any other method investigated (Figure 2.5), the results suggest that there is no difference in mortality between hemorrhage control techniques (REBOA and laparotomy) within the study population.

2.6 Discussion

Although methods exist to conduct propensity scoring in the presence of multinomial treatments (e.g. Seya and Yoshida, 2017; Rassen et al., 2013; Rubin, 1979; Zhao, 2004; Tu, Jiao and Koh, 2013; Lopez and Gutman, 2017; Zanutto, Lu and Hornik, 2005; Huang et al., 2005; Yang et al., 2016; Lechner, 2001; Feng et al., 2012), few methods have the capability and flexibility to estimate both ATE and ATT and correctly model data sources that present with large covariate spaces. Recently, researchers have advocated for the use of machine learning propensity models to produce more accurate GPS vectors especially in the presence of a large covariate space (Setoguchi et al., 2008; Guertin et al., 2016; Chen and Moskowitz, 2016). Although the benefits of using GBM and other machine learning methods as detailed

above, are apparent, there are still drawbacks to these methods that need to be addressed. Currently, the GPS vector produced by machine learning methods is adjusted in outcome analysis via IPTW. Although IPTW is an easily adaptable method in order to produce causal treatment effect estimates, Rubin (Rubin, 2004) suggests that weighting directly on the propensity score leads to a higher degree of sensitivity to model misspecification. Furthermore, via simulation, Yang et al. (2016) show that when presented with six treatment groups (not an impossibly large number when considering EHR-derived studies), implementation of the GPS via IPTW leads to extreme weights. For example, the maximum weights reported by Yang et al. (2016) are 95.8 within in a three treatment scenario and 185.1 within a six treatment scenario. Although these extreme weights may not adversely impact covariate balance, they will lead to inaccurate ATEs/ATTs. Given the published limitations for IPTW for multiple treatments propensity scoring, this paper derived and tested via simulation and practice, a novel multinomial propensity scoring technique that utilizes the entire GPS vector. The GPS-CDF method directly maps the GPS vector resulting from any propensity model to a scalar value that is easily used for matching and stratification to produce either ATEs or ATTs. As this method generates $K!$ balancing scores, it follows closely to the current opinion in the literature of the ‘covariate balancing propensity score’ (Fong, Hazlett and Imai, 2018; Imai and Ratkovic, 2014) and the ‘distance adjusted propensity score’ (Papadogeorgou, Choirat and Zigler, 2018).

The proposed method of mapping the CDF of the GPS vector is given by equation (2.4). While other methods may be used to map CDFs (e.g. Kolmogorov-Smirnov test statistic, isotonic regression, Kullback–Leibler divergence, etc.), they do not result in a scalar balancing score, analogous to the scalar value derived within binary treatment propensity scoring. For example, the Kolmogorov-Smirnov test statistic tests the equality of CDFs

through a distance based metric (Massey, 1951). Although this approach may be used to match subjects with similar CDFs, the resultant pairwise distances cannot be easily adapted to stratify subjects. Furthermore, other common one parameter functions (e.g. exponential, logarithmic, sigmoid) do not have the same flexibility as the power function for mapping CDFs that present with both concave and convex shapes. Thus the proposed GPS-CDF method utilizes a one parameter power function in order to accurately map CDFs via a scalar value, which may be used to match and stratify subjects.

The current simulation study closely followed several recently published simulations (Austin, Grootendorst and Anderson, 2007; Fong, Hazlett and Imai, 2018; Greene, 2017); matching and stratification via the GPS-CDF method produced better covariate balance than the original data in both the correctly and incorrectly specified treatment model. Although GBM weighting produced better covariate balance compared to GPS-CDF matching and stratification within the correctly specified treatment model, similar to results presented by Fong et al. (2018), this increased balance did not translate to more accurate ATE estimates. GBM weighting produced highly biased estimates compared to GPS-CDF stratification for each treatment comparison when both the treatment and outcome models were misspecified.

Unlike IPTW which has been shown to produce unreliable estimates in the presence of multiple treatment groups (Yang et al., 2016), the GPS-CDF method is still valid. Using data simulated under different multinomial treatment group scenarios, the GPS-CDF method was able to accurately map CDFs of the GPS even in the presence of numerous treatment groups. When presented with 10 treatments, the average difference between the true CDF of the GPS vector and the estimated power function was minimal. Since the mapping capabilities of the method was shown to still be valid even in the presence of numerous treatment groups, the GPS-CDF method may produce more accurate causal inference

estimates compared to those derived by IPTW, especially in the presence of multiple treatment groups.

Finally, the performance and utility of the GPS-CDF method was further demonstrated using two data applications. First, data from the Cerner Health Facts database were queried in order to assess the association between vasopressor choice and mortality in patients with non-traumatic SAH. Overall, the novel multinomial propensity analysis approach, GPS-CDF, had low computational burden and produced better covariate balance compared to the original (unadjusted) data when applied via matching. Additionally, this EHR data example demonstrates the easy applicability of the GPS-CDF approach. These results further indicate that prospective studies should be conducted in order to determine which vasopressor is the most efficacious for patients with non-traumatic SAH. Additionally, the GPS-CDF methods were applied to the ETHCS to determine whether emerging hemorrhage control interventions influence patient mortality. These results demonstrate that REBOA has a similar effect on patient mortality compared to laparotomy.

2.7 Conclusion

This paper details the derivation and application of the GPS-CDF method that removes covariate imbalance in observational studies with multinomial treatments. Currently, no methods exist that transform the GPS vector into a single number, analogous to the single scalar balancing score found in binary treatment propensity scoring. Using a NLS algorithm, the GPS-CDF method directly maps any GPS vector to a scalar value which easily facilitates either matching or stratification in order to produce causal treatment effect estimates. Importantly, the scalar value derived from the GPS-CDF method can be adapted to produce either ATE or ATT estimates. Our detailed simulation study found that

implementation of the GPS-CDF method via stratification may lead to less biased causal inference estimates compared to methods based on IPTW. Furthermore, when applied to an EHR data set, the GPS-CDF method indicates that phenylephrine may be the superior vasopressor choice for patients that present with non-traumatic SAH.

There are limitations of this study. First, the EHR data application was derived from the Cerner Health Facts database, which contained a large number of patients with complete covariate data. Patients were included in the analysis based on a new diagnoses of SAH, but their diagnosis could not be confirmed via imaging. Additionally, due to the absence of baseline diagnostic variables, there is of course the possibility of unmeasured confounding within the analysis, as with any propensity score analysis, especially one derived from EHR. Furthermore, within the ETHCS data application, 94% of patients who received thoracotomy died. Thus meaningful analyses were not able to be conducted using this treatment group.

The GPS-CDF method presented here gives researchers more options when conducting multinomial treatment propensity scoring. This novel method can be used in conjunction with current machine learning methods in order to better facilitate propensity score adjustment in the presence of big data. Future studies should further evaluate the use of the GPS-CDF method when conducting propensity scoring with multinomial treatments in the context of relevant research questions. Open-source software is available to help facilitate the use of the proposed method in practice (Brown et al., 2019).

| | Strongly Associated with Treatment | Moderately Associated with Treatment | Independent of Treatment |
|---------------------------------------|---------------------------------------|---|-----------------------------|
| Strongly Associated with Outcome | x_1 | x_2 | x_3 |
| Moderately Associated with Outcome | x_4 | x_5 | x_6 |
| Independent of Outcome | x_7 | x_8 | x_9 |

Table 2.1. True association between baseline covariates with treatment and outcome. Note, x_1 , x_2 , x_4 , and x_5 are simulated to be pretreatment confounders.

| Analytical Method | Dopamine vs Phenylephrine | | Norepinephrine vs Phenylephrine | | Dopamine vs Norepinephrine | |
|------------------------|---------------------------|---------|---------------------------------|---------|----------------------------|---------|
| | OR [95% CI] | p-value | OR [95% CI] | p-value | OR [95% CI] | p-value |
| Unadjusted | 3.06 [2.46-3.81] | <0.001 | 2.53 [2.08-3.09] | <0.001 | 1.21 [0.96-1.53] | 0.110 |
| Adjusted | 3.02 [2.42-3.76] | <0.001 | 2.63 [2.15-3.22] | <0.001 | 1.15 [0.91-1.45] | 0.253 |
| GBM Weighted | 2.19 [1.70-2.81] | <0.001 | 2.24 [1.80-2.80] | <0.001 | 0.97 [0.75-1.27] | 0.852 |
| GPS-CDF Greedy Matched | 1.53 [1.11-2.10] | 0.008 | 1.41 [1.00-1.99] | 0.051 | 1.09 [0.79-1.49] | 0.610 |
| GPS-CDF Stratification | 2.59 [2.03-3.31] | <0.001 | 3.21 [2.55-3.31] | <0.001 | 0.81 [0.61-1.07] | 0.138 |

Table 2.2. Model estimates after applying each analytical method to SAH patients to determine the association between vasopressor and mortality within the EHR dataset. Outcome models (with the exception of the unadjusted model) adjusted for age at baseline, gender, race, and marital status.

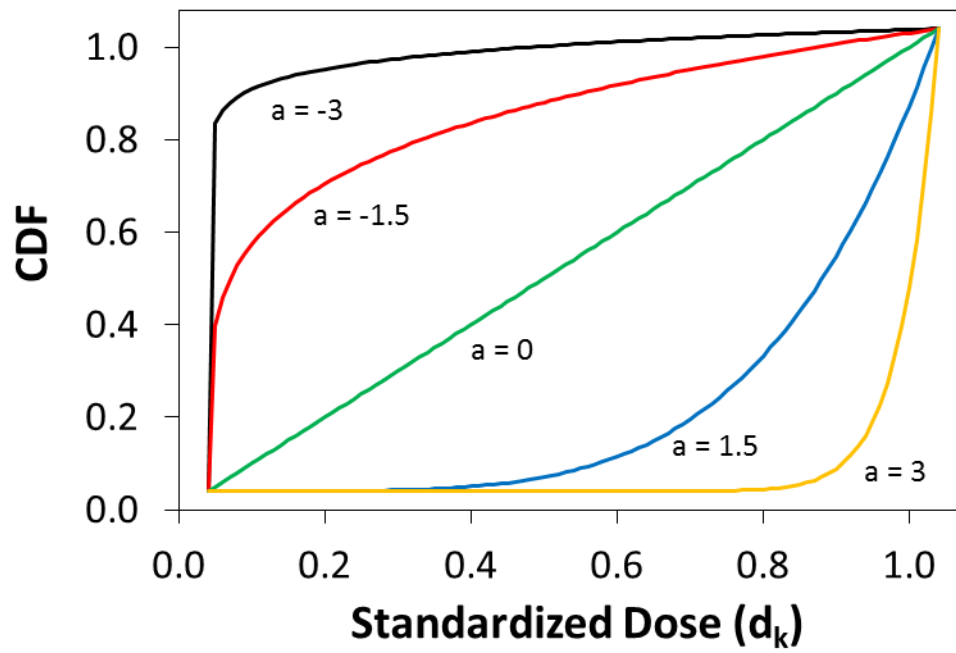


Figure 2.1. Graphical representation of the convex and concave modeling produced by the power function.

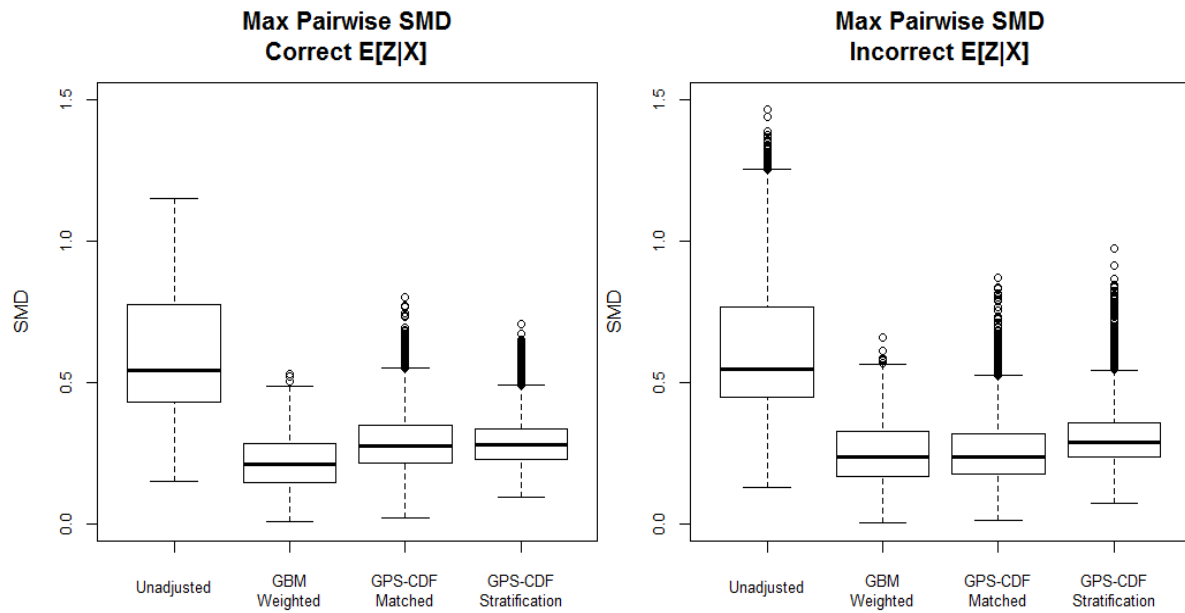


Figure 2.2. Graphical representation of the covariate balance achieved by each method under the correctly specified and incorrectly specified treatment assignment models. SMD was calculated for all baseline covariates within each treatment pair, and the maximum SMD across treatment pairs was retained.

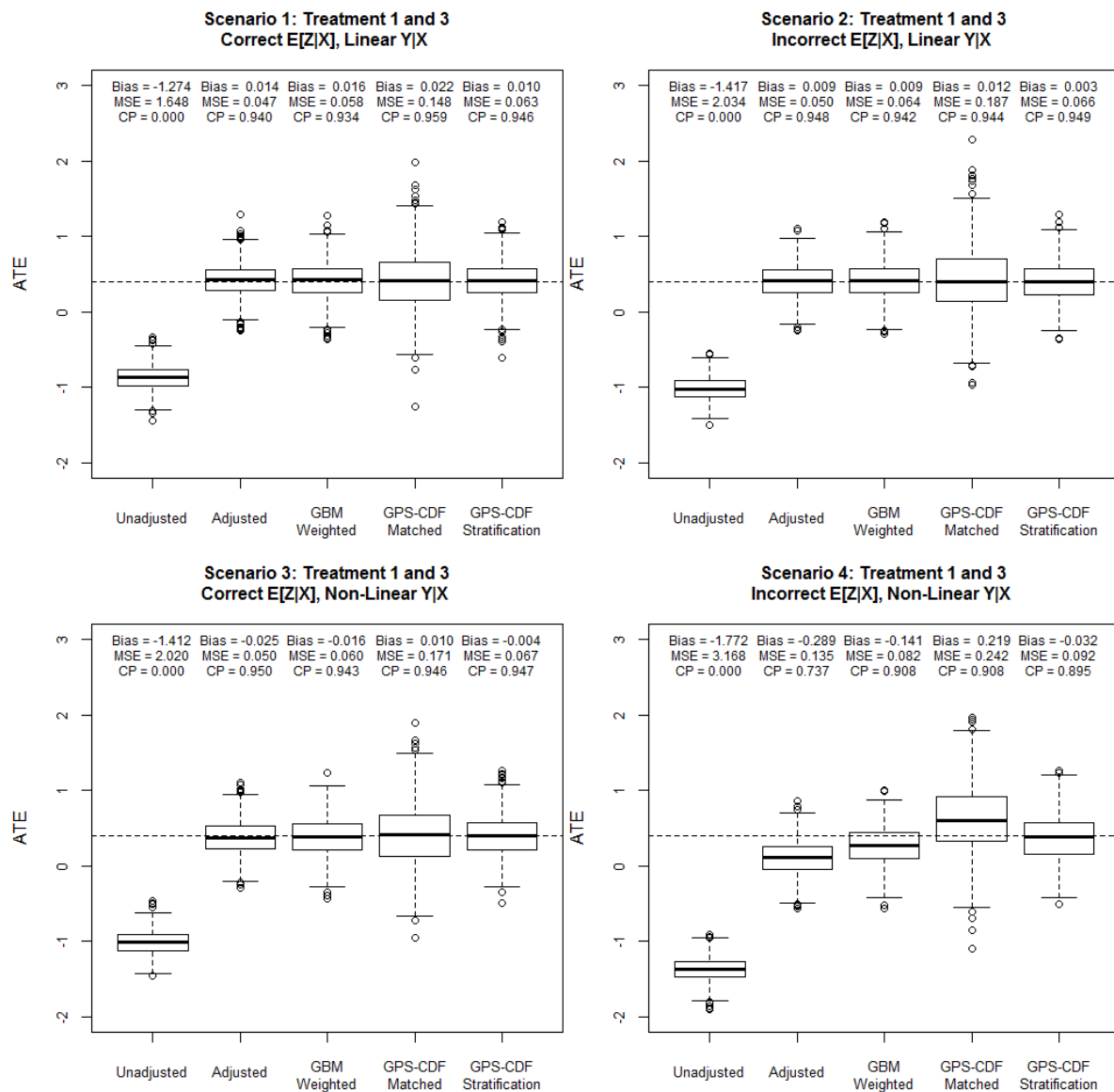


Figure 2.3. Distribution of the ATE for each method under each scenario between treatment 1 and treatment 3. The true ATE value of 0.4 is included as the dotted horizontal line.

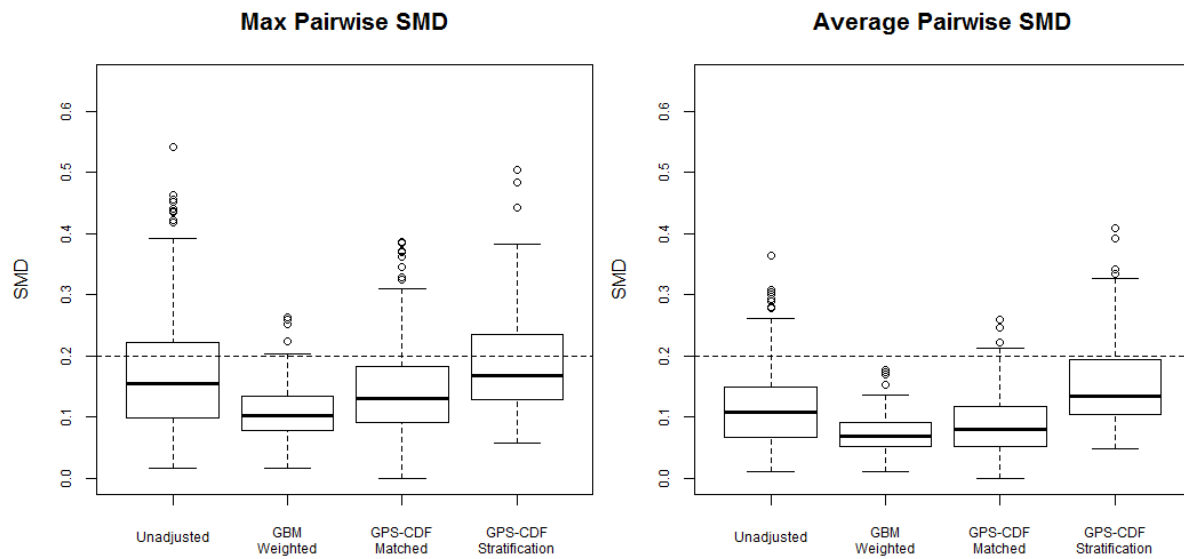


Figure 2.4. Graphical representation of the covariate balance achieved by each method for SAH patients within the Cerner Health Facts EHR database. The left plot presents the maximum pairwise SMD across treatment groups for each potential confounder. The right plot presents the average pairwise SMD across treatment groups for each potential confounder.

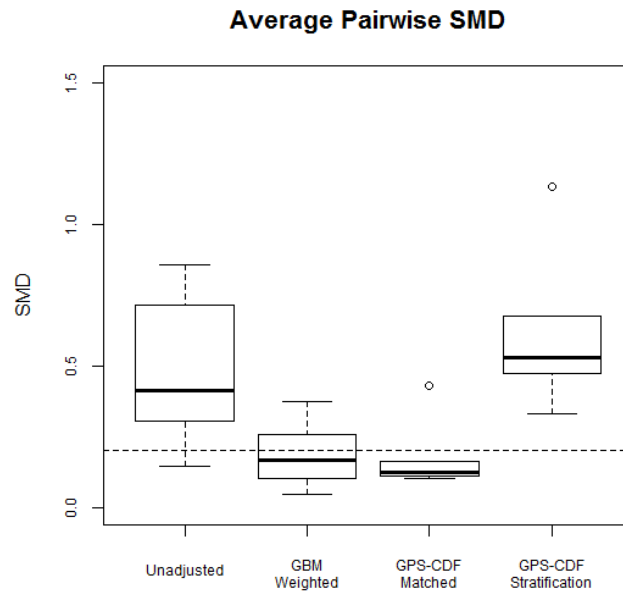


Figure 2.5. Graphical representation of the covariate balance achieved by each method for hemorrhage patients within the Emergency Truncal Hemorrhage Control Study. The plot presents the average pairwise SMD across treatment groups for each potential confounder.

References

- Austin, P. C. (2011). An introduction to propensity score methods for reducing the effects of confounding in observational studies. *Multivariate behavioral research* **46**, 399-424.
- Austin, P. C., Grootendorst, P., and Anderson, G. M. (2007). A comparison of the ability of different propensity score models to balance measured variables between treated and untreated subjects: a Monte Carlo study. *Stat Med* **26**, 734-753.
- Brookhart, M. A., Schneeweiss, S., Rothman, K. J., Glynn, R. J., Avorn, J., and Stürmer, T. (2006). Variable selection for propensity score models. *American journal of epidemiology* **163**, 1149-1156.
- Brown, D. W., Greene, T. J., and DeSantis S. M. (2019). GPSCDF: Generalized Propensity Score Cumulative Distribution Function. *R Package. Rand Corporation*.
- Burgette, L., Griffin, B. A., and McCaffrey, D. (2017). Propensity scores for multiple treatments: A tutorial for the mnps function in the twang package. *R package. Rand Corporation*.
- Busso, M., DiNardo, J., and McCrary, J. (2014). New evidence on the finite sample properties of propensity score reweighting and matching estimators. *Review of Economics and Statistics* **96**, 885-897.
- Chen, K. P., and Moskowitz, A. (2016). Comparative Effectiveness: Propensity Score Analysis. In *Secondary Analysis of Electronic Health Records*, 339-349. Cham: Springer International Publishing.
- Cochran, W. G. (1968). The effectiveness of adjustment by subclassification in removing bias in observational studies. *Biometrics*, 295-313.

- Cochran, W. G., and Rubin, D. B. (1973). Controlling Bias in Observational Studies: A Review. *Sankhyā: The Indian Journal of Statistics, Series A (1961-2002)* **35**, 417-446.
- Cohen, J. (1988). *Statistical power analysis for the behavioral sciences*. Hillsdale, N.J.: L. Erlbaum Associates.
- Eastridge, B. J., Mabry, R. L., Seguin, P., Cantrell, J., Tops, T., Uribe, P., . . . Rasmussen, T. E. (2012). Death on the battlefield (2001–2011): implications for the future of combat casualty care. *Journal of Trauma and Acute Care Surgery*, *73*(6), S431-S437.
- Feng, P., Zhou, X. H., Zou, Q. M., Fan, M. Y., and Li, X. S. (2012). Generalized propensity score for estimating the average treatment effect of multiple treatments. *Stat Med* **31**, 681-697.
- Fong, C., Hazlett, C., and Imai, K. (2018). Covariate balancing propensity score for a continuous treatment: Application to the efficacy of political advertisements. *The Annals of Applied Statistics* **12**, 156-177.
- Greene, T.J. (2017). Utilizing Propensity Score Methods for Ordinal Treatments and Prehospital Trauma Studies. *Texas Medical Center Dissertations (via ProQuest)*.
- Guertin, J. R., Rahme, E., Dormuth, C. R., and LeLorier, J. (2016). Head to head comparison of the propensity score and the high-dimensional propensity score matching methods. *BMC Med Res Methodol* **16**, 22.
- Gutman, R., and Rubin, D. B. (2015). Estimation of causal effects of binary treatments in unconfounded studies. *Stat Med* **34**, 3381-3398.

- Hirano, K., and Imbens, G. W. (2001). Estimation of causal effects using propensity score weighting: An application to data on right heart catheterization. *Health Services and Outcomes research methodology* **2**, 259-278.
- Huang, I.-C., Frangakis, C., Dominici, F., Diette, G. B., and Wu, A. W. (2005). Application of a propensity score approach for risk adjustment in profiling multiple physician groups on asthma care. *Health services research* **40**, 253-278.
- Imai, K., and Ratkovic, M. (2014). Covariate balancing propensity score. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **76**, 243-263.
- Imai, K., and Van Dyk, D. A. (2004). Causal inference with general treatment regimes: Generalizing the propensity score. *Journal of the American Statistical Association* **99**, 854-866.
- Imbens, G. W. (2000). The role of the propensity score in estimating dose-response functions. *Biometrika* **87**, 706-710.
- Joffe, M. M., and Rosenbaum, P. R. (1999). Invited commentary: propensity scores. *American journal of epidemiology* **150**, 327-333.
- Ju, C., Combs, M., Lendle, S. D., *et al.* (2019). Propensity score prediction for electronic healthcare databases using super learner and high-dimensional propensity score methods. *Journal of Applied Statistics*, 1-21.
- Kisat, M., Morrison, J. J., Hashmi, Z. G., Efron, D. T., Rasmussen, T. E., and Haider, A. H. (2013). Epidemiology and outcomes of non-compressible torso hemorrhage. *journal of surgical research*, 184(1), 414-421.

- Lechner, M. (2001). Identification and estimation of causal effects of multiple treatments under the conditional independence assumption. In *Econometric evaluation of labour market policies*, 43-58: Springer.
- Li, F., Morgan, K. L., and Zaslavsky, A. M. (2018). Balancing covariates via propensity score weighting. *Journal of the American Statistical Association* **113**, 390-400.
- Lopez, M. J., and Gutman, R. (2017). Estimation of causal effects with multiple treatments: a review and new ideas. *Statistical Science* **32**, 432-454.
- Low, Y. S., Gallego, B., and Shah, N. H. (2016). Comparing high-dimensional confounder control methods for rapid cohort studies from electronic health records. *J Comp Eff Res* **5**, 179-192.
- Lunt, M. (2014). Selecting an Appropriate Caliper Can Be Essential for Achieving Good Balance With Propensity Score Matching. *American journal of epidemiology* **179**, 226-235.
- Marquardt, D. W. (1963). An algorithm for least-squares estimation of nonlinear parameters. *Journal of the society for Industrial and Applied Mathematics* **11**, 431-441.
- Massey, F. J. (1951). The Kolmogorov-Smirnov Test for Goodness of Fit *Journal of the American Statistical Association* **46**, 68-78.
- McCaffrey, D. F., Griffin, B. A., Almirall, D., Slaughter, M. E., Ramchand, R., and Burgette, L. F. (2013). A tutorial on propensity score estimation for multiple treatments using generalized boosted models. *Stat Med* **32**, 3388-3414.
- McCaffrey, D. F., Ridgeway, G., and Morral, A. R. (2004). Propensity score estimation with boosted regression for evaluating causal effects in observational studies. *Psychological methods* **9**, 403.

- O'Quigley, J., Pepe, M., and Fisher, L. (1990). Continual reassessment method: a practical design for phase 1 clinical trials in cancer. *Biometrics* **46**, 33-48.
- Papadogeorgou, G., Choirat, C., and Zigler, C. M. (2018). Adjusting for unmeasured spatial confounding with distance adjusted propensity score matching. *Biostatistics*.
- Patorno, E., Glynn, R. J., Hernandez-Diaz, S., Liu, J., and Schneeweiss, S. (2014). Studies with many covariates and few outcomes: selecting covariates and implementing propensity-score-based confounding adjustments. *Epidemiology* **25**, 268-278.
- Rassen, J. A., Shelat, A. A., Franklin, J. M., Glynn, R. J., Solomon, D. H., and Schneeweiss, S. (2013). Matching by propensity score in cohort studies with three treatment groups. *Epidemiology* **24**, 401-409.
- Rosenbaum, P. R., and Rubin, D. B. (1983). The Central Role of the Propensity Score in Observational Studies for Causal Effects. *Biometrika* **70**, 41-55.
- Rosenbaum, P. R., and Rubin, D. B. (1984). Reducing bias in observational studies using subclassification on the propensity score. *Journal of the American Statistical Association* **79**, 516-524.
- Rosenbaum, P. R., and Rubin, D. B. (1985). Constructing a control group using multivariate matched sampling methods that incorporate the propensity score. *The American Statistician* **39**, 33-38.
- Roy, B., McCullough, L. D., Dhar, R., Grady, J., Wang, Y.-B., and Brown, R. J. (2017). Comparison of initial vasopressors used for delayed cerebral ischemia after aneurysmal subarachnoid hemorrhage. *Cerebrovascular Diseases* **43**, 266-271.

- Rubin, D. B. (1979). Using multivariate matched sampling and regression adjustment to control bias in observational studies. *Journal of the American Statistical Association* **74**, 318-328.
- Rubin, D. B. (2004). On principles for modeling propensity scores in medical research. *Pharmacoepidemiology and drug safety* **13**, 855-857.
- Schneeweiss, S., Rassen, J. A., Glynn, R. J., Avorn, J., Mogun, H., and Brookhart, M. A. (2009). High-dimensional propensity score adjustment in studies of treatment effects using health care claims data. *Epidemiology* **20**, 512-522.
- Schuemie, M. J., Coloma, P. M., Straatman, H., *et al.* (2012). Using electronic health care records for drug safety signal detection: a comparative evaluation of statistical methods. *Med Care* **50**, 890-897.
- Setoguchi, S., Schneeweiss, S., Brookhart, M. A., Glynn, R. J., and Cook, E. F. (2008). Evaluating uses of data mining techniques in propensity score estimation: a simulation study. *Pharmacoepidemiology and drug safety* **17**, 546-555.
- Seya, H., and Yoshida, T. (2017). Propensity score matching for multiple treatment levels: A CODA-based contribution. *arXiv preprint arXiv:1710.08558*.
- Stannard, A., Eliason, J. L., and Rasmussen, T. E. (2011). Resuscitative endovascular balloon occlusion of the aorta (REBOA) as an adjunct for hemorrhagic shock. *Journal of Trauma and Acute Care Surgery*, *71*(6), 1869-1872.
- Storer, B. E. (1989). Design and analysis of phase I clinical trials. *Biometrics* **45**, 925-937.
- Stuart, E. A. (2010). Matching methods for causal inference: A review and a look forward. *Statistical science: a review journal of the Institute of Mathematical Statistics* **25**, 1.

- Stuart, E. A., DuGoff, E., Abrams, M., Salkever, D., and Steinwachs, D. (2013). Estimating causal effects in observational studies using Electronic Health Data: Challenges and (some) solutions. *EGEMS (Wash DC)* **1**.
- Tu, C., Jiao, S., and Koh, W. Y. (2013). Comparison of clustering algorithms on generalized propensity score in observational studies: a simulation study. *Journal of Statistical Computation and Simulation* **83**, 2206-2218.
- Yang, S., Imbens, G. W., Cui, Z., Faries, D. E., and Kadziola, Z. (2016). Propensity score matching and subclassification in observational studies with multi-level treatments. *Biometrics* **72**, 1055-1065.
- Williams, G., Maroufy, V., Rasmy, L., *et al.* (2019). Mortality Can Be Significantly Reduced with Appropriate Vasopressor Choice in Patients with Non-Traumatic Subarachnoid Hemorrhage: A Nationwide EHR Analysis. *New England Journal of Medicine*, submitted.
- Zanutto, E., Lu, B., and Hornik, R. (2005). Using propensity score subclassification for multiple treatment doses to evaluate a national antidrug media campaign. *Journal of Educational and Behavioral Statistics* **30**, 59-73.
- Zhao, Z. (2004). Using matching to estimate treatment effects: Data requirements, matching metrics, and Monte Carlo evidence. *Review of Economics and Statistics* **86**, 91-107.

CHAPTER III

Sampling Based Propensity Score Stratification for Continuous Treatments

Proposed Journal: Biometrics

Derek W. Brown¹, Thomas J. Greene², Michael D. Swartz¹, Anna V. Wilkinson³, Stacia M. DeSantis¹

Author Affiliations

1. Department of Biostatistics and Data Science, University of Texas Health School of Public Health
2. GlaxoSmithKline, Division of Biostatistics
3. Department of Epidemiology, Human Genetics, and Environmental Sciences, University of Texas School of Public Health

Corresponding Author:

Derek Brown, PhD(c), MS
Department of Biostatistics and Data Science
University of Texas Health Science Center at Houston
1200 Pressler Street, Houston, TX 77030
[derek.brown@uth.tmc.edu]

Abstract

Continuous treatments propensity scoring remains understudied, as the majority of methods are focused on the binary treatment setting. Current propensity score methods for continuous treatments typically rely on weighting in order to produce causal estimates. It has been shown that in some cases, weighting methods can result in worse covariate balance than had no adjustments been made to the data. Furthermore, weighting is not always stable, and resultant estimates may be unreliable due to extreme weights. These issues motivate the current development of novel propensity score stratification techniques to be used with continuous treatments. Specifically, the generalized propensity score cumulative distribution function (GPS-CDF) and the nonparametric GPS-CDF (npGPS-CDF) approaches are introduced. Empirical CDFs are used to stratify subjects based on pretreatment confounders, in order to produce causal estimates. A detailed simulation study shows superiority of these new stratification methods based on the empirical CDF, when compared to standard weighting techniques. The proposed methods are applied to the “Mexican American Tobacco use in Children” (MATCH) study to determine the causal relationship between continuous exposure to smoking imagery in movies, and smoking behavior among Mexican-American adolescents. These promising results provide investigators with new options for implementing continuous treatment propensity scoring.

Keywords: Causal Inference, Continuous Treatment, Observational Study, Propensity Score

3.1 Introduction

Propensity scoring is often used to make causal inference about a treatment/exposure-outcome relationship in non-randomized observational studies. Although methods for binary and more recently, multiple treatments have been well-studied (e.g., Rosenbaum and Rubin, 1983, 1984, 1985; Joffe and Rosenbaum, 1999; Imbens, 2000; Imai and Van Dyk, 2004), there has been less research devoted to propensity score methods for continuous treatments. In this paper, continuous treatments refer to treatment assignment (e.g. dosing trials) or continuous exposures (e.g. environmental exposures). In the presence of continuous treatments, investigators may instead dichotomize or categorize the treatment in order to utilize more well-established propensity score techniques (e.g. Chertow, Normand, and McNeil, 2004; Davidson et al., 2006; Donohue and Ho, 2007; Flores-Lagunes, Gonzalez, and Neumann, 2007; Harder, Stuart, and Anthony, 2008; Boyd, Epstein, and Martin, 2010; Nielsen et al., 2011; Greene, 2017). However, it has been shown that categorization of a continuous treatment may lead to loss of information and subsequent decrease in power when conducting outcome analyses (Royston, Altman, and Sauerbrei, 2006; Zhu, Coffman, and Ghosh, 2015; Fong, Hazlett, and Imai, 2018). Also, not analyzing exposures on their original scale can produce clinical interpretations that are awkward to domain-area researchers.

Propensity scoring methods directly applicable to continuous exposures have been proposed. For example, maximum likelihood estimates (MLEs) derived from linear models have been used to estimate the generalized propensity score (GPS) (Robins, Hernan, and Brumback, 2000; Imai and Van Dyk, 2004; Hirano and Imbens, 2004). In practice, the GPS can be obtained by fitting a linear regression model of the form

$$T = \boldsymbol{\beta}\mathbf{X} + \varepsilon \tag{3.1}$$

where T is a continuous treatment, \mathbf{X} is a vector of potential confounders, and

$\varepsilon \sim N(0, \sigma^2)$. The GPS for individual i , is estimated as,

$$\hat{R}_i = \frac{1}{\sqrt{2\pi\hat{\sigma}^2}} \exp\left(-\frac{1}{2\hat{\sigma}^2}(T_i - \hat{T}_i)^2\right) \quad (3.2)$$

(Hirano and Imbens, 2004). The GPS is then used to remove covariate bias by first estimating the conditional expectation of the outcome as a function of the treatment level (T) and the GPS (R),

$$\beta(t, r) = E[Y|T = t, R = r] \quad (3.3)$$

(Hirano and Imbens, 2004). The dose-response function, i.e. the average response in the sample, is then estimated at a particular treatment level by averaging equation (3.3) over the GPS at that level of treatment,

$$\mu(t) = E[\beta(t, r(t, X))] \quad (3.4)$$

(Hirano and Imbens, 2004). By calculating the dose response function at two treatment levels, i.e. $\mu(t_1)$ and $\mu(t_2)$, the mean change in the outcome can be estimated (Austin 2018b). In practice, although the estimated GPS, \hat{R}_i , can be applied directly in regression adjustment (Hirano and Imbens, 2004), or the scalar value, $\hat{\beta}\mathbf{X}_i$, can be utilized for matching or stratification (Imai and Van Dyk, 2004), to produce causal estimates; using the GPS in a weighted outcome analyses has been prioritized recently (e.g., Robins et al., 2000; Zhu et al., 2015; Schuler, Chu, and Coffman, 2016; Fong et al., 2018; Austin, 2018a; Austin, 2018b).

Specifically, Robins et al. (2000) propose using the GPS to produce causal estimates using inverse probability weighting (IPW) (Robins et al., 2000). Briefly, IPW weights each individual with the inverse of the probability of receiving the treatment they actually received, given the covariates. By up-weighting those individuals less likely to receive the

treatment, IPW has the advantage of giving more weight in the analysis to subjects with dissimilar covariate distributions than subjects with similar covariate profiles (i.e. subject specific covariate values) within the same treatment level. In the calculation of \hat{R}_i , subjects with unexpected covariate distributions will have large estimates for $T_i - \hat{T}_i$, and conversely will have small values for \hat{R}_i . Thus, the IPW, given by

$$w_i = \frac{1}{\hat{R}_i} \quad (3.5)$$

will be higher, effectively giving more weight to subjects that have unexpected covariate distributions based on their continuous treatment level. Weights of the above form have infinite variance, so a stabilizing factor is applied to w_i in practice, called, $W(T_i)$, (Robins et al., 2000), given by the marginal density of T , which may be estimated by first fitting an intercept only model of the form

$$T = \beta_0 + \varepsilon \quad (3.6)$$

where $\varepsilon \sim N(0, \sigma_{sample}^2)$. The stabilizing factor is estimated as

$$\hat{W}(T_i) = \frac{1}{\sqrt{2\pi\hat{\sigma}_{sample}^2}} \exp\left(-\frac{1}{2\hat{\sigma}_{sample}^2}(T_i - \hat{\mu})^2\right) \quad (3.7)$$

where $\hat{\mu}$ is the mean treatment value of the sample (Austin, 2018b). The final estimated stabilized IPW is given by

$$sw_i = \frac{\hat{W}(T_i)}{\hat{R}_i}, \quad (3.8)$$

which is utilized in a weighted outcome regression of the form

$$E(Y|T) = \alpha + \beta_T T \quad (3.9)$$

in order to estimate the average treatment effect (ATE) (Schuler et al., 2016).

Although the calculation of these weights is straightforward, the MLE method detailed above relies heavily on correctly specifying the linear treatment model. If the model is not correctly specified or the model assumptions are not met (e.g., deviations from normality of errors), it has been shown that the MLE method can produce extreme weights that can lead to severely biased causal inference estimates (Fong et al., 2018). Therefore, methods that operate outside of the MLE framework may produce better weights, resulting in more covariate balance, and less biased estimates of the outcome.

Nonparametric methods of estimating the GPS vector have been shown to provide more accurate estimates of the GPS compared to parametric regression (Bia et al., 2014; Zhu et al., 2015). One such method that has gathered traction is the generalized boosted model (GBM) (Zhu et al., 2015; Fong et al., 2018). GBM fits a general model of the form,

$$T = m(\mathbf{X}) + \varepsilon \quad (3.10)$$

where $\varepsilon \sim N(0, \sigma^2)$ and $m(\mathbf{X})$ is the mean function of T given \mathbf{X} (Zhu et al., 2015). The mean function is estimated using a boosting algorithm that additively fits regression trees until the model is sufficiently flexible to fit the data (McCaffrey et al., 2013; Zhu et al., 2015). With the mean function derived, stabilized IPWs can be calculated and implemented just as in the MLE weighting procedure. Although it may appear as though GBM ultimately provides minimally biased causal inference estimates, there are still drawbacks that limit its usefulness. First, GBM does not afford users the ability to force variables into the final treatment model (Ridgeway et al., 2016), which is often appropriate in biomedical research (for example, age, gender, and other demographic or baseline clinical information). Additionally, although GBM has been shown to outperform MLE in simulation studies, covariate balance after GBM weighting can still remain poor, subsequently resulting in more

unstable estimates than if weights were not applied at all (Fong et al., 2018). Finally, the primary way to improve balance using GBM is by increasing the number of regression trees used by the method, which may not provide adequate control over sample imbalance (Fong et al., 2018).

Recent important extensions of the “covariate balancing propensity score,” which models treatment assignment while optimizing covariate balance, have been made for continuous treatments (Imai and Ratkovic, 2014; Fong et al., 2018). Specifically, the new covariate balancing generalized propensity score (CBGPS) uses the method of moments framework to derive IPWs such that the weighted correlation between \mathbf{X} and T is minimized (Fong et al., 2018). The nonparametric extension of this CBGPS (npCBGPS) places no parametric restrictions on the GPS, as weights are directly derived without giving a functional form to the propensity scores.

Although the CBGPS is a method for optimizing covariate balance while estimating the GPS, it is not without limitations as shown in Fong et al. (2018). In simulation, it was shown that GBM produces less biased causal estimates compared to CBGPS and npCBGPS when sample sizes are large ($\sim 1,000$). Additionally, since the nonparametric extension, npCBGPS, is based on an empirical likelihood approach, there is no guarantee that the optimization procedures find the global optimum. Furthermore, when the number of covariates is large, or if \mathbf{X} strongly predicts T , the npCBGPS may fail to find a solution, leaving the investigator to sacrifice covariate balance to derive weights. Moreover, even in scenarios where the CBGPS and npCBGPS methods produce the best covariate balance, they may not produce causal inference estimates with the lowest bias.

In sum, current methods for creating balanced data in the continuous treatment setting have relied heavily on weighting procedures, even though it has been well-studied that weighting methods may produce unreliable causal inference estimates due to extreme weights (Zhu et al., 2015; Fong et al., 2018). And although nonparametric methods have been proposed to derive weights in order to attenuate this issue, it has not yet been resolved (Zhu et al., 2015; Fong et al., 2018). Specifically, researchers have shown within simulations that when both treatment and outcome models are misspecified, all weighting propensity score methods fail to obtain accurate ATE estimates (Fong et al., 2018). The current literature indicates alternatives to weighting are desirable in some settings. Currently, fitting a parametric linear model and stratifying subjects based on the scalar value $\hat{\beta}X_i$, that is derived from the estimated model (Imai and Van Dyk, 2004), is the only, and seldom used (Elliott, Zhang, and Small, 2015), stratification method proposed to produce causal estimates for a continuous treatment. Although it is possible to successfully group subjects in this manner, there exists a possibility of subjects with similar values for $\hat{\beta}X_i$ presenting with different covariate distributions. Therefore, this paper seeks to derive more refined methods of stratification that neither utilize weighting nor rely on parametric assumptions in order to produce more reliable causal inference estimates (i.e., ATEs). Specifically, the current paper proposes two novel methodologies that produce causal estimates for continuous treatments: both the generalized propensity score cumulative distribution function (GPS-CDF) and the nonparametric GPS-CDF (npGPS-CDF) methods stratify subjects, based on pretreatment confounders, in order to produce causal estimates.

3.2 Methods

Stratifying subjects with the objective of achieving covariate balance based on pretreatment confounders amounts to creating groups of subjects with similar covariate distributions who received different treatments. With this goal in mind, this section describes two novel stratification methods that seek to refine current methods in order to produce better covariate balance and more accurate ATE estimates. Both proposed methods create subject specific covariate distributions that are used in order to create balancing strata. The first method (GPS-CDF) closely follows the stratification method introduced by Imai and Van Dyk (2004). The second method (npGPS-CDF) does not place any parametric restrictions on the relationship between T and \mathbf{X} .

3.2.1 GPS-CDF - parametric approach

Similar to the method proposed by Imai and Van Dyk (2004), the GPS-CDF approach creates balancing strata from a regression model in order to produce ATE estimates. However, to improve balance, a representation of the distribution of an individual's covariates is proposed, rather than using just the observed instance. This distribution, which is derived through bootstrapping, provides more detailed covariate information for each subject, which can lead to more accurate balancing strata and more accurate causal estimates.

The bootstrapping algorithm of the GPS-CDF method begins by fitting any regression model in the form of equation (3.1), that returns model estimates, in order to predict the continuous treatment (e.g. linear model, generalized linear model). \hat{T}_i , or a subject's predicted treatment, is calculated by multiplying the estimated model coefficients by each subject's covariate profile,

$$\hat{T}_i = \hat{\beta} X_i. \quad (3.11)$$

To further capture the complete covariate profile of a subject, B sets of $\hat{\beta}^*$ coefficients are resampled assuming a multivariate normal distribution,

$$\hat{\beta}_1^*, \dots, \hat{\beta}_b^* \sim MVN(\mu, \Sigma) \quad (3.12)$$

$$\mu = \begin{bmatrix} \hat{\beta}_1 \\ \vdots \\ \hat{\beta}_j \end{bmatrix} \quad \Sigma = \begin{bmatrix} \widehat{Var}(\hat{\beta}_1) & & \mathbf{0} \\ & \ddots & \\ \mathbf{0} & & \widehat{Var}(\hat{\beta}_j) \end{bmatrix}$$

where $\hat{\beta}_1, \dots, \hat{\beta}_j$ and $\widehat{Var}(\hat{\beta}_1), \dots, \widehat{Var}(\hat{\beta}_j)$ are the estimates derived from the original treatment regression model with j covariates for $b = 1, \dots, B$ (an arbitrarily large number, taken in this paper to be 10,000). $\hat{T}_{i,b}^*$ is then calculated for each subject using each of the B sets of sampled $\hat{\beta}^*$ coefficients,

$$\hat{T}_{i,b}^* = \hat{\beta}_b^* X_i \quad (3.13)$$

for $b = 1, \dots, B$. Placing $\hat{T}_{i,b}^*$ values in ascending order, individually for each subject, will create bootstrapped distributions of each subject's predicted treatment values. Each distribution is a separate unimodal probability density function (PDF) that fully encapsulates the covariate profile for each subject. The equation of a particular PDF is given by,

$$PDF_i = \hat{T}_{i,(1)}^*, \dots, \hat{T}_{i,(b)}^* \quad (3.14)$$

for individual i . As subject specific PDFs are derived through variation in $\hat{\beta}$, two subjects with identical covariate profiles will have identical PDFs; this would not be true if variation were introduced in relation to X . Therefore, subjects with similar PDFs will have, on average, similar covariate distributions. Thus, a function that accurately maps the PDF of each subject can subsequently be used to classify subjects into covariate balancing strata.

Unfortunately, mapping directly to this PDF is challenging. A PDF is not a monotone function; the shape of the PDF depends on the covariate distribution for each subject. Instead of mapping a function directly to the PDF, an empirical cumulative density function (eCDF) is estimated for each subject by summing across the subject specific PDF,

$$\hat{F}_{i,b}(t_i) = \hat{P}_{i,b}(\hat{T}_i^* < t_i) = b^{-1} \sum_{l=1}^b I(t_{i,l} \leq t_i) \quad (3.15)$$

where I is the indicator function. The shape of the eCDF for each subject is a strictly non-decreasing function. Furthermore, as the eCDF is a 1-to-1 function of the PDF, subjects with similar eCDFs will have similar PDFs, and resultantly, similar covariate distributions. One proposed equation used to map the eCDF of subject specific predicted treatments (\hat{T}_i^*) generated from this bootstrapping method is a 2-parameter logistic curve given by,

$$F(\hat{T}) = \frac{1}{1 + \exp(-k * (\hat{T} - T_0))} \quad (3.16)$$

where k represents the scale or shape parameter of the logistic curve, and T_0 is the location or midpoint of the sigmoid. Once eCDFs are calculated for each subject, a non-linear least squares (NLS) algorithm (Marquardt, 1963) is used to fit the logistic curve,

$$\min_{k, T_0} \sum_{b=1}^B \left(F(\hat{T}) - \hat{F}_b(t) \right)^2 \text{ for } b = 1, \dots, B. \quad (3.17)$$

The above NLS algorithm iteratively fits values for k and T_0 until the residual distance between the eCDF and fitted logistic curve is minimized. Based on the fitted logistic curve, subjects with similar values for k and T_0 will have similar eCDF vectors and thus similar covariate distributions. Although there are many ways to classify subjects into strata based on two variables, k-Means clustering (KMC) has been shown to provide the highest covariate

similarity within clusters (Tu, Jiao, and Koh, 2013). Additionally, while any number of strata can be formed, following the convention set within binary treatment propensity score analyses, 5 strata are created (Cochran, 1968; Rosenbaum and Rubin, 1984; Zanutto, Lu, and Hornik, 2005; Austin, 2011). Thus using KMC, subjects can be accurately placed into one of five strata with subjects with similar values for both k and T_0 , and subsequently similar covariate distributions.

3.2.2 *npGPS-CDF - nonparametric approach*

Although typical propensity scoring methods fit a treatment model in order to create covariate balance, this is not always beneficial or necessary in the continuous treatment setting. Unlike binary and multiple treatment settings where the GPS is typically estimated through a logistic, multinomial, or probit regression model, the continuous treatment setting typically utilizes a linear model to calculate a predicted treatment. Thus if model assumptions are not met (e.g., deviations from normality of errors), the predicted treatment value from the linear model will be inaccurate, which could lead to poor balance and poor ATE estimates. Instead, as every subject represents a unique treatment group in the continuous treatment setting, creating balance utilizing stratification reduces to grouping subjects with similar covariate distributions, independent of treatment assignment. Thus, a method that stratifies subjects directly using potential confounders without using predicted treatment may improve covariate balance.

Consider the extreme case where one has a non-randomized study with an outcome, a continuous treatment, and two binary potential confounders (sex (male, female), and age (<50 or ≥ 50)). Instead of fitting a model between treatment and the confounders, four strata

can intuitively be created: young males, young females, older males, and older females. Thus, without fitting a treatment model, the two confounders of interest are completely balanced within the four strata. Although this method of stratification may not be possible in applied contexts that include continuous confounders, it does illustrate that balancing strata can be created without fitting a treatment model.

Utilizing the extreme case as a heuristic, the nonparametric extension to the GPS-CDF method, the npGPS-CDF, does not place any parametric restrictions on the relationship between T and \mathbf{X} as it does not involve fitting a regression model for treatment. Instead, an eCDF based solely on the potential confounders of interest is calculated for each subject and used for stratification.

A covariate based distribution can be formed for each subject by sampling B sets of $\hat{\mathbf{T}}^*$ values assuming any continuous distribution centered at 0 (e.g. multivariate standard normal, multivariate T)

$$\hat{\mathbf{T}}_1^*, \dots, \hat{\mathbf{T}}_b^* \sim \mathbf{MVN}(\mathbf{0}_j, \mathbf{I}_j) \quad (3.18)$$

where j is the number of covariates and \mathbf{I}_j is the identity matrix for $b = 1, \dots, B$ (an arbitrarily large number, taken here to be 10,000). These $\hat{\mathbf{T}}^*$ values are then used in order to derive subject specific covariate distributions. Each of the B sets of sampled $\hat{\mathbf{T}}^*$ coefficients is used to calculate $\hat{Z}_{i,b}^*$ for each subject using,

$$\hat{Z}_{i,b}^* = \hat{\mathbf{T}}_b^* \mathbf{X}_i \quad (3.19)$$

for $b = 1, \dots, B$. Placing $\hat{Z}_{i,b}^*$ values in ascending order, individually for each subject, will create separate covariate distributions for each subject. The sampled distribution is given by,

$$PDF_i = \hat{Z}_{i,(1)}^*, \dots, \hat{Z}_{i,(b)}^* \quad (3.20)$$

for the i^{th} individual, which again can be thought of as a unimodal PDF. Once again, eCDFs can be estimated for each subject by summing across subject-specific PDFs,

$$\hat{F}_{i,b}(z_i) = \hat{P}_{i,b}(\hat{Z}_i^* < z_i) = b^{-1} \sum_{l=1}^b I(z_{i,l} \leq z_i) \quad (3.21)$$

where I is the indicator function.

Unlike the parametric GPS-CDF method that requires a location parameter (T_0) to accurately map each eCDF, in the nonparametric setting, all eCDFs are centered at 0. This is a direct byproduct of sampling $\hat{\mathbf{T}}^*$ values from a continuous distribution centered at 0 (e.g. $\mathbf{MVN}(\mathbf{0}_j, \mathbf{I}_j)$ distribution). The proposed 1-parameter logistic curve that can accurately map the eCDF of each subject is given by,

$$F(\hat{Z}) = \frac{1}{1 + \exp(-k * \hat{Z})} \quad (3.22)$$

where k represents the scale or shape parameter of the logistic curve. Similarly, once the eCDF has been calculated for each subject, an NLS algorithm (Marquardt, 1963) can be used to fit this 1-parameter logistic curve,

$$\min_k \sum_{b=1}^B \left(F(\hat{Z}) - \hat{F}_b(z) \right)^2 \text{ for } b = 1, \dots, B. \quad (3.23)$$

Importantly, the npGPS-CDF method results in a single scalar value, k , that fully describes the covariate distribution of each subject. This single scalar balancing score can then be used to stratify subjects into quintiles, such that subjects within a quintile will have similar values of k and thus similar covariate distributions.

3.3 Simulation Study

A simulation study is conducted to determine how the GPS-CDF stratification and npGPS-CDF stratification methods perform under different data scenarios with varying levels of model misspecification. The design of the current simulation follows very closely to several recently published simulations that strive to represent real data (Austin, Grootendorst, and Anderson, 2007; Fong et al., 2018; Greene, 2017). Four data scenarios are considered with one continuous treatment, one binary outcome, and nine baseline covariates, 4 of which are defined as pretreatment confounders of the treatment outcome relationship. A table describing the associations of the baseline covariates with the treatment and the outcome variables is shown in Table 3.1. From the table, it may be noted that x_1 , x_2 , x_4 , and x_5 are simulated to be pretreatment confounders.

The four data scenarios considered within this simulation are very similar to those of Fong et al. (2018) and Greene (2017) in that they vary whether treatment assignment or outcome assignment were correctly specified through inclusion of a non-linear term. Within all four data scenarios, x_1 , x_6 , x_8 , and x_9 are multivariate normally distributed with mean 0, variance 1, and covariances of 0.1, while all other baseline covariates (x_2 , x_3 , x_4 , x_5 , and x_7) were independently drawn from a Bernoulli($p = 0.5$) distribution.

In Scenario 1, both the treatment and outcome models are correctly specified, containing only linear terms. The true treatment and outcome models are given by equations (3.24) and (3.25), respectively:

$$T_i = 0.6(x_{i,1} + x_{i,4} + x_{i,7}) + 0.2(x_{i,2} + x_{i,5} + x_{i,8}) + \varepsilon_i \quad (3.24)$$

$$\log \left[\frac{\Pr(Y_i=1)}{1-\Pr(Y_i=1)} \right] = \alpha + 0.7T_i + 0.6(x_{i,1} + x_{i,2} + x_{i,3}) + 0.2(x_{i,4} + x_{i,5} + x_{i,6}) \quad (3.25)$$

where $\varepsilon_i \stackrel{iid}{\sim} N(0,1)$ is the error term, $\alpha = -5$, and the true ATE is set at 0.7. The binary outcome is simulated by sampling one value from a Bernoulli distribution using the probabilities calculated from equation (3.25) as the probability sampling weights.

Scenario 2 introduces a non-linear term based on a mis-measured variable, $(x_{i,1} + 0.5)^2$, into the treatment assignment model, while the outcome model remained the same as equation (3.25). The misspecified treatment model is given by:

$$T_i = 0.4(x_{i,1} + .5)^2 + 0.6(x_{i,1} + x_{i,4} + x_{i,7}) + 0.2(x_{i,2} + x_{i,5} + x_{i,8}) + \varepsilon_i \quad (3.26)$$

where $\varepsilon_i \stackrel{iid}{\sim} N(0,1)$.

Scenario 3 introduces a non-linear term based on a mis-measured variable, $(x_{i,1} + 0.5)^2$, into the outcome assignment model, while the treatment model remained the same as equation (3.24). The misspecified outcome model is given by:

$$\log \left[\frac{Pr(Y_i = 1)}{1 - Pr(Y_i = 1)} \right] = \alpha + 0.2(x_{i,1} + .5)^2 + 0.7T_i + 0.6(x_{i,1} + x_{i,2} + x_{i,3}) + 0.2(x_{i,4} + x_{i,5} + x_{i,6}) \quad (3.27)$$

where $\alpha = -5$.

Finally in Scenario 4, both the treatment and outcome models are misspecified using the treatment and outcome assignment models detailed in equations (3.26) and (3.27) with $\alpha = -6$.

3.4 Results

For each scenario, 1000 datasets each containing 1000 observations are generated. Five methods from the literature are applied to estimate and compare ATEs: GBM weighting, CBGPS weighting, $\hat{\beta}X_i$ stratification, GPS-CDF stratification, and npGPS-CDF stratification. The propensity model for each method includes all 9 baseline covariates. Outcome analyses to produce ATE estimates utilize survey-weighted generalized linear

models for GBM and CBGPS and conditional logistic regression for $\hat{\beta}X_i$ stratification, GPS-CDF stratification, and npGPS-CDF stratification. Furthermore, to ensure robust ATE estimates, the outcome models additionally adjusted for all first order covariates associated with outcome assignment (Rosenbaum and Rubin, 1984; Hirano and Imbens, 2001; Imai and Van Dyk, 2004).

Figure 3.1 is a graphical representation of covariate balance achieved by each propensity score method under the correctly specified and incorrectly specified treatment assignment models. The plots depict the distribution of F -statistics obtained from regressing T on X , in the overall (unweighted) dataset and using the weights or strata derived from each propensity score method, to give an overall covariate balance summary for the simulated datasets (e.g., as done in Fong et al., 2018). F -statistics were calculated using weighted generalized linear models for GBM and CBGPS. Stratified models, that pooled F -statistics via weighted averages, were used for $\hat{\beta}X_i$ stratification, GPS-CDF stratification, and npGPS-CDF stratification, as is common with stratified analyses (Rosenbaum and Rubin, 1984; Huang et al., 2005; Austin, 2011). Methods that achieved covariate balance have F -statistics closer to zero.

All methods compared achieve better balance, on average, compared to the original (unweighted) data. However, weights derived through GBM produce variable F -statistics, especially within the incorrectly specified treatment model (right plot). The balance achieved by CBGPS weighting is better compared to GBM weighting, but CBGPS is still prone to inadequate covariate balance in both treatment assignment scenarios. Alternatively, $\hat{\beta}X_i$ stratification and GPS-CDF stratification produce smaller F -statistics, which are less sensitive to model misspecification and less susceptible to F -statistic outliers, compared to

both GBM and CBGPS weighting. The balance achieved by npGPS-CDF stratification appears to be poorer overall in the correctly specified model, but better in the incorrectly specified model, compared to GBM and CBGPS weighting. Additionally, npGPS-CDF stratification produces F -statistics without outliers that are less sensitive to model misspecification.

Within each scenario, the five propensity score methods are compared via average bias, mean squared error (MSE), and coverage probability of the estimated ATE. Figure 3.2 depicts the distribution of the ATE for each method under each scenario. The true ATE value is 0.7 and is included as the dotted horizontal line.

In Scenario 1, both GBM and CBGPS produce estimates with increased bias and MSE, as well as decreased coverage probability compared to $\hat{\beta}X_i$ stratification, GPS-CDF stratification, and npGPS-CDF stratification. Additionally, GBM and CBGPS both produce severe ATE outliers, which was expected as the balance produced by these methods was not well controlled. Alternatively, even though the degree of balance produced by npGPS-CDF stratification is poorer than other methods, it performs the best in terms of bias in ATEs and produces the lowest MSE. When the treatment model is misspecified but the outcome model is correct (Figure 3.2, Scenario 2), results are similar to Scenario 1. Again, GBM produces the highest bias and MSE, and the lowest coverage probability. Although CBGPS produces the lowest bias among the five methods, it still produces high MSE and large ATE outliers. Again, npGPS-CDF stratification outperforms $\hat{\beta}X_i$ stratification in terms of bias, MSE, and coverage probability.

When the treatment model is correct, but the outcome model is misspecified (Figure 3.2, Scenario 3), application of GBM results in lower bias and MSE compared to CBGPS.

Consistent with the previous scenarios, CBGPS produces large ATE outliers. GPS-CDF stratification and npGPS-CDF stratification have lower bias and MSE compared to all other methods, with npGPS-CDF stratification producing the most accurate estimates. Finally, Figure 3.2 Scenario 4 further demonstrates that the weighting procedures, GBM and CBGPS, do not perform as well compared to the stratification procedures while the novel npGPS-CDF stratification vastly outperformed all other propensity score methods yet still maintaining a coverage probability equal to 0.95.

For completeness, npCBGPS weighting was additionally conducted under each simulation scenario, but the results obtained were worse than those for CBGPS weighting under all scenarios and are therefore not presented.

3.5 Data Application: Effect of exposure to smoking imagery on smoking initiation in youth

To assess the utility of the novel continuous propensity scoring techniques, GPS-CDF stratification and npGPS-CDF stratification are applied to the Mexican-American Tobacco use in Children (MATCH) study to determine whether exposure to smoking imagery in movies influences smoking initiation among Mexican-American adolescents (Wilkinson et al., 2008).

The MATCH study was a longitudinal population-based cohort study among Mexican-American teens in Houston, Texas, that aimed to measure factors that influence an adolescent's decision to experiment with cigarettes (Spelman et al., 2009). One of the predictors of interest, exposure to smoking imagery in movies (SIM), was measured using a previously validated method in which subjects indicate whether or not they had viewed 50

randomly selected movies from a pool of 250. A scaled continuous variable which quantifies a subject's exposure to SIM was then calculated (Sargent et al., 2008).

Typically, the continuous SIM exposure variable is categorized into four ordinal exposure groups. A previous ordinal propensity score analysis of these data determined that the odds of smoking initiation among teens significantly increased as their level of exposure to smoking imagery quartile increased (stratified ordinal propensity score OR=1.53, 95% CI [1.15, 2.03], $p=0.004$) (Greene, 2017). Although this method of categorization is not inappropriate, categorization of a continuous treatment variable may lead to loss of information during the outcome analysis (Zhu et al., 2015; Fong et al., 2018). The GPS-CDF stratification and npGPS-CDF stratification methods allow one to treat SIM exposure as a continuous covariate to assess its relationship with smoking initiation in adolescents.

Several potential pre-exposure confounders (that are associated with both the level of exposure to smoking imagery in movies and smoking initiation) are included in the current analyses (Table 3.2). Details of all variables included in the propensity models can be found in previous publications (Wilkinson et al., 2009; Greene, 2017). A visual representation of covariate balance is shown in Figure 3.3. The left plot shows the absolute Pearson correlations between each potential confounder (including square terms) and the treatment variable in the original (Unweighted) dataset as well as after utilization of each propensity score method (Zhu et al., 2015; Fong et al., 2018; Austin, 2018b). Zhu et al. (2015) suggest that correlation values less than 0.1 indicate that the confounding effect of the covariate is small. Based on this cutoff, all propensity score methods create better covariate balance compared to the original data.

The right plot of Figure 3.3 presents F -statistics that are calculated by regressing the continuous treatment variable against each potential confounder one at a time. The interquartile range of F -statistics within the figure are (2.52-20.08) for the original data, (0.02-0.54) for GBM, (0.00-0.00) for CBGPS, (0.31-0.79) for $\hat{\beta}X_i$ stratification, (0.39-0.92) for GPS-CDF stratification, and (0.75-2.21) for npGPS-CDF stratification. As all propensity score methods produce small F -statistics, they result in much better balance compared to the original data.

After stratification using GPS-CDF and npGPS-CDF, covariate imbalance within the sample is largely removed. As our simulations show GPS-CDF and npGPS-CDF perform the best in terms of ATE estimation, even in the presence of model misspecification, analyses of the MATCH study are conducted using GPS-CDF and npGPS-CDF stratification. Results from the analyses are shown in Table 3.2. The methods show that the odds of smoking initiation among teens significantly increases as exposure to smoking imagery in movies increases ($OR_{GPS-CDF} = 3.75$, 95% CI [1.50, 9.38], $p = 0.005$; $OR_{npGPS-CDF} = 3.84$, 95% CI [1.52, 9.68], $p = 0.004$).

Results are similar when the analysis are conducted using GBM, CBGPS, and $\hat{\beta}X_i$ stratification with ORs equal to 4.41, 4.46, and 3.34, respectively. All methods attenuated the relationship between exposure to smoking imagery in movies and the odds of smoking initiation among teens compared to the original unweighted data ($OR = 6.57$). Again, outcome analyses are conducted using survey-weighted generalized linear models and conditional logistic regression for weighting methods and stratification methods, respectively. Running the analyses using a dual-core Intel Core i3-3110M with 4 GB RAM,

results were available in 28 seconds, 5 seconds, 3 seconds, 8 seconds, and 7 seconds using GBM, CBGPS, $\hat{\beta}X_i$ stratification, GPS-CDF, and npGPS-CDF, respectively.

3.6 Discussion

Although weighting methods have been proposed to conduct propensity score analyses with continuous treatments (Robins et al., 2000; Zhu et al., 2015; Fong et al., 2018), these methods are not always stable and may produce unreliable estimates. Through simulation, Fong et al. (2018) showed that MLE (Robins et al., 2000) and GBM (Zhu et al., 2015) weights may result in worse covariate balance than had no adjustment been made. These authors further demonstrated that their newly developed weighting methods, CBGPS and npCBGPS, were able to produce better balance than both the MLE and GBM methods. Although the CBGPS methods aim to optimize covariate balance, this increased balance does not always provide more accurate estimates within the outcome analyses. When both treatment and outcome models were misspecified, they found that all weighting propensity score methods failed to obtain accurate ATE estimates. Although, a simplistic method has been proposed that operates without weighting, stratification based on the scalar value $\hat{\beta}X_i$ (Imai and Van Dyk, 2004), its performance has not been sufficiently evaluated through simulation, and is therefore seldom used in practice (Elliott et al., 2015). Based on the inability of weighting procedures to produce both stable and accurate ATE estimates and the underutilization of stratification methods for continuous treatments, we developed new continuous propensity score stratification techniques, GPS-CDF and npGPS-CDF, and investigated their performance against other continuous treatment propensity score methods.

Our simulation study is stronger than some previously conducted (Austin, 2018a; Fong et al., 2018), as it was representative of biomedical data through inclusion of both continuous and binary covariates. The inability of GBM weighting to produce reliable and accurate covariate balance was re-established within our simulation. When treatment assignment was both correctly and incorrectly specified, GBM weighting produced poor balance with patterns similar to the previous simulation (Fong et al., 2018). Unlike Fong et al. (2018), the CBGPS method did not optimize balance for all datasets within our simulation. Since CBGPS methods seek to minimize the weighted correlation between baseline covariates and the treatment, inclusion of binary covariates (as with our simulation) in the propensity score model may cause the CBGPS methods to fail, in terms of producing reliable covariate balance. Of note, we were able to fully replicate the results of Fong et al. (2018) using a simulation consisting of only continuous covariates (not presented), which in our opinion, is not generally applicable to biomedical research questions.

Failure to achieve covariate balance when presented with both continuous and binary pretreatment confounders did not arise in the stratification methods, $\hat{\beta}X_i$ and GPS-CDF. Both methods produced better balance than GBM and CBGPS within our simulation. Interestingly, GPS-CDF stratification produced better covariate balance than npGPS-CDF stratification. Rubin (2006) detailed, within a binary treatment setting, that matching on a regression based scalar value produces better covariate balance than methods that match directly on covariates. As GPS-CDF stratification is implemented using a regression model and npGPS-CDF stratification balances directly on potential confounders, our findings in a continuous treatment setting are analogous to those of Rubin (2006).

GPS-CDF and npGPS-CDF stratification performed well across Scenarios 1-3 for comparisons of the ATE. Similar to the previous simulation (Fong et al., 2018), CBGPS had the lowest average bias among all five methods in Scenario 2 even though the balance achieved by CBGPS weighting was worse than both GPS-CDF and npGPS-CDF, for the incorrectly specified treatment model. This finding may further demonstrate that better covariate balance does not always lead to less biased causal inference estimates (Lee, Lessler, and Stuart, 2010; Stuart, Lee, and Leacy, 2013). The superiority of CBGPS weighting within Scenario 2 did not extend outside of average bias, as CBGPS had MSE three times that of npGPS-CDF stratification. For Scenario 4, which contained misspecification in both the treatment and outcome models, GBM and CBGPS failed to obtain satisfactory ATE estimates, while our newly developed GPS-CDF methods were still robust. The npGPS-CDF method had minimal bias and MSE, and high coverage probability for models with high amounts of misspecification.

The utility and performance of the GPS-CDF methods was further demonstrated on the MATCH study. Our newly derived methods have similar computational burden as current methods. Additionally, GPS-CDF and npGPS-CDF stratification produced better covariate balance compared to the original (unweighted) dataset. Furthermore, our stratification methods showed a stronger association between the odds of smoking initiation and exposure to smoking imagery in movies in Mexican-American adolescents than previous ordinal propensity score analyses (Greene, 2017). Based on these causal findings, public health interventions, including anti-smoking ad campaigns, may be formulated and implemented to help prevent potentially at-risk youth from forming smoking habits.

3.7 Conclusion

This paper details the derivation and application of two propensity scoring methods that remove imbalance due to confounding in observational studies with continuous treatments. Unlike current methods of continuous treatment propensity scoring that utilize weighting, the GPS-CDF and npGPS-CDF methods presented here, create balancing strata that contain subjects with similar covariate distributions. Our simulation study shows that stratification methods may produce less biased causal inference estimates compared to methods that rely on weighting, since extreme weights lead to inaccurate estimates. Furthermore, when applied to the MATCH study, the GPS-CDF and npGPS-CDF methods found a significant association between exposure to smoking imagery in movies and smoking initiation among Mexican-American adolescents.

There are limitations within the current study. Primarily, only 4 data scenarios were considered within the simulation. Thus there exists a possibility that results could differ under different modeling assumptions. However, the simulation scenarios in this paper follow very closely to several recently published simulation studies by experts in the field of causal inference, and are representative of real-world data (e.g., Austin et al., 2007; Fong et al., 2018; Greene, 2017).

In summary, the novel methods presented here allow investigators additional options when conducting continuous treatment propensity scoring in both parametric (GPS-CDF) and nonparametric (npGPS-CDF) frameworks. As with all propensity score methods, investigators should select the method that creates the best covariate balance for their data. Future research should further investigate the use of stratification techniques when

conducting continuous treatment propensity scoring with applications to relevant public health research questions.

| | Strongly Associated with Treatment | Moderately Associated with Treatment | Independent of Treatment |
|---------------------------------------|---------------------------------------|---|-----------------------------|
| Strongly Associated with Outcome | x_1 | x_2 | x_3 |
| Moderately Associated with Outcome | x_4 | x_5 | x_6 |
| Independent of Outcome | x_7 | x_8 | x_9 |

Table 3.1. Association of covariates with treatment and outcome. From the table, it may be noted that x_1 , x_2 , x_4 , and x_5 are simulated to be pretreatment confounders.

| | OR _{GPS-CDF} | 95% CI | p-value | OR _{npGPS-CDF} | 95% CI | p-value |
|-----------------------------|-----------------------|--------------|---------|-------------------------|--------------|---------|
| Movie Exposure | 3.75 | [1.50, 9.38] | 0.005 | 3.84 | [1.52, 9.68] | 0.004 |
| Age | 1.25 | [0.81, 1.94] | 0.307 | 1.34 | [0.86, 2.09] | 0.196 |
| Gender | 0.73 | [0.36, 1.47] | 0.378 | 0.72 | [0.42, 1.25] | 0.248 |
| Born in USA | 0.92 | [0.45, 1.86] | 0.811 | 0.99 | [0.54, 1.82] | 0.969 |
| Level of Acculturation | 0.75 | [0.43, 1.30] | 0.298 | 0.78 | [0.51, 1.18] | 0.241 |
| Parental Education | | | | | | |
| Less than HS | Ref | - | - | Ref | - | - |
| Completed some HS | 0.96 | [0.49, 1.89] | 0.908 | 1.00 | [0.51, 1.96] | 0.996 |
| More than HS | 0.67 | [0.30, 1.52] | 0.341 | 0.69 | [0.35, 1.38] | 0.295 |
| Household Members who Smoke | | | | | | |
| None | Ref | - | - | Ref | - | - |
| One | 0.81 | [0.42, 1.55] | 0.524 | 0.82 | [0.45, 1.50] | 0.521 |
| Two or More | 0.71 | [0.26, 1.97] | 0.510 | 0.74 | [0.28, 1.99] | 0.555 |
| Close Peer who Smokes | 1.40 | [0.68, 2.87] | 0.364 | 1.60 | [0.80, 3.20] | 0.183 |
| Served in Detention | 1.39 | [0.71, 2.70] | 0.334 | 1.39 | [0.82, 2.38] | 0.221 |
| Cognitively Susceptible | 1.09 | [0.53, 2.21] | 0.817 | 1.11 | [0.62, 2.00] | 0.720 |
| Risk Taking Behavior Score | 1.17 | [0.78, 1.77] | 0.453 | 1.24 | [0.87, 1.77] | 0.241 |
| POE Average | 1.41 | [0.68, 2.91] | 0.354 | 1.45 | [0.75, 2.82] | 0.271 |
| TAS | 1.00 | [0.89, 1.12] | 0.960 | 1.03 | [0.84, 1.28] | 0.760 |
| DAA | 1.38 | [1.08, 1.77] | 0.010 | 1.40 | [1.14, 1.71] | 0.001 |
| SD | 1.30 | [1.05, 1.61] | 0.016 | 1.38 | [1.12, 1.69] | 0.003 |
| SSS | 0.92 | [0.76, 1.12] | 0.424 | 0.94 | [0.72, 1.23] | 0.664 |

Table 3.2. Model estimates after GPS-CDF and npGPS-CDF stratification from the MATCH study. Note: HS = High School, POE = Positive outcome expectation, TAS = Thrill and adventure seeking, DAA = Drug and alcohol, SD = Social disinhibition score, SSS = Subjective Social Status.

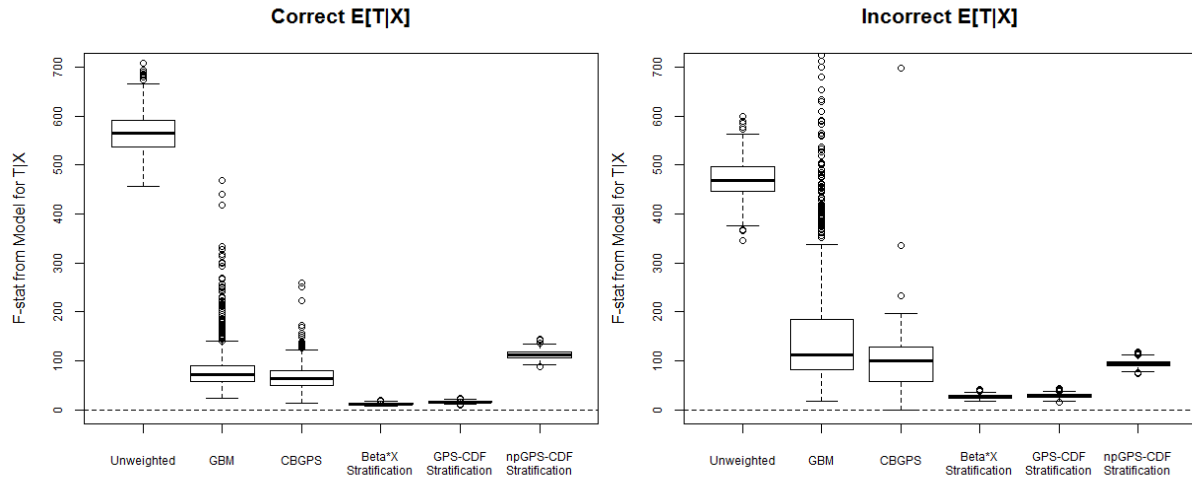


Figure 3.1. Graphical representation of the covariate balance achieved by each propensity score method under the correctly specified and incorrectly specified treatment assignment models. F -statistics obtained from regressing T on X .

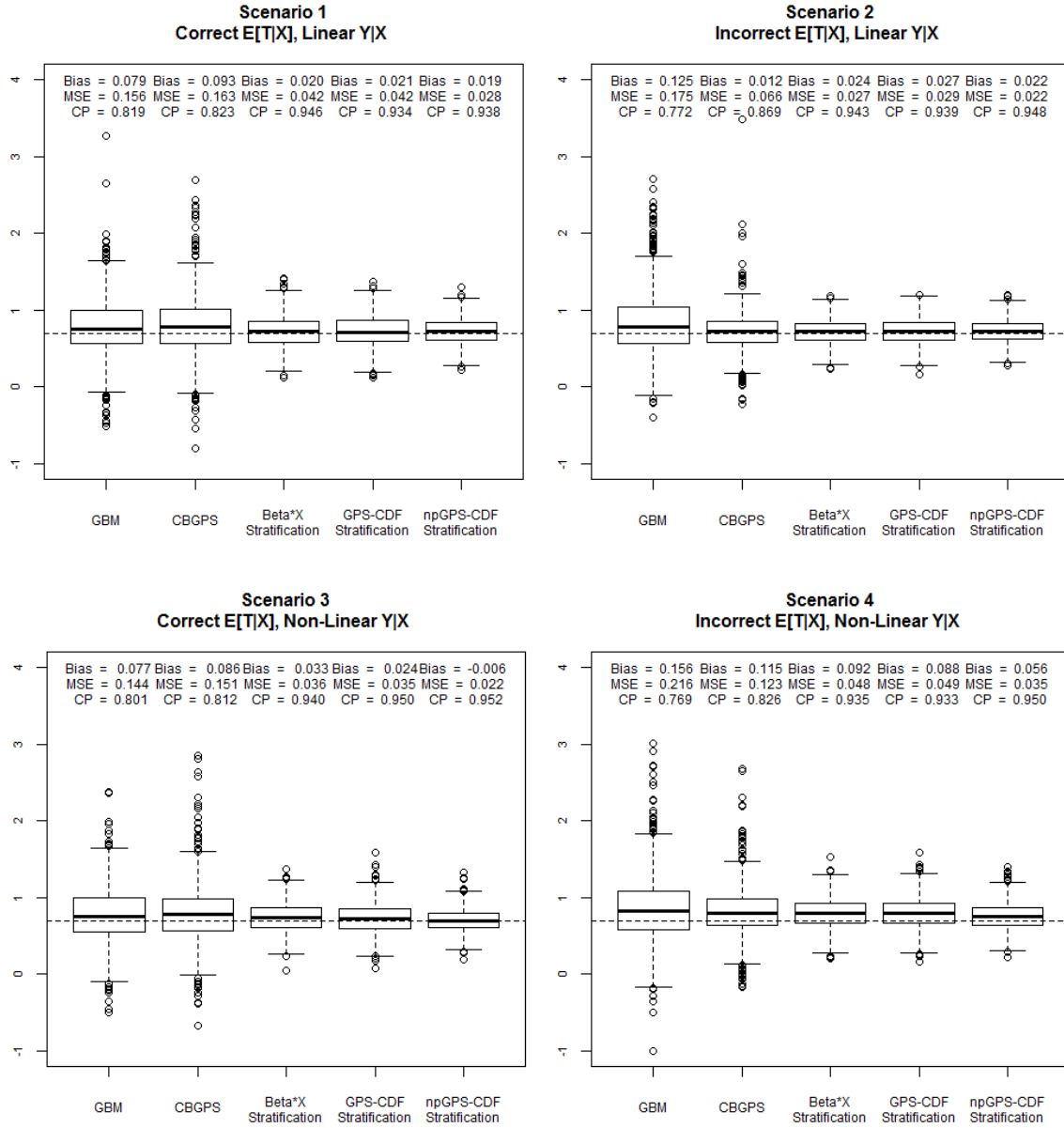


Figure 3.2. Distribution of the ATE for each method under each scenario. The true ATE value is included as the dotted horizontal line.

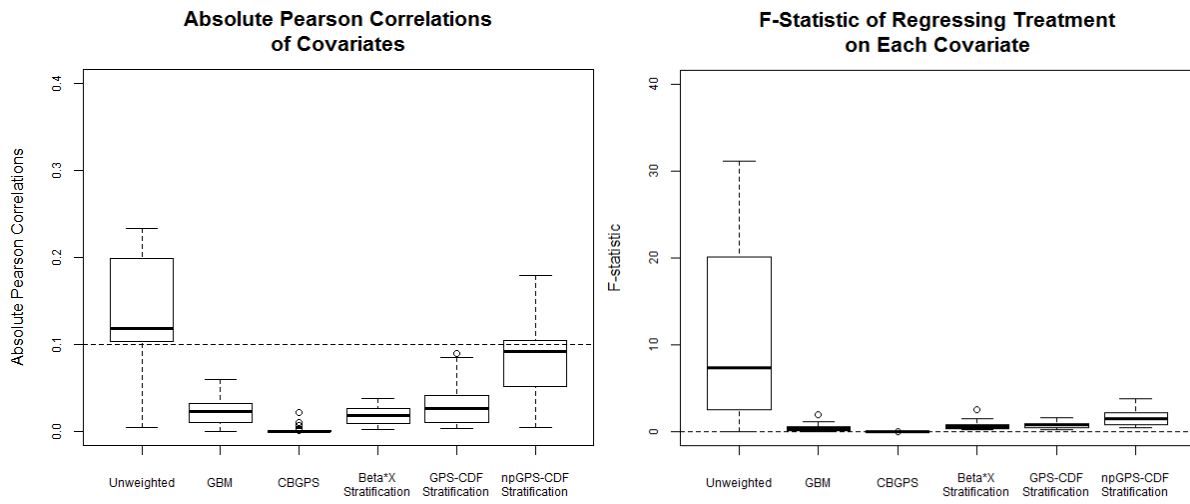


Figure 3.3. Graphical representation of the covariate balance achieved by each propensity score method within the MATCH study. The left plot presents the absolute Pearson correlation between treatment and each potential confounder (including square terms). The right plot presents F -statistics obtained from regressing T on each potential confounder one at a time.

References

- Austin, P. C. (2011). An Introduction to Propensity Score Methods for Reducing the Effects of Confounding in Observational Studies. *Multivariate behavioral research* **46**, 399-424.
- Austin, P. C. (2018a). Assessing covariate balance when using the generalized propensity score with quantitative or continuous exposures. *Statistical methods in medical research*, 0962280218756159.
- Austin, P. C. (2018b). Assessing the performance of the generalized propensity score for estimating the effect of quantitative or continuous exposures on binary outcomes. *Stat Med* **37**, 1874-1894.
- Austin, P. C., Grootendorst, P., and Anderson, G. M. (2007). A comparison of the ability of different propensity score models to balance measured variables between treated and untreated subjects: a Monte Carlo study. *Stat Med* **26**, 734-753.
- Bia, M., Flores, C. A., Flores-Lagunes, A., and Mattei, A. (2014). A Stata package for the application of semiparametric estimators of dose–response functions. *Stata Journal* **14**, 580-604.
- Boyd, C. L., Epstein, L., and Martin, A. D. (2010). Untangling the causal effects of sex on judging. *American journal of political science* **54**, 389-411.
- Chertow, G. M., Normand, S.-L. T., and McNeil, B. J. (2004). “Renalism”: inappropriately low rates of coronary angiography in elderly individuals with renal insufficiency. *Journal of the American Society of Nephrology* **15**, 2462-2468.
- Cochran, W. G. (1968). The effectiveness of adjustment by subclassification in removing bias in observational studies. *Biometrics*, 295-313.

- Davidson, M. B., Hix, J. K., Vidt, D. G., and Brotman, D. J. (2006). Association of impaired diurnal blood pressure variation with a subsequent decline in glomerular filtration rate. *Archives of internal medicine* **166**, 846-852.
- Donohue, J. J., and Ho, D. E. (2007). The Impact of Damage Caps on Malpractice Claims: Randomization Inference with Difference-in-Differences. *Journal of Empirical Legal Studies* **4**, 69-102.
- Elliott, M. R., Zhang, N., and Small, D. S. (2015). Application of Propensity Scores to a Continuous Exposure: Effect of Lead Exposure in Early Childhood on Reading and Mathematics Scores. *Observational Studies* **1**, 30-55.
- Flores-Lagunes, A., Gonzalez, A., and Neumann, T. (2007). Estimating the effects of length of exposure to a training program: the case of Job Corps.
- Fong, C., Hazlett, C., and Imai, K. (2018). Covariate balancing propensity score for a continuous treatment: Application to the efficacy of political advertisements. *The Annals of Applied Statistics* **12**, 156-177.
- Greene, T.J. (2017). Utilizing Propensity Score Methods for Ordinal Treatments and Prehospital Trauma Studies. *Texas Medical Center Dissertations (via ProQuest)*.
- Harder, V. S., Stuart, E. A., and Anthony, J. C. (2008). Adolescent cannabis problems and young adult depression: male-female stratified propensity score analyses. *American journal of epidemiology* **168**, 592-601.
- Hirano, K., and Imbens, G. W. (2001). Estimation of causal effects using propensity score weighting: An application to data on right heart catheterization. *Health Services and Outcomes research methodology* **2**, 259-278.

- Hirano, K., and Imbens, G. W. (2004). The propensity score with continuous treatments. *Applied Bayesian modeling and causal inference from incomplete-data perspectives* **226164**, 73-84.
- Huang, I.-C., Frangakis, C., Dominici, F., Diette, G. B., and Wu, A. W. (2005). Application of a propensity score approach for risk adjustment in profiling multiple physician groups on asthma care. *Health services research* **40**, 253-278.
- Imai, K., and Ratkovic, M. (2014). Covariate balancing propensity score. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **76**, 243-263.
- Imai, K., and Van Dyk, D. A. (2004). Causal inference with general treatment regimes: Generalizing the propensity score. *Journal of the American Statistical Association* **99**, 854-866.
- Imbens, G. W. (2000). The role of the propensity score in estimating dose-response functions. *Biometrika* **87**, 706-710.
- Joffe, M. M., and Rosenbaum, P. R. (1999). Invited commentary: propensity scores. *American journal of epidemiology* **150**, 327-333.
- Lee, B. K., Lessler, J., and Stuart, E. A. (2010). Improving propensity score weighting using machine learning. *Stat Med* **29**, 337-346.
- Marquardt, D. W. (1963). An algorithm for least-squares estimation of nonlinear parameters. *Journal of the society for Industrial and Applied Mathematics* **11**, 431-441.
- McCaffrey, D. F., Griffin, B. A., Almirall, D., Slaughter, M. E., Ramchand, R., and Burgette, L. F. (2013). A tutorial on propensity score estimation for multiple treatments using generalized boosted models. *Stat Med* **32**, 3388-3414.

- Nielsen, R. A., Findley, M. G., Davis, Z. S., Candland, T., and Nielson, D. L. (2011). Foreign aid shocks as a cause of violent armed conflict. *American journal of political science* **55**, 219-232.
- Ridgeway, G., McCaffrey, D., Morral, A., Griffin, B., and Burgette, L. (2016). Toolkit for Weighting and Analysis of Nonequivalent Groups (Version 1.4-9.5).
- Robins, J. M., Hernan, M. A., and Brumback, B. (2000). Marginal structural models and causal inference in epidemiology. LWW.
- Rosenbaum, P. R., and Rubin, D. B. (1983). The Central Role of the Propensity Score in Observational Studies for Causal Effects. *Biometrika* **70**, 41-55.
- Rosenbaum, P. R., and Rubin, D. B. (1984). Reducing bias in observational studies using subclassification on the propensity score. *Journal of the American Statistical Association* **79**, 516-524.
- Rosenbaum, P. R., and Rubin, D. B. (1985). Constructing a control group using multivariate matched sampling methods that incorporate the propensity score. *The American Statistician* **39**, 33-38.
- Royston, P., Altman, D. G., and Sauerbrei, W. (2006). Dichotomizing continuous predictors in multiple regression: a bad idea. *Stat Med* **25**, 127-141.
- Rubin, D. B. (2006). *Matched Sampling for Causal Effects*: Cambridge University Press.
- Sargent, J. D., Worth, K. A., Beach, M., Gerrard, M., and Heatherton, T. F. (2008). Population-based assessment of exposure to risk behaviors in motion pictures. *Communication Methods and Measures* **2**, 134-151.

- Schuler, M. S., Chu, W., and Coffman, D. (2016). Propensity score weighting for a continuous exposure with multilevel data. *Health Services and Outcomes research methodology* **16**, 271-292.
- Spelman, A. R., Spitz, M. R., Kelder, S. H., *et al.* (2009). Cognitive susceptibility to smoking: Two paths to experimenting among Mexican origin youth. *Cancer Epidemiol Biomarkers Prev* **18**, 3459-3467.
- Stuart, E. A., Lee, B. K., and Leacy, F. P. (2013). Prognostic score-based balance measures for propensity score methods in comparative effectiveness research. *Journal of clinical epidemiology* **66**, S84-S90.e81.
- Tu, C., Jiao, S., and Koh, W. Y. (2013). Comparison of clustering algorithms on generalized propensity score in observational studies: a simulation study. *Journal of Statistical Computation and Simulation* **83**, 2206-2218.
- Wilkinson, A. V., Spitz, M. R., Prokhorov, A. V., Bondy, M. L., Shete, S., and Sargent, J. D. (2009). Exposure to smoking imagery in the movies and experimenting with cigarettes among Mexican heritage youth. *Cancer Epidemiol Biomarkers Prev* **18**, 3435-3443.
- Wilkinson, A. V., Waters, A. J., Vasudevan, V., Bondy, M. L., Prokhorov, A. V., and Spitz, M. R. (2008). Correlates of susceptibility to smoking among Mexican origin youth residing in Houston, Texas: a cross-sectional analysis. *BMC Public Health* **8**, 337.
- Zanutto, E., Lu, B., and Hornik, R. (2005). Using propensity score subclassification for multiple treatment doses to evaluate a national antidrug media campaign. *Journal of Educational and Behavioral Statistics* **30**, 59-73.

Zhu, Y., Coffman, D. L., and Ghosh, D. (2015). A boosting algorithm for estimating generalized propensity scores with continuous treatments. *Journal of causal inference* **3**, 25-40.

CHAPTER IV

Generalized Propensity Score Cumulative Distribution Function (GPS-CDF) R Package

The Comprehensive R Archive Network (Accepted)

Derek W. Brown¹, Thomas J. Greene², Stacia M. DeSantis¹

Author Affiliations

1. Department of Biostatistics and Data Science, University of Texas Health School of Public Health
2. GlaxoSmithKline, Division of Biostatistics

Corresponding Author:

Derek Brown, PhD(c), MS
Department of Biostatistics and Data Science
University of Texas Health Science Center at Houston
1200 Pressler Street, Houston, TX 77030
[derek.brown@uth.tmc.edu]

4.1 Introduction

A freely downloadable R software package (R Foundation, Vienna, Austria) was created to facilitate the distribution of the newly created GPS-CDF propensity score methods. The *GPSCDF* R package (Brown et al., 2019) includes both the ordinal (Greene, 2017) and multinomial (as detailed in Chapter II) GPS-CDF propensity score methods. This package allows researchers to input a GPS vector of length >2 , and outputs \tilde{a} , the single scalar balancing score that dictates the shape of the CDF. Additional functionality of the package allows researchers to automatically match (both optimal and greedy matching) and stratify subjects based on \tilde{a} . The R documentation for the *GPSCDF* R package (Brown et al., 2019) is given below in Section 4.2, and an illustrative data example is detailed in Section 4.3. The package code used to implement the GPS-CDF method is given in Appendix B.

4.2 R Documentation

GPSCDF {GPSCDF}

R Documentation

Generalized Propensity Score Cumulative Distribution Function (GPS-CDF)

Description

GPSCDF takes in a generalized propensity score (GPS) object with length >2 and returns the GPS-CDF balancing score.

Usage

```
GPSCDF(pscores = NULL, data = NULL, trt = NULL, stratify = FALSE,  
       nstrat = 5, optimal = FALSE, greedy = FALSE, ordinal = FALSE,  
       multinomial = FALSE, caliper = NULL)
```

Details

The *GPSCDF* method is used to conduct propensity score matching and stratification for both ordinal and multinomial treatments. The method directly maps any GPS vector (with length >2) to a single scalar value that can be used to produce either average treatment effect (ATE) or average treatment effect among the treated (ATT) estimates. For the K multinomial treatments setting, the balance achieved from each $K!$ ordering of the GPS should be assessed to find the optimal ordering of the GPS vector (see Examples for more details).

4.2.1 User Defined Inputs

Arguments

| | |
|--------------------------|--|
| <code>pscores</code> | The object containing the treatment ordered generalized propensity scores for each subject. |
| <code>data</code> | An optional data frame to attach the calculated balancing score. The data frame will also be used in stratification and matching. |
| <code>trt</code> | An optional object containing the treatment variable. |
| <code>stratify</code> | Option to produce strata based on the power parameter (<code>ppar</code>). Default is <code>FALSE</code> . |
| <code>nstrat</code> | An optional parameter for the number of strata to be created when <code>stratify</code> is set to <code>TRUE</code> . Default is 5 strata. |
| <code>optimal</code> | Option to perform optimal matching of subjects based on the power parameter (<code>ppar</code>). Default is <code>FALSE</code> . |
| <code>greedy</code> | Option to perform greedy matching of subjects based on the power parameter (<code>ppar</code>). Default is <code>FALSE</code> . |
| <code>ordinal</code> | Specifies ordinal treatment groups for matching. Subjects are matched based on the ratio of the squared difference of power parameters for two subjects, <code>ppar_i</code> and <code>ppar_j</code> , in the numerator and the squared difference in observed treatment received, <code>trt_i</code> and <code>trt_j</code> , in the denominator: $(ppar_i - ppar_j)^2 / (trt_i - trt_j)^2$. Default is <code>FALSE</code> . |
| <code>multinomial</code> | Specifies multinomial treatment groups for matching. Subjects are matched based on the absolute difference of power parameters for two subjects, <code>ppar_i</code> and <code>ppar_j</code> , who received different treatments: $ ppar_i - ppar_j $. Default is <code>FALSE</code> . |
| <code>caliper</code> | An optional parameter for the caliper value used when performing greedy matching. Used when <code>greedy</code> is set to <code>TRUE</code> . Default is <code>.25*sd(ppar)</code> . |

4.2.2 GPS-CDF Package Outputs

Value

| | |
|---------------------------|--|
| <code>ppar</code> | The power parameter scalar balancing score to be used in outcome analyses through stratification or matching. |
| <code>data</code> | The user defined dataset with power parameter (<code>ppar</code>), strata, and/or optimal matching variables attached. |
| <code>nstrat</code> | The number of strata used for stratification. |
| <code>strata</code> | The strata produced based on the calculated power parameter (<code>ppar</code>). |
| <code>optmatch</code> | The optimal matches produced based on the calculated power parameter (<code>ppar</code>). |
| <code>optdistance</code> | The average absolute total distance of power parameters (<code>ppars</code>) for optimally matched pairs. |
| <code>caliper</code> | The caliper value used for greedy matching. |
| <code>grddata</code> | The user defined dataset with greedy matching variable attached. |
| <code>grdmatch</code> | The greedy matches produced based on the calculated power parameter (<code>ppar</code>). |
| <code>grdydistance</code> | The average absolute total distance of power parameters (<code>ppars</code>) for greedy matched pairs. |

4.2.3 GPS-CDF Package Examples

Examples

```
### Example: Create data example
N<- 100

set.seed(18201) # make sure data is repeatable
Sigma <- matrix(.2,4,4)
diag(Sigma) <- 1
data<-matrix(0, nrow=N, ncol=6,dimnames=list(c(1:N),
      c("Y","trt",paste("X",c(1:4),sep=""))))
data[,3:6]<-matrix(MASS::mvrnorm(N, mu=rep(0, 4), Sigma,
      empirical = FALSE) , nrow=N, ncol = 4)

dat<-as.data.frame(data)

#Create Treatment Variable
tlogits<-matrix(0,nrow=N,ncol=2)
tprobs<-matrix(0,nrow=N,ncol=3)

alphas<-c(0.25, 0.3)
strongbetas<-c(0.7, 0.4)
modbetas<-c(0.2, 0.3)

for(j in 1:2){
  tlogits[,j]<- alphas[j] + strongbetas[j]*dat$X1 + strongbetas[j]*dat$X2+
    modbetas[j]*dat$X3 + modbetas[j]*dat$X4
}

for(j in 1:2){
  tprobs[,j]<- exp(tlogits[,j])/(1 + exp(tlogits[,1]) + exp(tlogits[,2]))
  tprobs[,3]<- 1/(1 + exp(tlogits[,1]) + exp(tlogits[,2]))
}

for(j in 1:N){
  ylogits[j,1]<- -1.1 + 0.7*data[j,2] + 0.6*dat$X1[j] + 0.6*dat$X2[j] +
    0.4*dat$X3[j] + 0.4*dat$X4[j]

  yprobs[j,2]<- 1/(1+exp(-ylogits[j,1]))

  yprobs[j,1]<- 1-yprobs[j,2]
}

set.seed(91187)
for(j in 1:N){
  data[j,1]<-sample(c(0,1),size=1,prob=yprobs[j,])
}

dat<-as.data.frame(data)
```

```

### Example: Using GPSCDF

#Create the generalized propensity score (GPS) vector using any parametric or
#nonparametric model

glm<- nnet::multinom(as.factor(trt)~ X1+ X2+ X3+ X4, data=dat)
probab<- round(predict(glm, newdata=dat, type="probs"),digits=8)
gps<-cbind(probab[,1],probab[,2],1-probab[,1]-probab[,2])

#Create scalar balancing power parameter
fit<-GPSCDF(pscores=gps)

## Not run:
  fit$ppar

## End(Not run)

#Attach scalar balancing power parameter to user defined data set
fit2<-GPSCDF(pscores=gps, data=dat)

## Not run:
  fit2$ppar
  fit2$data

## End(Not run)

### Example: Ordinal Treatment

#Stratification
fit3<-GPSCDF(pscores=gps, data=dat, stratify=TRUE, nstrat=5)

## Not run:
  fit3$ppar
  fit3$data
  fit3$nstrat
  fit3$strata

  library(survival)
  modell<-survival::clogit(Y~as.factor(trt)+X1+X2+X3+X4+strata(strata),
                        data=fit3$data)

  summary(modell)

## End(Not run)

```

```

#Optimal Matching
fit4<- GPSCDF(pscores=gps, data=dat, trt=dat$trt, optimal=TRUE, ordinal=TRUE)

## Not run:
fit4$ppar
fit4$data
fit4$optmatch
fit4$optdistance

library(survival)
model2<-survival::clogit(Y~as.factor(trt)+X1+X2+X3+X4+strata(optmatch),
                        data=fit4$data)

summary(model2)

## End(Not run)

#Greedy Matching
fit5<- GPSCDF(pscores=gps, data=dat, trt=dat$trt, greedy=TRUE, ordinal=TRUE)

## Not run:
fit5$ppar
fit5$data
fit5$caliper
fit5$grddata
fit5$grdmatch
fit5$grdydistance

library(survival)
model3<-survival::clogit(Y~as.factor(trt)+X1+X2+X3+X4+strata(grdmatch),
                        data=fit5$grddata)

summary(model3)

## End(Not run)

### Example: Multinomial Treatment

#Create all K! orderings of the GPS vector
gps1<-cbind(gps[,1],gps[,2],gps[,3])
gps2<-cbind(gps[,1],gps[,3],gps[,2])
gps3<-cbind(gps[,2],gps[,1],gps[,3])
gps4<-cbind(gps[,2],gps[,3],gps[,1])
gps5<-cbind(gps[,3],gps[,1],gps[,2])
gps6<-cbind(gps[,3],gps[,2],gps[,1])

gpsarry<-array(c(gps1, gps2, gps3, gps4, gps5, gps6), dim=c(N,3,6))

#Create scalar balancing power parameters for each ordering of the GPS vector
fit6<- matrix(0,nrow=N,ncol=6,dimnames=list(c(1:N),c("ppar1","ppar2","ppar3",
"ppar4","ppar5","ppar6"))))

## Not run:
for(i in 1:6){
  fit6[,i]<-GPSCDF(pscores=gpsarry[,i])$ppar
}

fit6

#Perform analyses (similar to ordinal examples) using each K! ordering of the
#GPS vector. Select ordering which achieves optimal covariate balance
#(i.e. minimal standardized mean difference).

## End(Not run)

```

4.3 Illustrative Data Example

The methodology presented in Chapter II for estimating multiple treatments propensity scoring is available in the *GPSCDF* package in R (Brown et al., 2019). Below is an illustration of how to implement the package in practice in order to estimate ATEs using the Cerner Health Facts database (as detailed in Chapter II). This data example aimed to analyze the relationship between vasopressor choice and mortality in patients with non-traumatic aneurysmal subarachnoid hemorrhage (SAH). We begin by loading the required packages and reading in the SAH data.

```
> library(GPSCDF)
> library(twang)
> library(tableone)

> ehr<- read.csv(file="EHR_pre_treat_merged_counts.csv",
+               header=TRUE, sep=",")

> dim(ehr)

[1] 2417 275
```

The EHR dataset contains records for 2,417 patients with complete data for all 273 pretreatment variables (demographics and medication variables). Variables were selected for inclusion into the GPS vector using L1-penalized generalized linear models (GLM Lasso) (Mee Young and Hastie, 2007). This procedure was conducted in order to identify the significant confounders associated with the choice of vasopressor treatments. After utilizing GLM Lasso, 170 variables were selected and entered into a GBM model in order to derive

subject specific GPS vectors. After the GPS vector is calculated for each subject, each ordering of the GPS vector can be created.

```
> length(x_1)

[1] 170

> xt1m<- paste(x_1,collapse="+" )

> tmodpaste<- as.formula(paste("as.factor(VASPRESOR_Class)~",
+                               xt1m, sep = ""))

> #GBM model
> ps1<- mnps(tmodpaste, data=ehr, estimand="ATE",verbose =
+            FALSE, stop.method = c("es.mean"), n.trees=6000)

> #Order: 1-2-3
> pscores1<-cbind(ps1$psList$`1`$ps$es.mean.ATE,
+                 ps1$psList$`2`$ps$es.mean.ATE,
+                 ps1$psList$`3`$ps$es.mean.ATE)
> pscoresnorm1<- pscores1/rowSums(pscores1)

> #Order: 1-3-2
> pscores2<-cbind(ps1$psList$`1`$ps$es.mean.ATE,
+                 ps1$psList$`3`$ps$es.mean.ATE,
+                 ps1$psList$`2`$ps$es.mean.ATE)
> pscoresnorm2<- pscores2/rowSums(pscores2)

> #Order: 2-1-3
> pscores3<-cbind(ps1$psList$`2`$ps$es.mean.ATE,
+                 ps1$psList$`1`$ps$es.mean.ATE,
+                 ps1$psList$`3`$ps$es.mean.ATE)
> pscoresnorm2<- pscores2/rowSums(pscores2)

> #Order: 2-3-1
> pscores4<-cbind(ps1$psList$`2`$ps$es.mean.ATE,
+                 ps1$psList$`3`$ps$es.mean.ATE,
+                 ps1$psList$`1`$ps$es.mean.ATE)
> pscoresnorm4<- pscores4/rowSums(pscores4)

> #Order: 3-1-2
> pscores5<-cbind(ps1$psList$`3`$ps$es.mean.ATE,
+                 ps1$psList$`1`$ps$es.mean.ATE,
+                 ps1$psList$`2`$ps$es.mean.ATE)
```

```

> pscor norms5<- pscor es5/rowSums(pscor es5)

> #Order: 3-2-1
> pscor es6<-cbind(psl$psList$`3`$ps$es.mean.ATE,
+                  psl$psList$`2`$ps$es.mean.ATE,
+                  psl$psList$`1`$ps$es.mean.ATE)
> pscor norms6<- pscor es6/rowSums(pscor es6)

```

4.3.1 Obtaining the Scalar Balancing Score

The function GPSCDF is called by the following, with the arguments described below:

```

> gp scdf.ehr <- GPSCDF(pscor es = pscor norms1, data = ehr,
+                       trt = ehr$VASPRESOR_Class, greedy = TRUE,
+                       stratify = TRUE, multinomial = TRUE)

```

The main argument of the GPSCDF function is `pscores` which indicates the ordering of the GPS vector to be used to create \tilde{a} , the single scalar balancing score that dictates the shape of the CDF. Other key arguments include `data`, which indicates the name of the dataset to attach \tilde{a} ; `stratify`, which instructs the function to create strata based on the calculated \tilde{a} ; `optimal` and `greedy`, which produce either optimal or greedy matches based on \tilde{a} , respectively; `ordinal` and `multinomial`, which indicate if matches are selected from either ordinal or multinomial treatments, respectively. The below procedure calculates \tilde{a} and additionally creates greedy matches and strata based on the GPSCDF balancing score obtained from the initial ordering of the GPS vector. The multinomial option was specified to ensure matches are based on the absolute difference of \tilde{a} .

A key component of using the GPS-CDF method for multinomial treatments propensity scoring is selecting the ordering of the GPS vector that creates the best balance in the data. We do this by selecting the ordering that minimizes the standardized mean difference (SMD) within matches.

```
> SMDdat <- gpscdf.ehr$grddata
> SMDdat$trtc <- 0

> for(i in 1:(dim(SMDdat)[1]/2)){
>   matchpair<-SMDdat[which(SMDdat$grdmatch==i),]
>   matchpair<-matchpair[order(matchpair$VASPRESOR_Class),]
>   trtc<-paste(matchpair[1,4],matchpair[2,4],sep=" ")
>
>   SMDdat$trtc[i]<-trtc
>   SMDdat$trtc[i+dim(SMDdat)[1]/2]<-trtc
> }
> mvars=paste(x_1, sep=" " )
> fvars=paste(x_1[c(-1)])
> MatchTab12<- CreateTableOne(vars=mvars, factorVars = fvars,
+   strata=c("VASPRESOR_Class"),
+   data=SMDdat[which(SMDdat$trtc==12),], test=F)
> SMDMatch12<-abs(ExtractSmd(MatchTab12))
> SMD12<-mean(SMDMatch12)

> MatchTab13<- CreateTableOne(vars=mvars, factorVars = fvars,
+   strata=c("VASPRESOR_Class"),
+   data=SMDdat[which(SMDdat$trtc==13),], test=F)
> SMDMatch13<-abs(ExtractSmd(MatchTab13))
> SMD13<-mean(SMDMatch13)

> MatchTab23<- CreateTableOne(vars=mvars, factorVars = fvars,
+   strata=c("VASPRESOR_Class"),
+   data=SMDdat[which(SMDdat$trtc==23),], test=F)
> SMDMatch23<-abs(ExtractSmd(MatchTab23))
> SMD23<-mean(SMDMatch23)

> averageSMD<-(SMD12+SMD13+SMD23)/3
> averageSMD

[1] 0.1115303
```

As shown above, the overall SMD from the initial ordering of the GPS is 0.1115303. This SMD procedure is further applied to each additional ordering of the GPS vector to determine the ordering which creates matches that minimizes the SMD among all covariates.

```
> averageSMD
```

| | 1-2-3 | 1-3-2 | 2-1-3 | 2-3-1 | 3-1-2 | 3-2-1 |
|---|-----------|-----------|------------|-----------|----------|-----------|
| 1 | 0.1115303 | 0.1132709 | 0.09815014 | 0.1151472 | 0.088769 | 0.1201286 |

Based on these results, the fifth ordering of the GPS vector (i.e. 3-1-2) produced greedy matches which minimizes the SMD and is therefore retained for the outcome analysis. A similar procedure can be conducted to select the ordering which minimizes SMD among strata.

```
> strataSMD
```

| | 1-2-3 | 1-3-2 | 2-1-3 | 2-3-1 | 3-1-2 | 3-2-1 |
|---|-----------|-----------|-----------|-----------|-----------|-----------|
| 1 | 0.1518437 | 0.1626902 | 0.1550208 | 0.1631293 | 0.1531595 | 0.1568014 |

Based on these results, the initial ordering of the GPS vector (i.e. 1-2-3) produced strata with the minimum SMD and should be retained for outcome analyses.

4.3.2 Outcome Analyses

The orderings of the GPS vector that produce matches and strata with the best covariate balance are retained for outcome analyses. As the outcome of interest within the EHR dataset is mortality (i.e. a binary variable), conditional logistic regression models are used in order to obtain ATE estimates. The below procedure is used to obtain the effect

estimates from the greedy matched data based on the selected ordering of the GPS vector. A similar procedure may be conducted in order to obtain ATE estimates using the selected stratified data.

```
> dat <- gpscdf.ehr$grddata

> #Phenylephrine As reference
> dat$VASPRESOR_Class_2 <- relevel(as.factor(dat$VASPRESOR_Class),
+                                ref = "2")

> model1<- clogit(Mortality~ as.factor(VASPRESOR_Class_2)+
+                  AGE_IN_YEARS + RACE + MARITAL_STATUS + GENDER +
+                  strata(grdmatch), data=dat)

> coefs1<-summary(model1)$coefficients[,1]
> secoef1<-summary(model1)$coefficients[,3]

> #Norepinephrine As reference
> dat$VASPRESOR_Class_3 <- relevel(as.factor(dat$VASPRESOR_Class),
+                                ref = "3")

> model2<- clogit(Mortality~ as.factor(VASPRESOR_Class_3)+
+                  AGE_IN_YEARS + RACE + MARITAL_STATUS + GENDER +
+                  strata(grdmatch), data=dat)

> coefs2<-summary(model2)$coefficients[1,1]
> secoef2<-summary(model2)$coefficients[1,3]

> #Dopamine vs Phenylephrine:
> OR12<-exp(coefs1[1])
> LCL12<-exp(coefs1[1]-qnorm(.975)*secoef1[1])
> UCL12<-exp(coefs1[1]+qnorm(.975)*secoef1[1])

> OR12
as.factor(VASPRESOR_Class_2)1
1.52918
> LCL12
as.factor(VASPRESOR_Class_2)1
1.114712
> UCL12
as.factor(VASPRESOR_Class_2)1
2.097755
```

```

> #Norepinephrine vs Phenylephrine:
> OR32<-exp(coefs1[2])
> LCL32<-exp(coefs1[2]-qnorm(.975)*secoef1[2])
> UCL32<-exp(coefs1[2]+qnorm(.975)*secoef1[2])

> OR32
as.factor(VASPRESOR_Class_2)3
1.408009

> LCL32
as.factor(VASPRESOR_Class_2)3
0.9984772

> UCL32
as.factor(VASPRESOR_Class_2)3
1.985512

> #Dopamine vs Norepinephrine:
> OR13<-exp(coefs2)
> LCL13<-exp(coefs2-qnorm(.975)*secoef2)
> UCL13<-exp(coefs2+qnorm(.975)*secoef2)

> OR13
[1] 1.086059
> LCL13
[1] 0.7911163
> UCL13
[1] 1.490961

```

The ATE estimates obtain above are identical to those detailed in Chapter II (Table 2.2). These results show that phenylephrine is superior to dopamine in relation to mortality in patients with non-traumatic SAH, but the comparison between phenylephrine and norepinephrine remains unclear.

4.4 Conclusion

The package detailed here may be freely downloaded and easily adapted for various research projects that present with either ordinal or multinomial treatments. We have

presented an example that shows the ease of which GPS-CDF can be applied to a large data set, and outline key considerations when using the GPSCDF package to estimate either the average treatment effect, or the average treatment effect among the treated.

References

- Brown, D.W., Greene, T.J., and DeSantis, S.M. (2019). GPSCDF: Generalized Propensity Score Cumulative Distribution Function. *R Package*, version 0.1.1. <https://CRAN.R-project.org/package=GPSCDF>.
- Greene, T.J. (2017). Utilizing Propensity Score Methods for Ordinal Treatments and Prehospital Trauma Studies. *Texas Medical Center Dissertations (via ProQuest)*.
- Mee Young, P., and Hastie, T. (2007). L₁-Regularization Path Algorithm for Generalized Linear Models. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)* **69**, 659-677.

CHAPTER V

Conclusions and Future Work

5.1 Conclusion

Propensity score methods are used to make causal inference in non-randomized observational studies. The goal of these methods is to create covariate balance among treatment groups, thereby mimicking randomized control trials. Although there are countless methods and techniques to implement propensity scoring when presented with binary treatments, extensions to multinomial and continuous treatments are under-studied. The work presented here fills significant gaps within the propensity score literature by introducing two novel methodologies that remove imbalance due to confounding in observational studies with either multinomial or continuous treatments.

As discussed in Chapter I and II, current multinomial propensity score methods do not have the same flexibility as the scalar value derived in binary treatment settings. Therefore, the goal of Aim 1 was to develop a novel methodology of propensity score analysis that derives a single scalar balancing score for multinomial treatments. The proposed method, the GPS-CDF method, accurately maps the GPS vector, produced by either parametric or non-parametric models, to a scalar value that can be used to match or stratify subjects. The utility of the GPS-CDF method, when presented with multinomial treatments, was assessed via simulation and through application using an electronic health records data

set. The flexibility and application of the GPS-CDF method provides researchers with a new option, more relatable to standard binary propensity score techniques, when conducting multinomial treatment propensity scoring.

Current methods for conducting propensity score analysis in the presence of continuous treatments, as detailed in Chapter I and III, rely on weighting. Although these methods are not inappropriate, they do not always produce accurate effect estimates due to extreme weights. The goal of Aim 2 was to develop a novel method of propensity score analysis for continuous treatments that does not rely on weighting. Both the GPS-CDF and npGPS-CDF methods were derived to stratify subjects, based on pre-treatment confounders, in order to create covariate balance in the presence of continuous treatments. Simulations as well as an application to the MATCH data showed that these newly developed stratification methods, GPS-CDF and npGPS-CDF, performed better than standard continuous treatment propensity score weighting methods. Our novel methodologies allow researchers to conduct propensity score analyses without utilizing weighting, in both parametric (GPS-CDF) and nonparametric (npGPS-CDF) frameworks, when presented with continuous treatments.

Finally, Aim 3 provides an R package to implement the multinomial (and ordinal) GPS-CDF method detailed in Chapter II. Although the multinomial GPS-CDF method is straightforward to implement in practice, few researchers, especially those with non-computational backgrounds, will take the time to implement the method for themselves. Therefore, having a standard R package available that implements the novel GPS-CDF method will hopefully allow more robust propensity score analyses by researchers, when presented with multiple treatments.

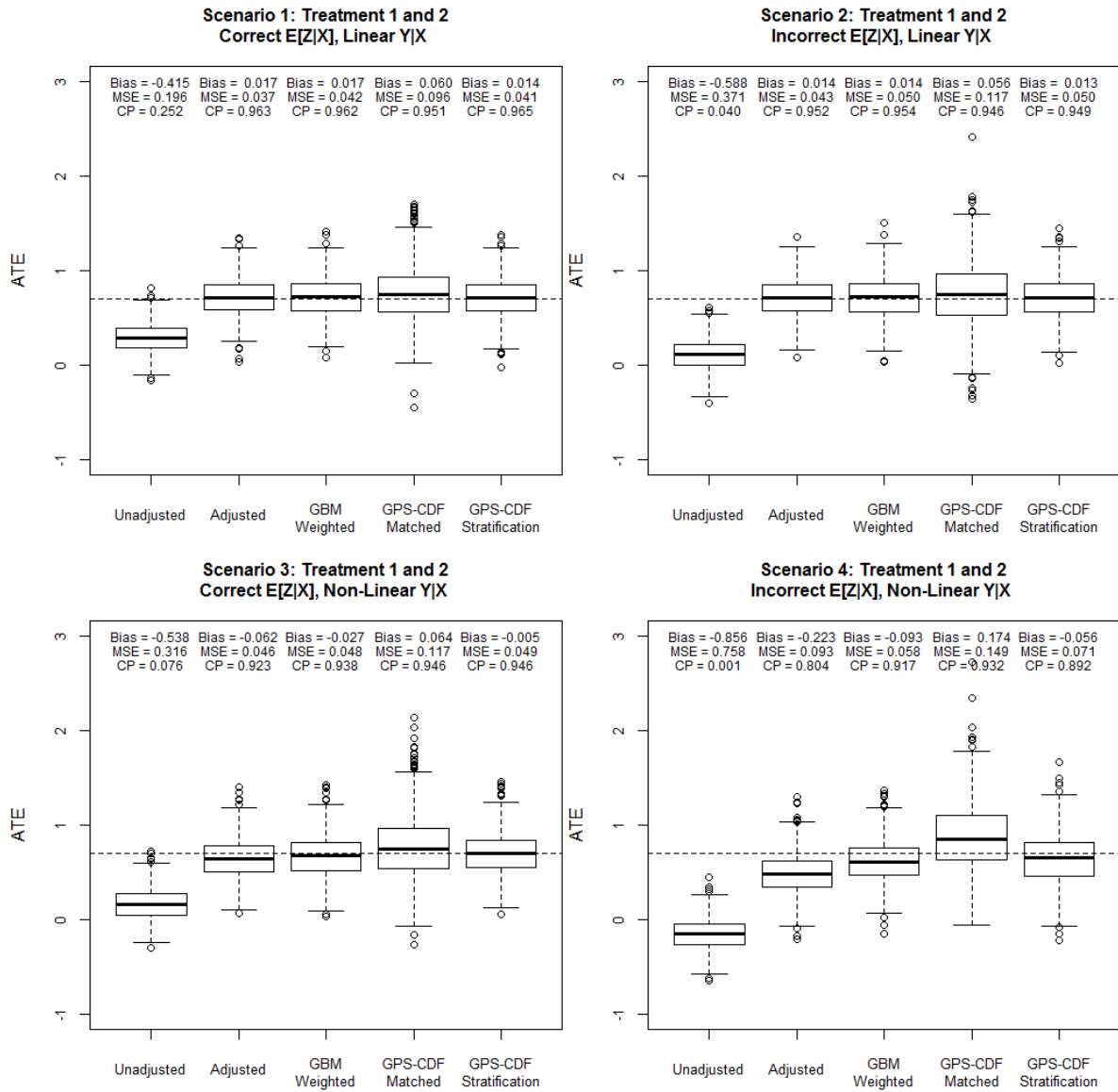
The overall strength of this research is the novelty of the proposed methods. Currently, methods of implementing multinomial and continuous treatment propensity scoring are not as well developed or studied as those for standard binary propensity scoring. Aim 1 and 2 provided new multiple treatment propensity score methodologies and further tested these novel methods in simulation. Each simulation contained multiple treatment and outcome scenarios that were representative of real data. These detailed simulations ensure that the methodologies developed in Aim 1 and 2 will translate well to real data applications. Furthermore, when our novel methodologies were applied to real data applications (i.e. an EHR data set, and the MATCH study), they out performed current standard methods in terms of achieved covariate balance. Furthermore, the utility of any new methodology is directly related to the ease at which it can be implemented. Therefore, the R package developed in Aim 3 will hopefully lead researchers to use of the GPS-CDF method in practice.

There are limitations within the current work that should be acknowledged. Both simulation studies conducted within Aim 1 and 2 only considered four different data scenarios. Thus, the results detailed in Chapters II and III could differ under different modeling assumptions. Additionally, the multinomial GPS-CDF method detailed in Chapter II only utilized one non-parametric method, GBM. Therefore, it is unclear if other non-parametric methods of deriving the GPS vector will produce better balance and outcome estimates when used in tangent with the GPS-CDF method. Finally, the continuous GPS-CDF method detailed in Chapter III assumed normal distributions when bootstrapping/re-sampling subject specific covariate distributions. Other distributions, including the T-distribution or non-parametric distributions, may lead to more accurate balancing strata.

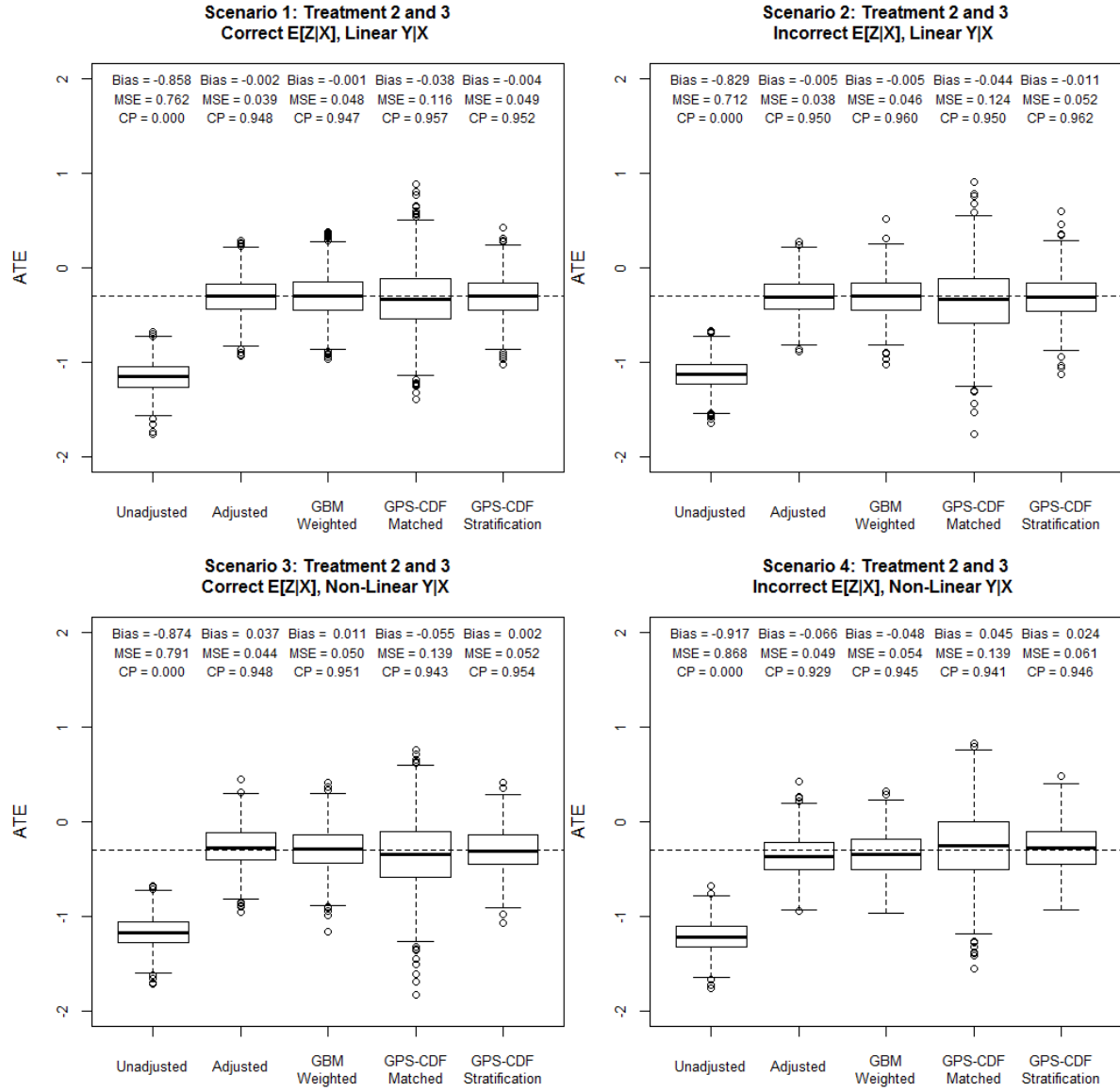
There are still many avenues for future research into propensity score methodologies for multiple treatments. As indicated above, other non-parametric machine learning techniques should be tested and examined to find a methodology that most accurately derives the GPS vector for multinomial treatments. Additionally, as our method was able to accurately derive effect estimates when presented with three treatment groups, the multinomial GPS-CDF method may be extended to genetic applications when analyzing SNP data. Furthermore, through our work, it appears that stratification methods may outperform weighting methods within the continuous treatment setting. Therefore, other stratification techniques may be derived and tested in the continuous treatment setting. For example, stratification based on the predicted value derived from the CBGPS method may prove more accurate than stratification based on a linear model. Overall, propensity scoring is a growing tool for researchers to implement when working with observational data. It is up to subject specific researchers to determine which method creates the optimal balance in their data. Future propensity score research should continue to evaluate the merit and application of the GPS-CDF methodologies when working with multinomial and continuous treatments.

Appendices

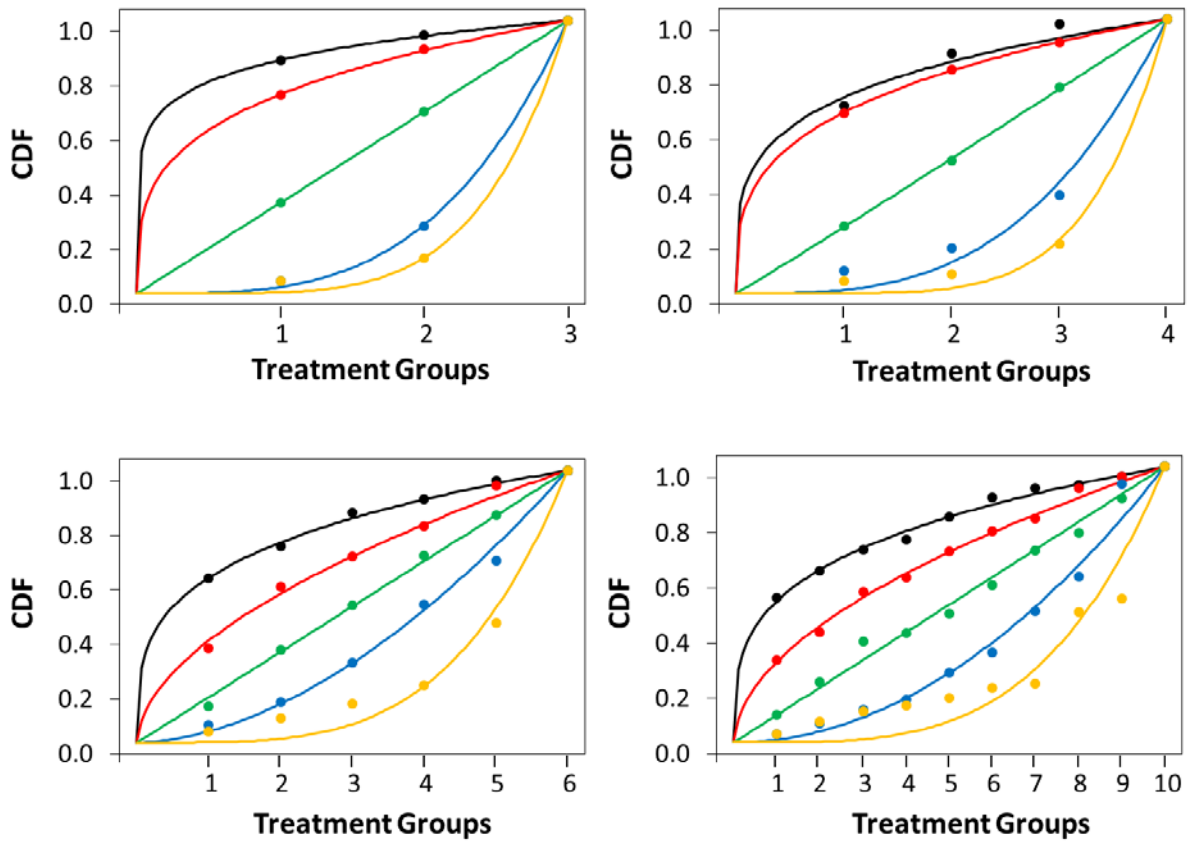
Appendix A. Supplemental Figures



Supplemental Figure 1. Distribution of the ATE for each method under each scenario between treatment 1 and treatment 2. The true ATE value of 0.7 is included as the dotted horizontal line.



Supplemental Figure 2. Distribution of the ATE for each method under each scenario between treatment 2 and treatment 3. The true ATE value of -0.3 is included as the dotted horizontal line.



Supplemental Figure 3. Graphical representation of the CDF mapping produced by the GPS-CDF method under 3, 4, 6, and 10 multinomial treatment group scenarios, respectively.

Appendix B. Code for implementation of GPS-CDF method using R software

```
#####
# R Function to create Generalized Propensity Score #
#      Cumulative Distribution Function (GPS-CDF)    #
#####

GPSCDF<-function(pscores=NULL, data=NULL, trt=NULL, stratify=FALSE, nstrat=5,
  optimal=FALSE, greedy=FALSE, ordinal=FALSE, multinomial=FALSE,
  caliper=NULL){

  if(is.null(stratify)){
    stratify<- FALSE
  }

  if(is.null(optimal)){
    optimal <- FALSE
  }

  if(is.null(greedy)){
    greedy <- FALSE
  }

  N<-dim(pscores)[1] # Number of subjects
  size<-dim(pscores)[2] # Number of treatments

  cpscores<-t(apply(pscores[,], 1,cumsum))
  Z<- seq(1, dim(pscores)[2], by=1)

  if(sum(pscores)/N == 1){

    Znorm<-sort(unique(Z))/max(unique(Z))
    ppar<-rep(0,N)

    for( i in 1:N){
      y<-cpscores[i,]
      mod<-stats::nls(y~I(Znorm^exp(power)), control = stats::nls.control(maxiter =
        150, tol = 1e-05, minFactor = 1/1024,printEval = FALSE, warnOnly = TRUE),
        start = list(power = 0),trace = F)
      parm<-summary(mod)$coefficients[1]

      mod<-stats::nls(y~I(Znorm^exp(power)), control = stats::nls.control(maxiter =
        150, tol = 1e-05, minFactor = 1/1024,printEval = FALSE, warnOnly = TRUE),
        start =list(power =parm),trace = F)
      parm<-summary(mod)$coefficients[1]

      ppar[i] <- parm
    }

    if (!is.null(data)){
      data$a<- ppar
      data2<- data
    }
  }
}
```

```

#Set up Stratification
if(stratify==TRUE){
  strata<-dplyr::ntile(ppar, n=nstrat)

  if (!is.null(data)){
    data$strata<-strata
  }
} else{
  strata<- NULL
  nstrat<- NULL}

#Set up Optimal Matching
if(optimal==TRUE){

  if (is.null(data)){
    stop('Specify a dataframe to attach matches')
  } else{

    if (is.null(trt)){
      stop('Specify a treatment variable to proceed with matching')
    } else{

      if (ordinal==FALSE & multinomial==FALSE){
        stop('Specify Ordinal or Multinomial treatments')
      } else{

        # Set up matching score
        epsilon=1e-5
        deltammat<-matrix(0,nrow=N,ncol=N)
        deltammat2<-matrix(0,nrow=N,ncol=N)

        # Loops to set up delta matrix
        if(ordinal==TRUE){
          for(i in 1:N){
            deltammat[i,]<-((ppar[i]-ppar)^2+epsilon)/((trt[i]-trt)^2)
          }
        }

        if(multinomial==TRUE){
          for(i in 1:N){
            for(k in 1:N){
              if(trt[i]==trt[k]){deltamat[i,k]<-999} else{
                deltammat[i,k]<- abs((ppar[i]-ppar[k]))}
            }
          }
        }

        for(i in 1:N){
          deltammat2[i,]<-abs((ppar[i]-ppar))
        }

        #Get rid of Inf and put in 999999
        deltammat[!is.finite(deltamat)]<-99
        diag(deltamat)<-99
      }
    }
  }
}

```

```

# Derigs algorithm only works with integers so we can multiply all
distances # by 10,000 to get accuracy to 4 decimal places
deltamatint<-deltamat*100000

# Use distancematrix function to reform so we can do NBP matching
suppressWarnings(distmat<-nbpMatching::distancematrix(deltamatint))

# Set up matches
invisible(utils::capture.output(matchset<-
nbpMatching::nonbimatch(distmat)))

#Remove row if N is odd
matches1<-matchset$halves[ grep("ghost", matchset$halves$Group2.ID,
invert = TRUE) , ]
matches<- matches1[ grep("ghost", matches1$Group1.ID, invert = TRUE) , ]

#Find distances of matches
matchmat<-matrix(NA, nrow=round(dim(matches)[1]), ncol=3)
for(i in 1:dim(matches)[1]){
  pair<-matches[i,c(2,4)]
  value<- deltammat2[pair[1,1],pair[1,2]]

  matchmat[i,1]<-pair[1,1]
  matchmat[i,2]<-pair[1,2]
  matchmat[i,3]<-value
}

npairs<-dim(matchmat)[1]

#Calculate Average Total Distance of Matched Pairs
optdistance<- sum(matchmat[,3])/npairs

data$optmatch<-0
# Attach matches to data
for(i in 1:dim(data)[1]){
  data$optmatch[matches[i,2]]<-i
  data$optmatch[matches[i,4]]<-i
}
optmatch<-data$optmatch
}
}
} else{
  optmatch<- NULL
  optdistance<- NULL}

#Set up Greedy Matching
if(greedy==TRUE){

  if (is.null(data)){
    stop('Specify a dataframe to attach matches')
  } else{

    if (is.null(trt)){
      stop('Specify a treatment variable to proceed with matching')

```

```

} else{

  if (ordinal==FALSE & multinomial==FALSE){
    stop('Specify Ordinal or Multinomial treatments')
  } else{

    # Set up matching score
    epsilon=1e-5
    deltammat<-matrix(0,nrow=N,ncol=N)
    deltammat2<-matrix(0,nrow=N,ncol=N)

    #Set Caliper
    if (is.null(caliper)){
      caliper<-0.25*stats::sd(ppar)
    }

    # Loops to set up delta matrix
    if(ordinal==TRUE){
      for(i in 1:N){
        deltammat[i,]<-((ppar[i]-ppar)^2+epsilon)/((trt[i]-trt)^2)
      }
    }

    if(multinomial==TRUE){
      for(i in 1:N){
        for(k in 1:N){
          if(trt[i]==trt[k]){deltamat[i,k]<-999} else{
            deltammat[i,k]<- abs((ppar[i]-ppar[k]))
          }
        }
      }
    }

    for(i in 1:N){
      deltammat2[i,]<-abs((ppar[i]-ppar))
    }

    #Get rid of Inf and put in 999
    deltammat[!is.finite(deltamat)]<-999
    diag(deltamat)<-999

    # Use Greedy matching to get matches
    # Set up matches

    # Set up holding for matches
    matchmat<-matrix(NA, nrow=round(dim(deltamat)[1]/2), ncol=3)

    #Replace matched pairs with maximum of delta matrix so it wont be used
    again
    repnum<-max(deltamat)

    i<-0
    while(min(deltamat) < caliper){
      i<-i+1
      inds = which(deltamat== min(deltamat), arr.ind=TRUE)
      value= deltammat2[inds[1,1],inds[1,2]]

      pair<-inds[1,1:2]

```

```

        matchmat[i,1:2]<-pair
        matchmat[i,3]<-value

        deltammat[pair,]<-repnum
        deltammat[,pair]<-repnum
    }

    matchmat<-matchmat[is.na(matchmat[,1])!=F,]

    npairs<-dim(matchmat)[1]

    #Calculate Average Total Distance of Matched Pairs
    grdydistance<- sum(matchmat[,3])/npairs

    # Attach matches to data
    data2<- data2[matchmat,]

    data2$grdmatch<-0
    for(j in 1:(dim(data2)[1]/2)){
        data2$grdmatch[j]<-j
        data2$grdmatch[j+dim(data2)[1]/2]<-j
    }
    grdmatch<-data2$grdmatch
    grddata<-data2
    }
}
} else{
    caliper<- NULL
    grdmatch<- NULL
    grddata<- NULL
    grdydistance<-NULL}

} else{
    stop('Pscores must add to 1')
}

returnlist<-list(ppar=ppar, data=data, nstrat=nstrat, strata=strata,
    optmatch=optmatch, optdistance=optdistance, caliper=caliper,
    grddata=grddata, grdmatch=grdmatch, grdydistance=grdydistance, NULL=NULL)
returnlistfinal<- returnlist[-which(sapply(returnlist, is.null))]
return(returnlistfinal)

}

#####
# END GPS-CDF Function #
#####

```