Distribution Agreement

In presenting this thesis or dissertation as a partial fulfillment of the requirements for an advanced degree from Emory University, I hereby grant to Emory University and its agents the non-exclusive license to archive, make accessible, and display my thesis or dissertation in whole or in part in all forms of media, now or hereafter known, including display on the world wide web. I understand that I may select some access restrictions as part of the online submission of this thesis or dissertation. I retain all ownership rights to the copyright of the thesis or dissertation. I also retain the right to use in future works (such as articles or books) all or part of this thesis or dissertation.

Signature:

_____ _____

Name: Haocan Song Date: 04/09/2018

Assess Improvement of Balancing Covariates by Propensity Score approach using Generalized Boosted Model (GBM) and Application Based on National Cancer Database

By

Haocan Song
MPH

Biostatistics and Bioinformatics Department

Thesis Advisor: Yuan Liu, PHD

Reader: Kundu Suprateek, PHD

Assess Improvement of Balancing Covariates by Propensity Score approach using Generalized Boosted Model (GBM) and Application Based on National Cancer Database

By

Haocan Song

B.A., Southwest University, 2016

Thesis Committee Chair: Yuan Liu, Doctor

An abstract of
A thesis submitted to the Faculty of the
Rollins School of Public Health of Emory University
in partial fulfillment of the requirements for the degree of
Master of Public Health
in
Biostatistics and Bioinformatics Department
2018

Abstract

Comparison of Covariate Balance by Three Propensity Score Estimation Approaches: An Application Based on National Cancer Database
By Haocan Song

Background: Observational study is one of the most commonly used study designs in many medical research, but they have a major limitation of getting vulnerable to selection bias to make valid causal inference. Propensity score (PS) matching and weighting are popular methods that can be applied to reduce the bias and estimating causal effects in observational studies. In this work, we focused on General Boosted Method (GBM), a tree-based approach to obtain more accurate estimated PS score without specifying the form of prediction function, and we further compared its performance in terms of covariate balancing with the conventional model-based approach, such as logistic regression.

Method and Study Design: In this study, we tested 3 alternative methods for propensity score (PS) estimation: main-effect logistic regression model (model 1: LOGREG), comprehensive logistic regression model with all two-way interactions and polynomial terms (model 2: LOGREG(INT)), and GBM (model 3). Implemented these algorithms for an application based on prostate cancer from NCDB dataset, where we aimed to conduct an effect comparison of overall survival between proton radiation therapy and conventional x-ray based radiation therapy. Matching was performed to eliminate confounding effect via PSM with caliper and different matching ratio up to 1:5. Balance was evaluated before and after matching by standardized difference. The proportional hazard model was carried out to estimate the hazard ratio of proton therapy with 95% confidence interval in the matched sample.

Conclusion: The study reveals that covariate balancing can be improved by a more accurate PS estimation model through GBM or comprehensive logistic regression, and both approaches should be encouraged in the practice. In case study, we also found that proton radiation therapy hold an improved clinical benefit for prostate cancer patients for long-term survival.

KEYWORDS: Observational study, Propensity score, Matching, GBM, Covariates Balance Check

1. INTRODUCTION

1.1 Observational Study

In many clinical studies, making comparison about the effect of different interventions or treatments is commonly desired. Randomized studies are the gold standard for determining whether one treatment is superior to another. However, for practical or ethical reasons, randomized studies are not always possible. In these situations, observational studies could provide an important source of information when randomized controlled trials or case control studies cannot or should not be used. Observational studies are always the only feasible options in many clinical studies, but the studies have a major limitation of getting vulnerable to selection bias, a situation where individual characteristics (covariates) are related to the likelihood of receiving the treatment, and such relations lead to an inaccurate estimate of the treatment effect (Rosenbaum, 2002). In other words, the baseline characteristics of the population under one treatment could dramatically differ from the other one (Yuan, Dana, & Joseph, 2013) If this problem for observational study couldn't be addressed sufficiently, the heterogeneity of characteristic for different treatment group will introduce confounding effects into a causal-effect relationship and result in bias in the estimation of treatment effect.

1.2 Propensity Score

To deal with the issue of confounding due to nonrandom treatment assignment, several statistical and econometric techniques are commonly employed, including multivariable regression analyses that attempt to control for potential confounders in observational studies (Reshma, 2014) Standard multiple regression techniques are limited in deciding the relationship between covariates and outcome and not all confounders may be observable in specific studies. In these situations, propensity score (PS) method (Paul & Donald, 1983) can play an important role in improving the accuracy of statistical inferences. The propensity

score (PS) is defined as a subject's probability of receiving a specific treatment assignment conditional on the observed baseline covariates:

$$e_i = \Pr\,(Z_i = 1|X_i)$$

Where e_i denotes the propensity score for the i-th subject, Z_i denotes the indicator variable whether or not the ith subject was in the treatment group. X_i denotes the ith subject. The propensity score is a balancing score: conditional on the propensity score, the distribution of measured baseline covariates is similar between treatment group and comparison group. Thus, ideally, in a set of subjects all of whom have the same propensity score, the distribution of observed baseline covariates will be the same between the treated and untreated subjects. It is one of the most applicable approaches that have been widely used in practice to reduce the selection bias and build up casual inference based on observational data.

1.3 Variable Selection for the Propensity Score Model

As propensity score is defined as the probability of treatment assignment, statisticians are more in favor of the inclusion of only those variables that affect treatment assignment. However, there is not an exact instruction or definition in the applied literature about which variables to include in the propensity score model. But a few general guidelines for covariate selection have been offered. Through experience, the possible sets of variables for inclusion in the propensity score model include the following parts: all measured baseline covariates, all baseline covariates that are associated with treatment assignment, all covariates that affect the outcome (i.e., the potential confounders), and all covariates that affect both treatment assignment and the outcome (i.e., the true confounders) (Austin, 2011).

1.4 Propensity Score Calculation

Once the covariates have been selected, the PS could be estimated for each individual. The value of propensity score estimation is so important, since the accuracy would directly affect the PS analysis later in the procedure. The estimation of PS is typically done parametrically via generalized linear modeling (i.e., logistic regression, probit regression, or discriminant analysis) where treatment status is regressed on the covariates. This method can theoretically eliminate the confounds for observed covariates, but accurate estimation of propensity scores is impeded by large numbers of covariates, uncertain functional forms for their associations with treatment selection, and other problems. A popular and advanced alternative approach of accounting for propensity score is the non-parametric estimation via generalized boosted modeling (GBM), which can overcome many of these obstacles.

1.4.1 Main-effect Logistic Regression Model (LOGREG)

The majority of published propensity score analyses use logistic regression method to estimate. Through the literature review, we found the method of logistic regression is really attractive for probability prediction since it is mathematically constrained to produce probabilities in the range (Gail, Krickeberg, Samet, Tsiatis, & Wong, 2002) and generally converges on parameter estimates relatively easily. Further, logistic regression is also a familiar and reasonably well-understood tool of researchers and statisticians in a variety of disciplines and is easy to implement in most statistical packages (Westreich, Lessler, & Funk, 2010). Even logistic regression is still wide used in practice, its accuracy for PS estimation can be impacted by the final model specified, and it is highly criticize that the interaction terms or the polynomial function for continuous variables should be added into the predicting model for PS.

1.4.2 Comprehensive Logistic Regression Model with all Two-way Interactions and Polynomial Terms (LOGREG(INT))

Since most of researchers are used to using the ordinary and simple logistic regression model for data analysis, this familiarity will predispose investigators to using logistic regression even when better alternatives may be available, like the model with interactions and polynomial. Adding products (including polynomial or interaction) terms of the covariates in the PS estimation is also an applicable method. For application of this method, it could be more comprehensive and reliable since it contains more main effects. However, sometimes this kind of PS model can easily become very complex and hard to decide which interactions to be included, especially when the number of covariates becomes large.

1.4.3 Generalized Boosted Models (GBM)

In order to increase the accuracy of propensity score estimation, statisticians developed the general bootstrap method (GBM) to calculate the propensity score. GBM is an general, automated, data adaptive modeling algorithm that can estimate the nonlinear relationship between a variable of interest and a large number of covariates, for iteratively forming a collection of simple regression tree models to add together to estimate the propensity score. As a modern statistical technique, the GBM has been used in many statistical procedures to improve the validity and accuracy of statistical analyses through estimating more accurate standard errors than traditional statistical techniques (Efron & Tibshirani, 1993). Also, because the final GBM model is a sum of regression trees, it inherits many of their advantageous properties for estimating propensity scores. Trees are computationally fast to fit (Breiman, 1984) and trees could handle continuous, nominal, ordinal, and missing independent variables. Especially for existing missing values in the dataset, GBM automatically adds indicators for missing values and includes them in the model.

It is appealing in the context of case-mix adjustment since it can predict treatment assignment from a large number of pretreatment covariates while also allowing for flexible, non-linear relationships between the covariates and the propensity score (McCaffrey DF, 2004).

1.5 Propensity Score Matching

There are several methods to apply the propensity scores to compare the treatment and comparison groups. The most popular choices are matching (Stuart & Rubin, 2008) (Stuart & Rubin, 2008) subclassification (Lunceford & Davidian, 2004) and weighting (Hirano K, 2003) (Robins JM, 2000). In this article, we put more strength on Propensity score matching. The PS can be utilized to form matches of treated and comparison cases for which the treatment effect is examined. Several matching algorithms are available in the published literature before, researchers are encouraged to try out different algorithms to see which one serves best for the particular dataset. The commonly used propensity score matching methods are: nearest neighbor matching, caliper matching, and Mahalanobis metric matching.
In this article, simple greedy matching method and 1-1, 1-N caliper matching method are used to test the difference.

1.5.1 Greedy Matching

The greedy matching method, described by Parsons (LS, 2001), is rounding the propensity score to 5 significant figures and randomly selecting pairs that match exactly on this score. For the unmatched subjects, the score is then rounded to 4 significant figures and exact matches selected, with the process continuing until subjects are matched to 1 significant figure. However, for the subject that once a match is made, it is never reconsidered, which means the controlled subjects are considered without replacement. This matching method is the best match currently available (Parsons, 2001).

1.5.2 1-1 to 1-N Caliper Matching

Nearest neighbor matching within a specified caliper distance has the further restriction that the absolute difference in the propensity scores of matched subjects must be below some pre-specified threshold (the caliper distance). In practice, a wide variety of calipers distance is

used (Austin P. , 2008). Previous article states that the usage of a caliper equal to 0.25 standard deviations is recommended (Cochran & Rubin, 1973), however, with the exception of Austin (Austin P. , 2011), reducing the caliper from 0.25 standard deviations to 0.2 standard deviations is more accurate suggested.

And for 1-1 matching, each treated subject is matched to its nearest one control, while one-to-many (1-N) matching (Kewei Ming, 2001) matches each member of the treatment group to a fixed or variable number of persons in the comparison group. 1-N matching is particularly useful when the size of the groups differs largely in the original sample; it can increase the overall size of the matched sample and thus efficiency in the estimation of treatment effect. A downside is the risk of bias stemming from the additional matches that are not always as close as the first match.

1.6 Treatment Effect

Causal effects for individuals generally cannot be estimated due to the fundamental problem of causal inference: we cannot observe an individual under each of the multiple treatments being compared (P., 1986). Instead, we only observe what happens to an individual under the treatment condition they actually received (McCaffrey DF, 2004).

Every member in the population has two potential value of treatment for any outcome. One is the treatment condition, t_1; and the other one is comparison group, t_0. Only one of these values is observed for each individual. The treatment effect is $E(t_1) - E(t_0)$, where expectation is over the entire population. Let Z be an indicator variable denoting the treatment received (Z_0 for control treatment vs. Z_1 for active treatment). Thus, only one outcome, t_i ($t_i = Z_i * t_1 + (1 - Z_i) * t_0$) is observed for each subject: the outcome under the actual treatment received. Here we consider 2 treatment effect methods:

1.6.1 Average Treatment Effect (ATE): The ATE of treatment t_1 relative to treatment t_0 is the comparison of mean outcomes had the entire population been observed under one treatment t_1 versus had the entire population been observed under another treatment t_0(Wooldridge, 2002). More formally, the ATE for comparing treatment t_0 and t_1 equals $E(D[t_1, t_0]) = E(Y[t_1] - Y[t_0]) = E(Y[t_1]) - E(Y[t_0])$, where expectation is over the entire population.

1.6.2 Average Treatment Effect Among the Treated (ATT): The ATT of treatment t_1 among those treated with treatment t_0 is the comparison, among study participants who were treated with t_0, of their mean outcome when treated with treatment t_0, as they were, with the mean outcome they would have had if they had instead been treated with treatment t_1 (Wooldridge, 2002). More formally, the ATT for comparing treatment t_0 and t_1 equals $E(D[t_1, t_0]) = E(Y[t_1] - Y[t_0] \mid Z_1) = E(Y[t_1] \mid Z_1) - E(Y[t_0] \mid Z_1)$, where expectation is over the treatment group.

The ATEs and ATTs can differ when the treatment effects are not constant across individuals. The choice of estimand depends on the substantive questions a study hopes to address and the population that is the target of the treatment. A study can estimate both ATE and ATT, but one or the other typically is better suited for any particular situation. The ATEs are more likely to be of interest compared with ATTs if every treatment potentially might be offered to every member of the population. Conversely, if the research question focuses on the effectiveness of one treatment program, then the ATT would be of interest because it measures the relative effectiveness of programs t_0 and t_1 on the population receiving program t_1.

1.7 Checking balance on the covariates before and after matching

The next step after the PS estimation is to check the balance before and after matching for the selected covariates. A straightforward for statisticians to test the balance is to test the covariates for absolute standard differences(ASD) between the treatment and comparison groups. As stated in the literature, the absolute standard differences(ASD) is an alternative, more commonly accepted measure of covariance balance nowadays (G.W, 2004). This method can be calculated as the group difference in means of a continuous covariate and the categorical covariates. Imbalance would be expected for some covariates; even in randomized trials exact balance is a large-sample theory (Austin, 2011).

For a continuous covariate, the standardized difference is defined as:

$$d = \frac{(\bar{x}_{treatment} - \bar{x}_{control})}{\sqrt{\frac{s^2_{treatment} + s^2_{control}}{2}}}$$

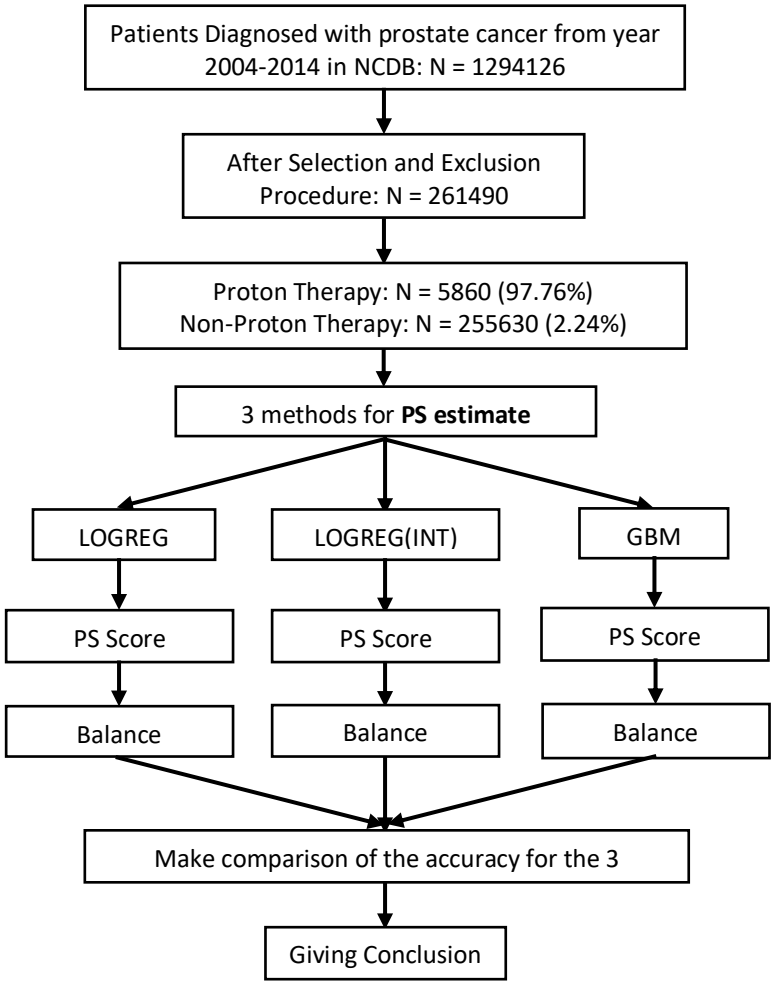
where $\bar{x}_{treatment}$ and $\bar{x}_{control}$ denote the sample mean of the covariate in treated and comparison subjects, respectively. Whereas $s^2_{treatment}$ and $s^2_{control}$ denote the sample variance of the covariate in treated and comparison subjects, respectively.

For dichotomous variables, the standardized difference is defined as:

$$d = \frac{(\hat{p}_{treatment} - \hat{p}_{control})}{\sqrt{\frac{\hat{p}_{treatment}(1 - \hat{p}_{treatment}) + \hat{p}_{control}(1 - \hat{p}_{control})}{2}}}$$

where $\hat{p}_{treatment}$ and $\hat{p}_{control}$ denote the prevalence or mean of the dichotomous variable in treated and untreated subjects, respectively. The standardized difference compares the difference in means in units of the pooled standard deviation. Moreover, it is not influenced by sample size and allows for the comparison of the relative balance of variables measured in different units. Although there is no universally agreed upon criterion about what's the exact value of threshold of the standardized difference, which can be used to indicate important imbalance, a standard difference that is less than 0.1 has been taken to indicate a negligible

difference in the mean or prevalence of a covariate between treatment groups (Normand, 2001)



2. CASE STUDY

2.1 Study Objective

Prostate cancer is the most common non-cutaneous malignancy diagnosed among men in the United States. Recent advances in medical technology have introduced many new forms of therapy for the treatment of prostate cancer that are frequently the subject of comparative effectiveness research (Nguyen, Gu, & Lipsitz, 2001) (Wisenbaugh, Andrews, & Ferrigni, 2014). Perhaps the most controversial form of definitive treatment for prostate cancer is proton therapy. The unique dose distribution properties of proton therapy theoretically allow clinicians to increase target dosage while reducing exposure to surrounding normal anatomy (Wisenbaugh, Andrews, & Ferrigni, 2014).

Despite the advantage of proton therapy over photon external beam radiotherapy, to the best of our knowledge there is little consensus regarding whether significant toxicity and/or outcome benefits exist and whether the benefits are worth the cost of adopting an expensive new technology (Efstathiou JA, 2013). However, the recent dissemination, adoption, and marketing of proton therapy for prostate cancer has left many patients seeking proton therapy and clinicians with the difficult task of evaluating whether it is a cost-effective option for their patients (Shah A, 2013).

In this case study, we set the treatment group with Proton therapy V.S. comparison therapy with non-Proton treatment (including treatments like: external beam(NOS), Photons(2-5MV), Photons(6-10MV), Photons(11-19MV), Photons(>19MV), Photons(mixed energies), Intensity Modulated Radiation Therapy(IMRT), Conformal or 3-D therapy). The goal of this study is to use a propensity score matched (PSM) analysis with the National Cancer Database (NCDB) for the comparison of Proton therapy and Non-Proton therapy for organ confined prostate cancer. And make comparison on the 3 methods of calculating propensity score.

2.2 NCDB database

One empirical example about the propensity score method is given to demonstrate a comprehensive process of PS analysis. In this paper, we introduce an observational study from National Cancer Database(NCDB) that motivates our methodology. The NCDB is jointly sponsored by the American College of Surgeons and the American Cancer Society. It is a clinical oncology database sourced from hospital registry data collected in more than 1,500 commission on cancer-accredited facilities. NCDB data are used to analyze and track patients with malignant neoplastic diseases, their treatments, and outcomes. Data represent approximately 70% of newly diagnosed cancer cases nationwide and 34 million historical records (Bilimoria KY, 2008). At each hospital, certified tumor registers abstract data from patient medical records, and registrars are required to obtain and submit patient treatment and follow-up data even if part of the care is received at another (e.g. non-CoC-accredited) hospital. Annually, registrars upload data to the NCDB on incident cancer cases as well as follow-up information on existing patients. Data contained in the NCDB include: demographics (age, gender, race, marital status, medical insurance), comorbidity status, stage, treatments received, recurrence and survival (Reshma, 2014). The NCDB includes prostate cancer patients treated from 2004 to 2014 providing information on demographics, risk factors specific to prostate cancer, staging information, treatment, and survival data for de-identified Patients.

2.3 Define study population

The study population consists of 1294126 patients from year 2004-2014 who have a prostate cancer in this Data Dictionary. Before PS calculation, we do the selection and exclusion procedure. We initially included all patients diagnosed between 2004 and 2014; excluded those patients with carcinoma in situ (not invasive); specify desired radiation modality

format: external beam(NOS), Photons(2-5MV), Photons(6-10MV), Photons(11-19MV), Photons(>19MV), Photons(mixed energies), Intensity Modulated Radiation Therapy(IMRT), Conformal or 3-D therapy, Protons; include desired radiation volume: Pelvis(NOS), Prostate and pelvis, Prostate; exclude stage IV and metastasis cases; and exclude the missing outcome from Last Contact or Death and Month. Finally leaving 261460 patients with our desired character for the study.

Table 1: Selection/Exclusion Diagram

Selection and Exclusion Criteria	Sample Size	Excluded
NCDB Prostate PUF Cancer Cases	1294126	-
Include YOD 2004-2014	1294126	0
Include Invasive cases	1293888	238
Include Desired Radiation Modality	319086	974802
Include Desired Radiation Volume	305147	13939
Exclude Stage IV and Metastasis Cases	280888	24259
Exclude missing outcome	261490	19398

2.4 Select the covariates

The first step of a PS method analysis is to decide which covariates should be included in estimating the PS for each participant. After the evaluation of the covariates, there are 124 variables totally in the initial dataset. And we decided to include 16 covariates in the final model. Demographic variables evaluable from the NCDB include 11 variables: age, year of diagnosis(quantile), race, Hispanic or not, insurance status, median income quartiles, patient comorbidity via the Charlson–Deyo comorbidity score, facility type, facility location, Percent

No High School Degree Quartiles, Urban/Rural area and Great circle distance. Tumor and treatment specific factors evaluable from the NCDB include 5 variables: prostate-specific antigen (PSA), Gleason score and Grade level as well as AJCC Analytic Stage Group.

2.5 Statistical methods

Statistical analysis was conducted using SAS Version 9.4, and SAS macros or software developed at the Biostatistics and Bioinformatics of Emory University at Winship Cancer Institute. The significant level was set at 0.05. Descriptive statistics for each variable were reported. The univariate association between each covariate and study cohorts were assessed using the χ^2test for categorical covariates and ANOVA for numerical covariates. The univariate association between each covariate including study cohorts and study outcome (OS) were assessed using Cox proportional hazards models and log-rank tests. A multivariable Cox proportional hazard model was fit by a backward variable selection method applying an alpha =.20 removal criteria. The stratified analysis was conducted by including the interaction term between study cohorts and a stratified variable in a multivariable model and then hazard ration was estimated for study cohorts in each level of the strata variable. KM plots were produced to compare the survival curves by subgroups along with log-rank p-value.

3. RESULTS

3.1 Patients characteristics

First we make the description table for all the 16 variables and treatment variable "Proton"(Appendix-Table1). Among the table, a total of 261490 patients are included. Still, we found some basic descriptive and distributive character of covariance: 97.8% of patients are treated as Non-proton therapy with only 2.2% of patients are in Proton therapy group;

69.3% of patients accept the therapy in Non-Academic/Research Program; most of patients are White and Hispanic; 59.5% patients have Medicare insurance type; patients with larger income are more prefer to take the therapy; 82.5% of patients are form Metro part; 87.0% patients has Charlson-Deyo Score=0; 82.7% patients with AJCC Analytic Stage Group Stage II, etc.

Then we get Patient and treatment characteristics by treatment group (Appendix Table2). There are 255630 Non-proton therapy patients and 5860 Proton therapy patients. Significant differences groups existed on the all of the 17 variables with $p<0.05$.

By multivariable logistic regression model to predict proton therapy vs. non-proton therapy following the backward selection, no variables were removed from the model. Multivariate logistic regression Significant difference for all the 17 variables with type3 P-value<0.05. Among the proton therapy (treatment group), we got the odds ratio for 16 categorical variables and 1 continuous variable: patients with facility type from Academic/Research Program are much more likely to get Proton treatment (OR=41.34) compare with Non Academic/Research Program; patients in the west, White patients, no Hispanic patients, patients with Medicare insurance type, patients live in Metro part of city, patients with Charlson-Deyo Score equals to +1, patients get the treatment after 2011, AJCC Analytic Stage Group at Stage I, PSA less than 10, Gleason Score between 2-7 and lower age are more likely to get Proton treatment. The table is provided in Appendix Table3.

Last, run the Multivariable Survival Analysis of OS Main effect and gives out HR value for 17 variables. After the backward selection, no variables were removed from the model. Significant difference for all the 17 variables with type3 P-value<0.05. Among the proton therapy (treatment group), we got the odds ratio for 1 treatment, 16 categorical variables and 1 continuous variable: patients with proton therapy has less hazard death than non-proton therapy (HR=0.55), race rather than black and white, Hispanic ethnicity, private insurance

type, people who earned more, patients in metro area, Charlson-Deyo Score equals to 0, previous year diagnosis, patients with well differentiated Grade, AJCC Analytic Stage I group, patients with PSA less than 10, patients with Gleason Score 2-7, circle distance>30 mile and lower age has less hazard death. The table is provided in Appendix Table4.

3.2 Estimating propensity scores

The propensity score is the probability of treatment assignment conditional on observed baseline characteristics. In this case study, we apply 3 estimation models to get the propensity score: Logistic Regression Model, Logistic Regression Model with interactions and polynomial and Generalized Boosted Models (GBM). For GBM, we employ the twang package provided by Doctor McCaffery for calculation (Greg Ridgeway, 2017). And then give out the distributions of PS value and logit-PS value are showed below (Figure 2 – Figure 4), where $logitPS = l(X_i) = \log\{[1 - p(X_i)]/p(X_i)\}$. We see that there is a thin overlap about the distribution of PS or logit(PS) between two treatment group, which also indicates a substantial study population background difference between the two groups.

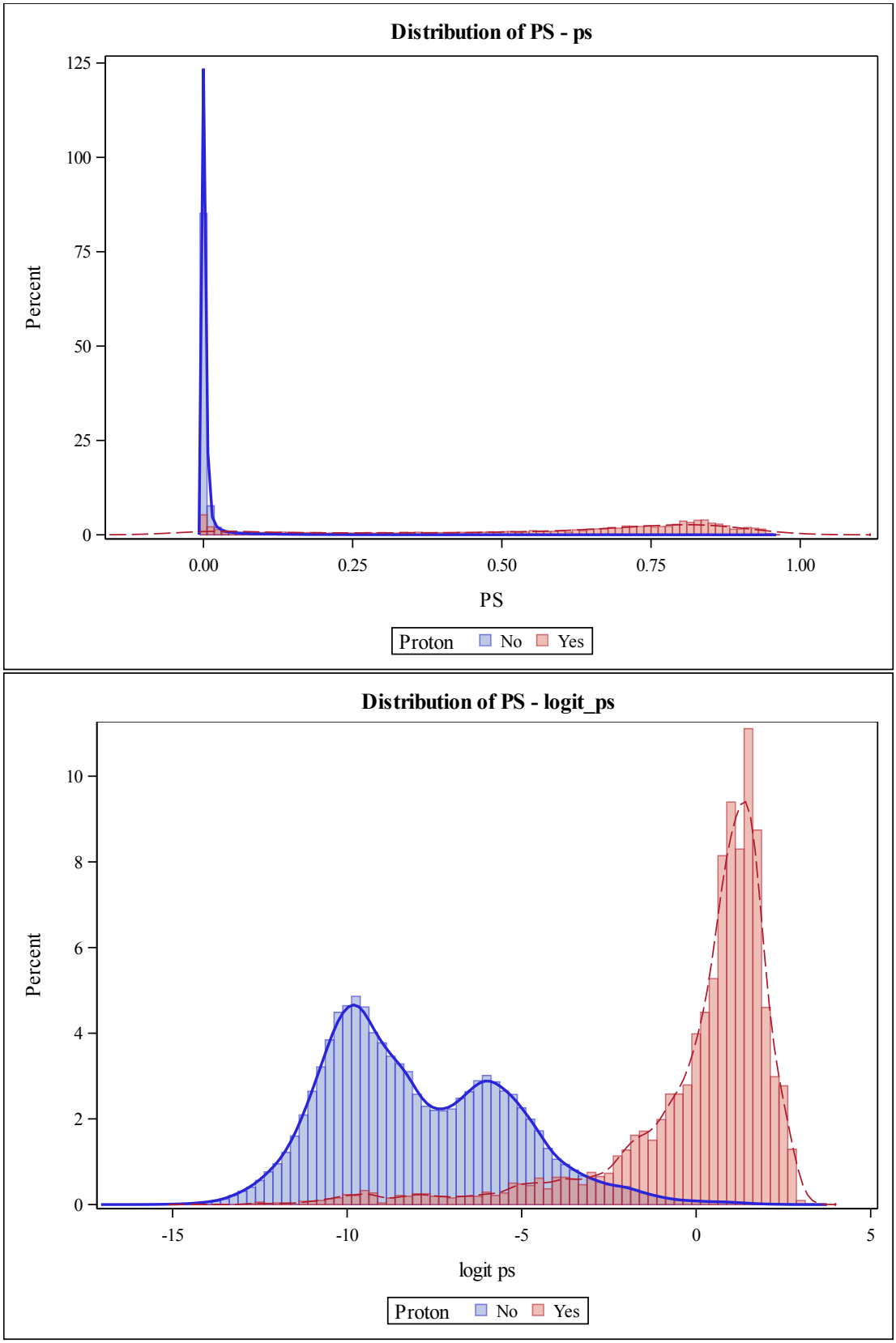


Figure 2: LOGREG - distribution of PS/ logit-PS Proton vs non-Proton.

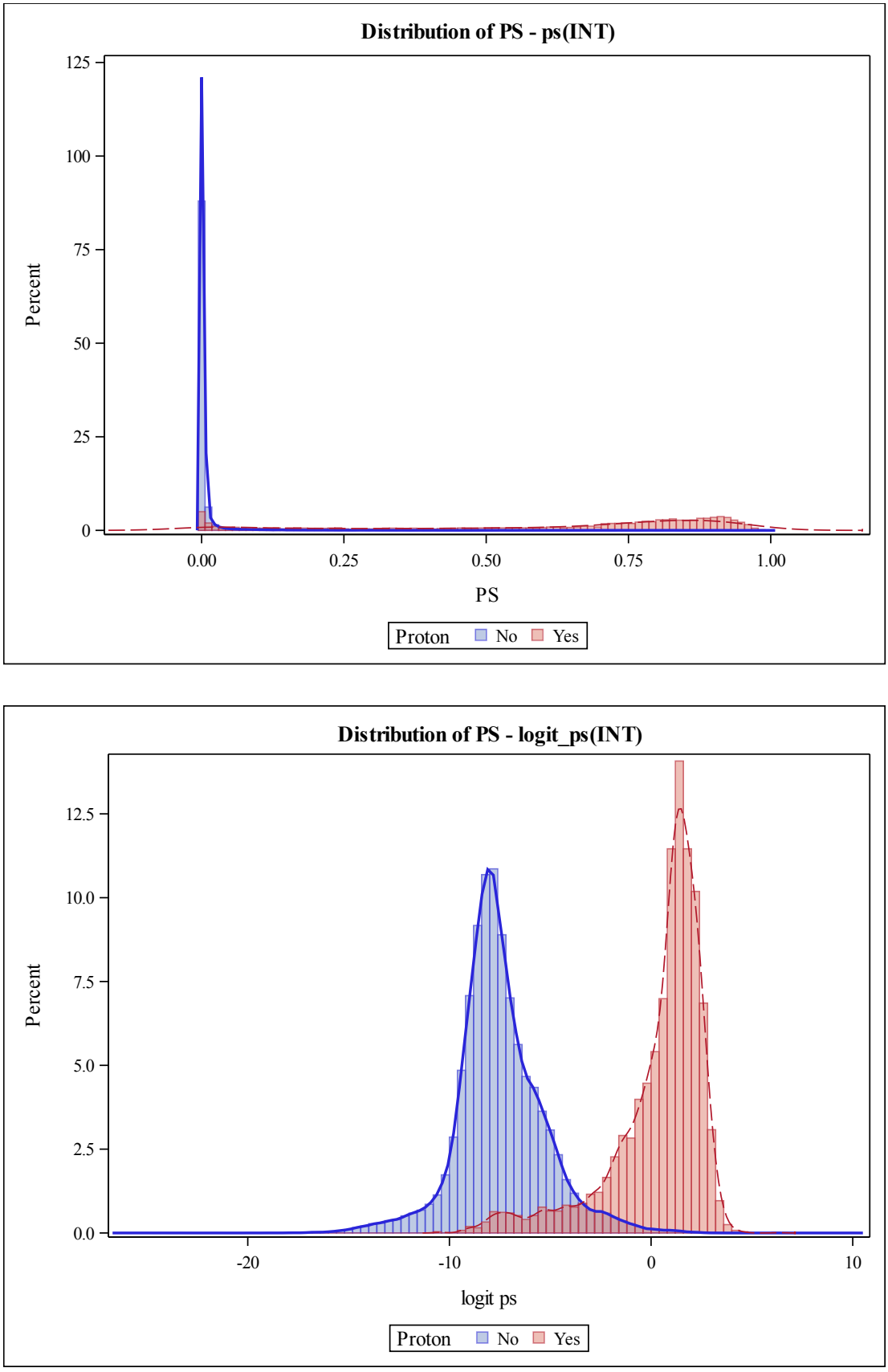


Figure 3: LOGREG(INT) - distribution of PS/ logit-PS Proton vs non-Proton.

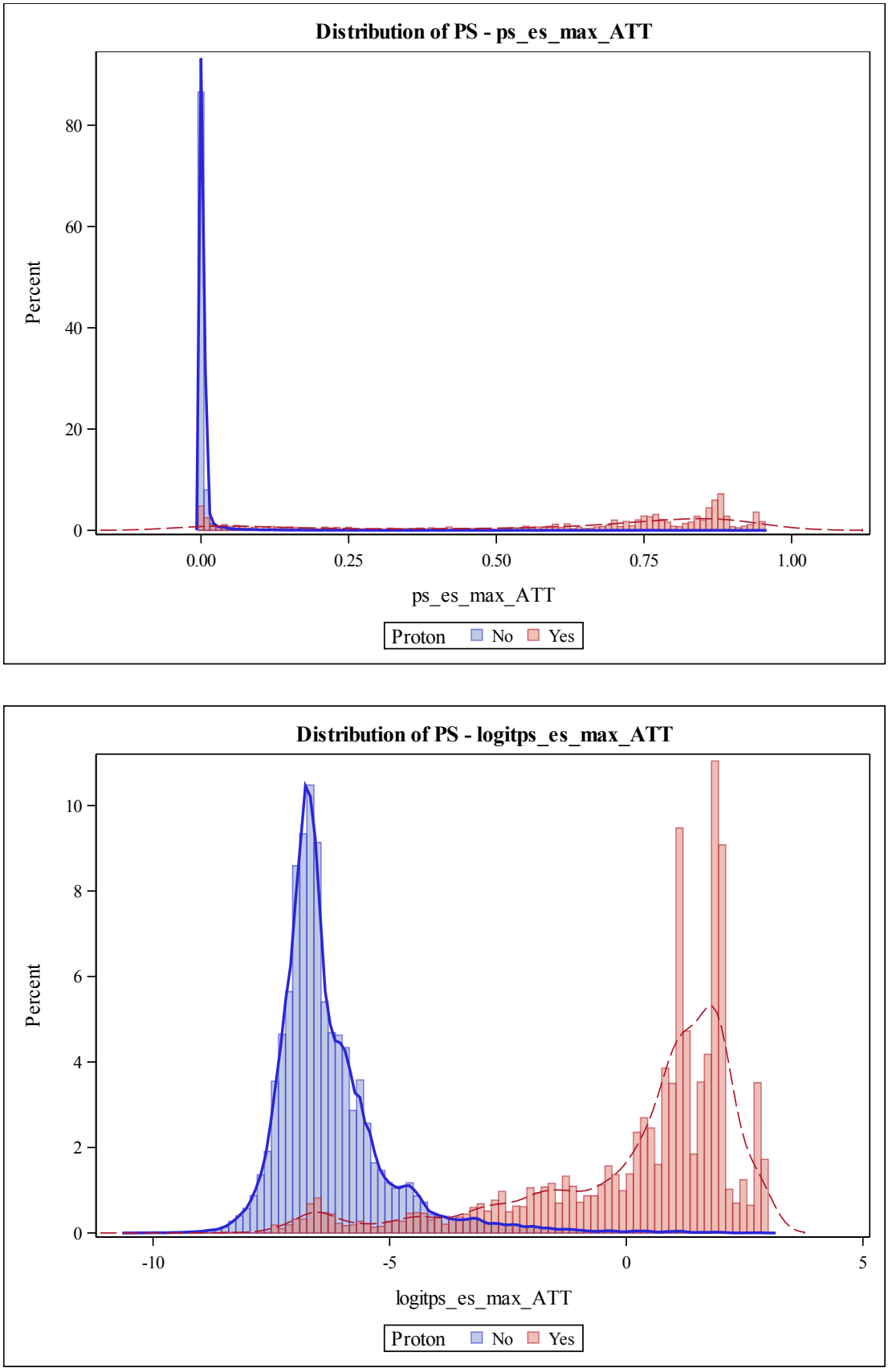


Figure 4: GBM - distribution of PS/ logit-PS Proton vs non-Proton.

3.3 PS Matching

After creating propensity scores value, the next step is to choose how to apply the propensity scores to compare the treatment and comparison groups. We are more interested in the effectiveness of Proton treatment method, so we use ATT because it measures the relative effectiveness of programs Proton and Non-Proton on the population receiving proton.

In this article, simple greedy matching method and 1-1, 1-N caliper matching method are used to test the difference. We apply 1-N caliper matching method here is because it is particularly useful when the size of the groups differs largely in the original sample, which appears significantly that the proton therapy group and Non-proton therapy group have really different size number. The 1-N caliper matching method can increase the overall size of the matched sample and thus efficiency in the estimation of treatment effect.

Propensity Score is a continuous variable, so we use ADS-formula 2 to get the value of ASD for PS in the 3 method groups: Logistic Regression Model, Logistic Regression Model with interactions and polynomial and Generalized Boosted Models (GBM). As article mentioned before, we prefer to set the caliper value equals to 0.2 standard deviation of logit (PS). For LOGREG, caliper value = 0.63209; For LOGREG(INT), caliper value = 0.69319; For GBM, caliper value = 0.71048.

3.4 Checking balance on the covariates before and after matching

From the figures2-4, for all the 3 methods, we get the propensity score for proton and Non-proton treatments are extremely distributed far from each other. The propensity score for non-proton is very low while the distribution of proton propensity score is relatively high near 0.9. The small area of common support indicates that the observed effect would be only valid for a small subgroup of the population. In addition to overlapping, the PS distributions are not similar between the treatment and comparison groups.

Absolute standard differences (ASDs) were used as a balance statistic for individual covariates, where an ASD below 0.10 is desirable for all variables.

The figures (Table 2 and Table 3) below show the dramatic changes of ASD before and after matching for the three applicable methods.

3.4.1 Greedy Matching

Table 2: Before Matching Procedure

PS Estimation Method	ASD-MAX	ASD-MIN	ASD-MEAN	ASD-STD	Number Before Matching	Matched HR with 95% CI (non-proton vs. proton)
LOGREG	2.1917	0.0368	0.5734	0.7089	213319(208126: 5193)	3.11 (2.78-3.48)
LOGREG(INT^2)	2.1917	0.0368	0.5734	0.7089	213319(208126: 5193)	3.11 (2.78-3.48)
GBM(ATT-es.max)	2.2517	0.0397	0.5761	0.7126	261490(255630: 261490)	3.07 (2.76-3.40)

Table 3: Greedy 5-1 digits Matching Method

PS Estimation Method	ASD-MAX	ASD-MIN	ASD-MEAN	ASD-STD	Number of Matching	Matched HR with 95% CI (non-proton vs. proton)
LOGREG	0.1064	0.0018	0.0337	0.0270	4900(2450:2450)	1.41 (1.17-1.70)
LOGREG(INT^2)	0.0830	0.0071	0.0349	0.0213	4596(2298:2298)	1.27(1.05-1.54)
GBM(ATT-es.max)	0.0787	0.0008	0.0288	0.0200	5626(2813:2813)	1.39 (1.17-1.64)

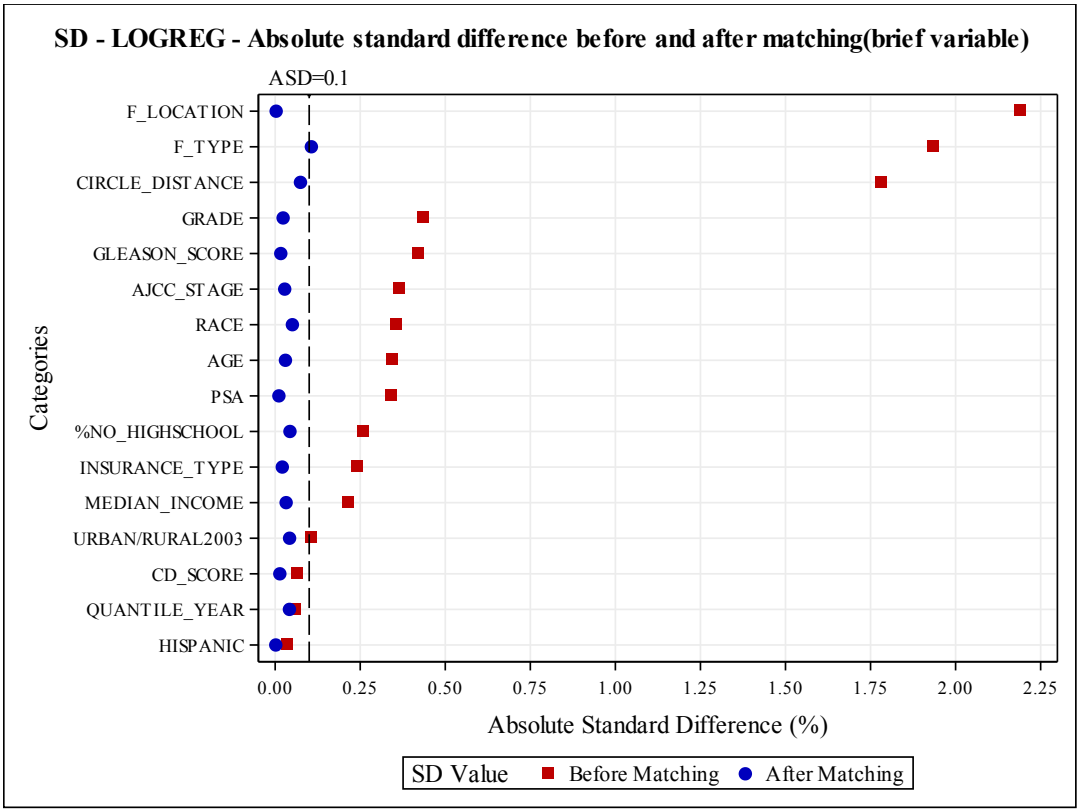


Figure 5: LOGREG – Absolute Standard Difference before and after Matching

Simple caliper matching for **Logistic Regression model** resulted in a matched sample of 4900 with 2450 in each group. Most of the covariates in these groups were well matched on the basis of ASDs below 0.1 after matching, except for covariate "facility type". Comparisons between treatment groups (Proton, Non-Proton) could be made using Hazard-Ratio. The value corresponding to the HR is 1.41 with p-value<0.001 and 95% CI is equal to (1.17, 1.71), indicating that non-proton group turns to have worse long-term survival comparing to proton patients.

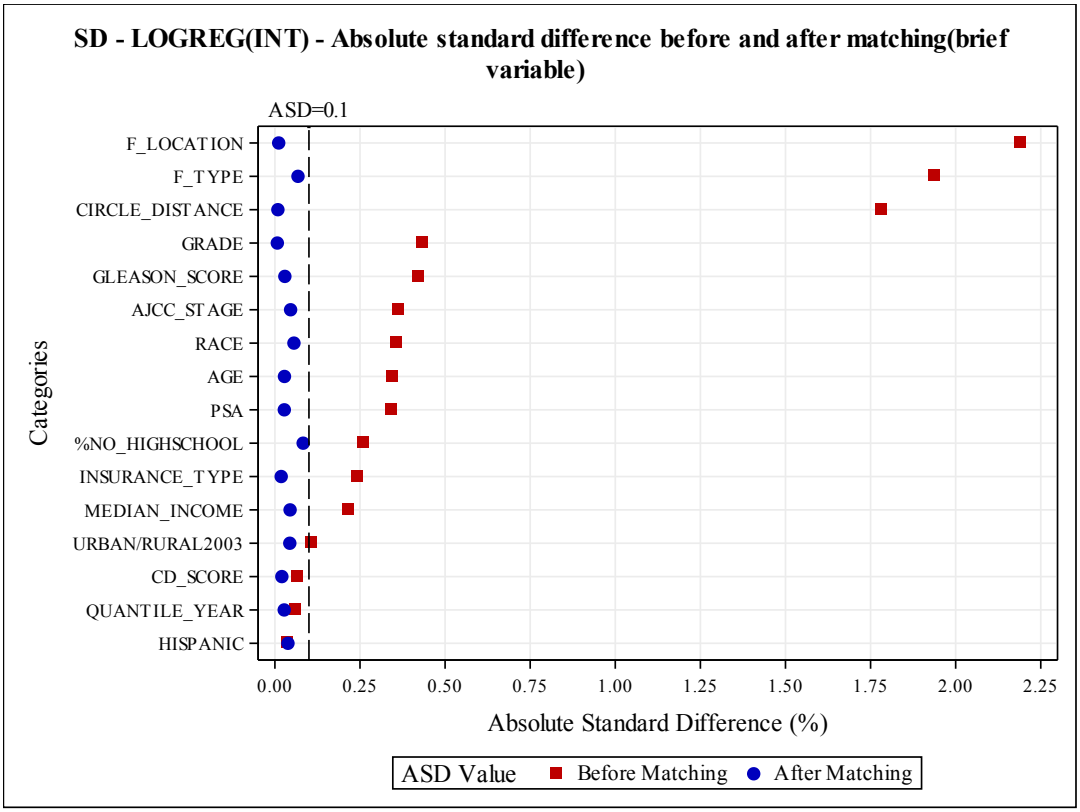


Figure 6: LOGREG(INT) – Absolute Standard Difference before and after Matching

Simple caliper matching for **Comprehensive Logistic Regression Model with all Two-way Interactions and Polynomial Terms** resulted in a matched sample of 4596 with 2298 in each group. All of the covariates in these groups were well matched on the basis of ASDs below 0.1. Comparisons between treatment groups (Proton, Non-Proton) could be made using Hazard-Ratio. The value corresponding to the HR is 1.27 with p-value=0.016 and 95% CI is equal to (1.05, 1.54), indicating that non-proton group turns to have worse long-term survival comparing to proton patients.

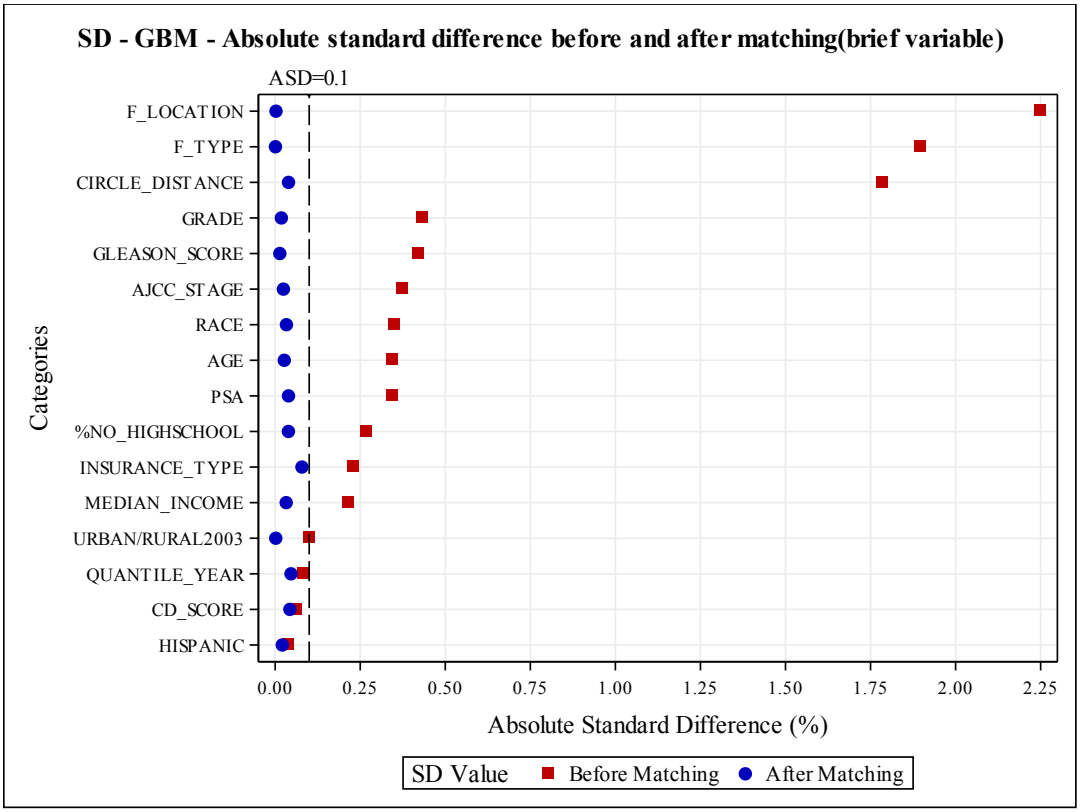


Figure 7: GBM – Absolute Standard Difference before and after Matching

Simple caliper matching for **GBM** resulted in a matched sample of 5626 with 2813 in each group. All of the covariates in these groups were well matched on the basis of ASDs below 0.1. Comparisons between treatment groups (Proton, Non-Proton) could be made using Hazard-Ratio. The value corresponding to the HR is 1.39 with p-value<0.001 and 95% CI is equal to (1.17, 1.64), indicating that non-proton group turns to have worse long-term survival comparing to proton patients.

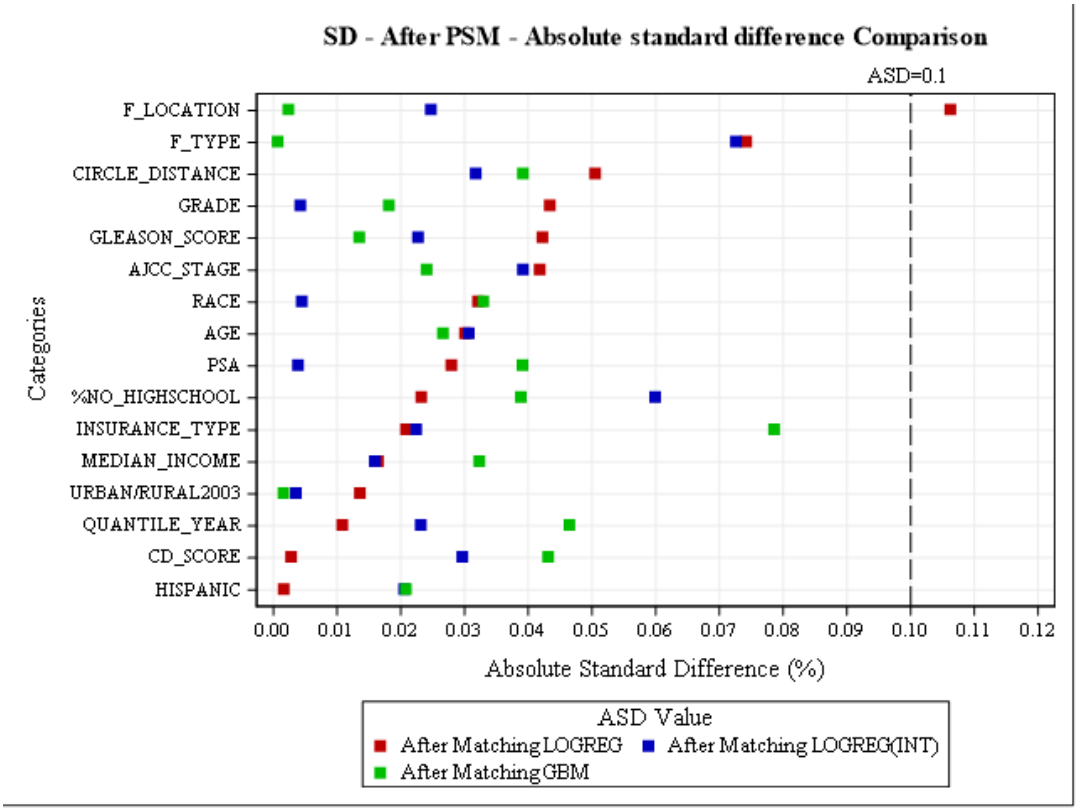


Figure 8: Absolute Standard Difference After Matching

Here is an image for absolute standard difference after simple Caliper PS Matching of 3 methods. We set the Range of absolute standard difference from 0 to 0.12, and sort the 16 variables by the value of ASD value LOGREG method. Most of the covariates in these 3 groups were well matched on the basis of ASDs below 0.1, except for covariate "facility type" of LOGREG method. We could also find that most covariates ASD for LOGREG(INT) and GBM is less than LOGREG method. For instance, for covariate QUANTILE_YEAR, CD_SCORE, RACE, GRADE, GLEASON_SCORE, URBAN_RURAL2003, LOGRAG method has the largest value of ASD. But sometimes LOGREG(INT) and GBM method could change the ASD to a larger value as well. For INSTURANCE_TYPE, ASD value is much larger for GBM (around 0.08) compare with LOGREG (around 0.003). For AJCC_STAGE, ASD value is Larger for LOGREG(INT) (around 0.06) compare with LOGREG (around 0.043). In general, LOGREG(INT) and GBM method have better ASD value after covariance balance check compare with LOGREG method.

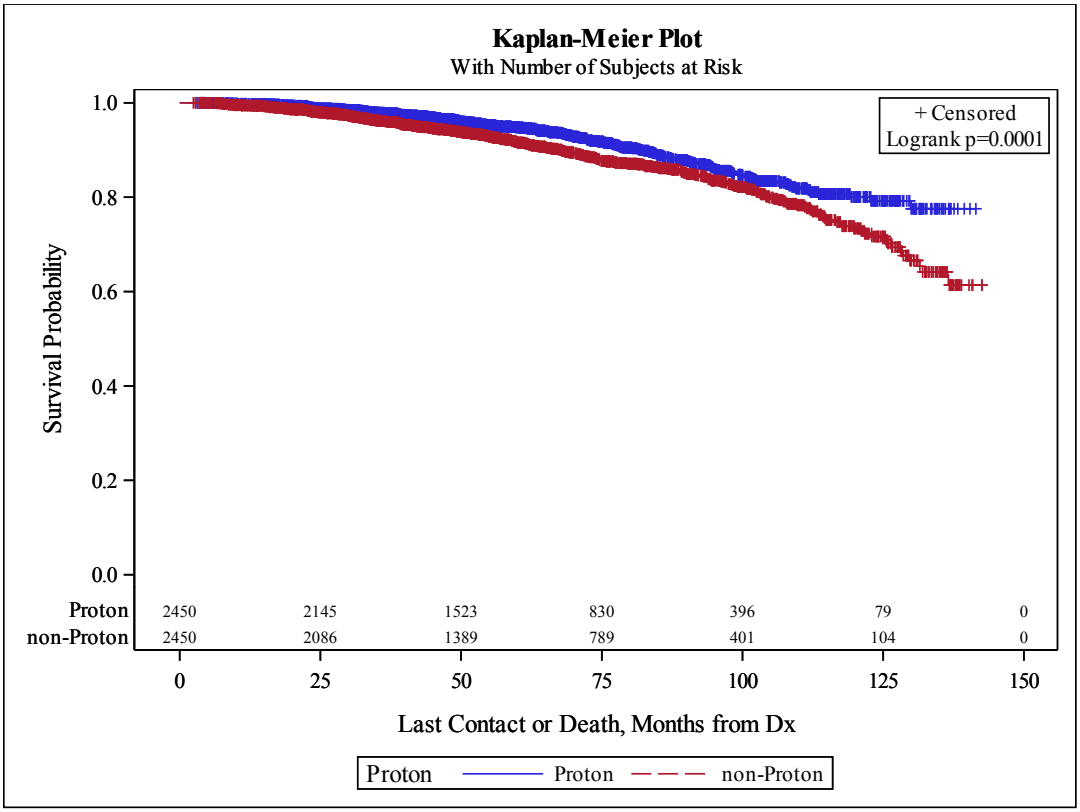


Figure 9: Kaplan–Meier estimates by treatment group in propensity score matched sample - LOGREG

Proton	No. of Subject	Event	Censored	Median Survival (95% CI)	60 Month Survival	120 Month Survival
Proton	2450	193 (8%)	2257 (92%)	NA (NA, NA)	94.8% (93.7%, 95.7%)	80.1% (76.5%, 83.1%)
non-Proton	2450	267 (11%)	2183 (89%)	NA (NA, NA)	91.6% (90.2%, 92.9%)	73.4% (69.3%, 77.0%)

Figure 9 provides **Main-effect Logistic Regression Model** Kaplan–Meier OS estimates by treatment group. Statistically significant overall differences are observed ($p<0.001$), survival for Proton therapy is better and the estimated OS is above the median follow up. And the estimated OS at 60 months (5 years) for Proton therapy and Non-proton therapy patients was 94.8% and 91.6%, respectively; the estimated OS at 120 months (10 years) for Proton therapy and Non-proton therapy patients was 80.1% and 73.4%, respectively.

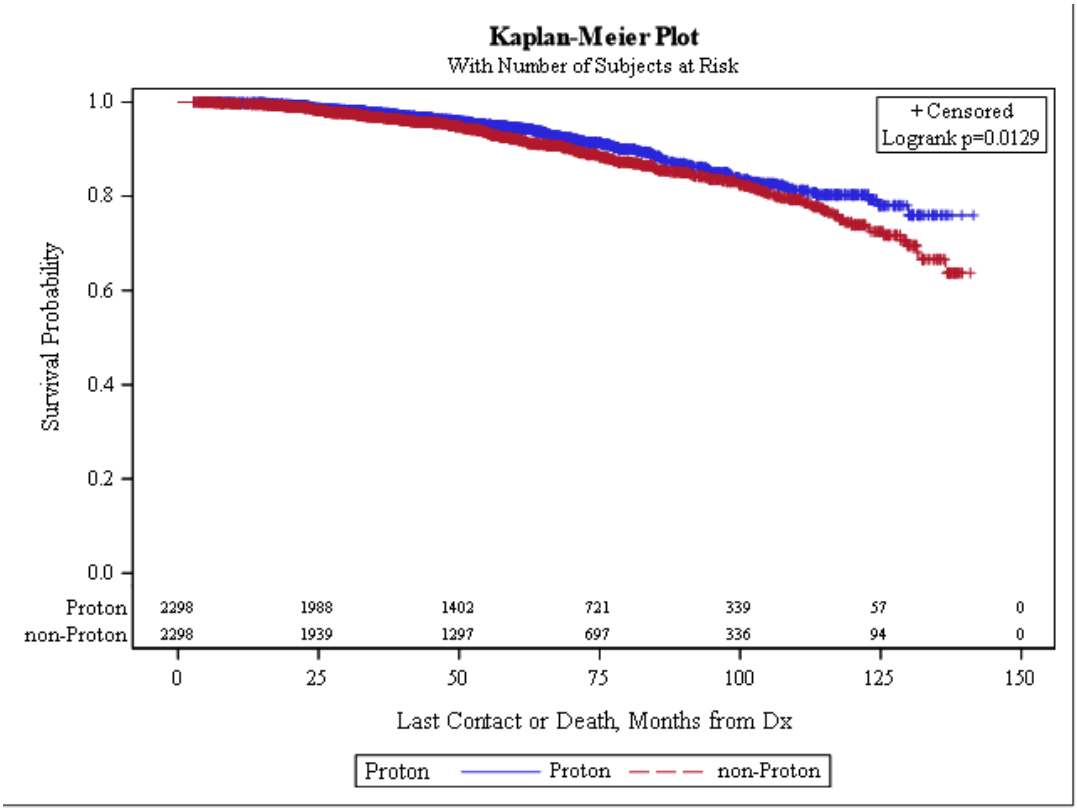


Figure 10: Kaplan–Meier estimates by treatment group in propensity score matched sample – LOGREG(INT)

Proton	No. of Subject	Event	Censored	Median Survival (95% CI)	60 Month Survival	120 Month Survival
Proton	2298	183 (8%)	2115 (92%)	NA (NA, NA)	94.5% (93.3%, 95.5%)	80.3% (76.7%, 83.3%)
non-Proton	2298	228 (10%)	2070 (90%)	NA (NA, NA)	92.2% (90.7%, 93.4%)	73.9% (69.5%, 77.8%)

Figure 10 provides **Logistic Regression model with interactions and polynomial** Kaplan–Meier OS estimates by treatment group. Statistically significant overall differences are observed ($p<0.001$), survival for Proton therapy is better and the estimated OS is above the median follow up. And the estimated OS at 60 months (5 years) for Proton therapy and Non-proton therapy patients was 94.5% and 92.2%, respectively; the estimated OS at 120 months (10 years) for Proton therapy and Non-proton therapy patients was 80.3% and 73.9%, respectively.

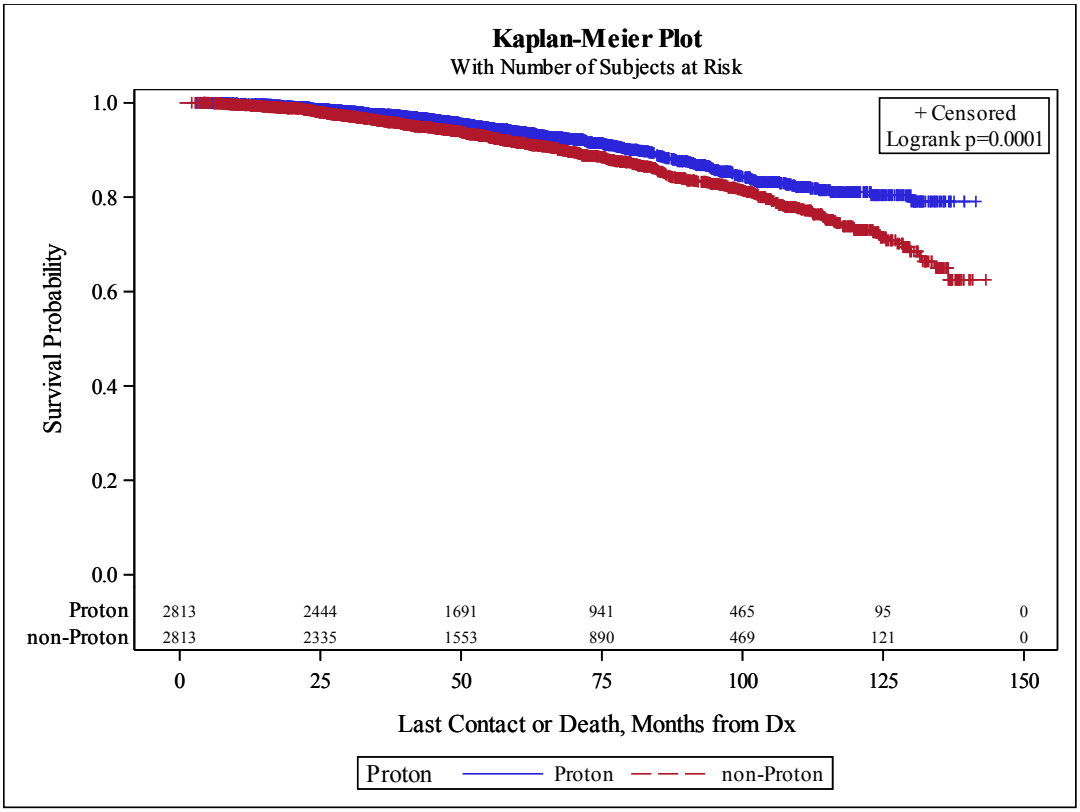


Table 11: Kaplan–Meier estimates by treatment group in propensity score matched sample – GBM

Proton	No. of Subject	Event	Censored	Median Survival (95% CI)	60 Month Survival	120 Month Survival
Proton	2813	227 (8%)	2586 (92%)	NA (NA, NA)	93.9% (92.7%, 94.9%)	81.1% (78.1%, 83.7%)
non-Proton	2813	305 (11%)	2508 (89%)	NA (NA, NA)	91.6% (90.3%, 92.8%)	73.1% (69.4%, 76.4%)

Figure11 provides **GBM** Kaplan–Meier OS estimates by treatment group. Statistically significant overall differences are observed ($p<0.001$), survival for Proton therapy is better and the estimated OS is above the median follow up. And the estimated OS at 60 months (5 years) for Proton therapy and Non-proton therapy patients was 93.9% and 91.6%, respectively; the estimated OS at 120 months (10 years) for Proton therapy and Non-proton therapy patients was 81.1% and 73.1%, respectively.

3.4.2 1-1 to 1-N Caliper Matching

In this part, we apply the caliper matching to these 3 methods and stop running after N=5. From the 3 tables below, we could find that with the increase of N value, the ASD is increasing as well. But all the matched HR are statistically significant.

Table 4: 1-N Caliper Matching for LOGREG

Matching Method(P:NP)	ASD-MAX	ASD-MIN	ASD-MEAN	ASD-STD	Number of Matching	Matched HR with 95% CI
LOGREG(Greedy)	0.1064	0.0018	0.0337	0.0270	4900(2450:2450)	1.41 (1.17-1.70)
LOGREG(1:1)	0.0957	0.0036	0.0349	0.0256	5178(2589: 2589)	1.50 (1.25-1.80)
LOGREG(1:2)	0.1881	0.0002	0.0409	0.0537	6436(2453: 3983)	1.63 (1.37-1.93)
LOGREG(1:3)	0.2467	0.0015	0.0508	0.0780	7550(2381: 5169)	1.66 (1.40-1.95)
LOGREG(1:4)	0.2884	0.0023	0.0646	0.0994	8655(2392: 6263)	1.56 (1.34-1.83)
LOGREG(1:5)	0.3188	0.0074	0.0769	0.1101	9643(2374: 7269)	1.59 (1.36-1.86)

Table 5: 1-N Caliper Matching for LOGREG(INT)

Matching Method(P:NP)	ASD-MAX	ASD-MIN	ASD-MEAN	ASD-STD	Number of Matching	Matched HR with 95% CI
LOGREG(INT^2) (Greedy)	0.0830	0.0071	0.0349	0.0213	4596(2298:2298)	1.27(1.05-1.54)
LOGREG(INT^2) (1:1)	0.0862	0.0006	0.0341	0.0223	4892 (2446: 2446)	1.45 (1.19-1.76)
LOGREG(INT^2) (1:2)	0.1507	0.0022	0.0458	0.0506	6109 (2308: 3801)	1.59 (1.33-1.91)
LOGREG(INT^2) (1:3)	0.2180	0.0038	0.0579	0.0777	7206 (2275: 4931)	1.70 (1.43-2.03)
LOGREG(INT^2) (1:4)	0.2755	0.0077	0.0738	0.0963	8165 (2253: 5912)	1.70 (1.44-2.02)
LOGREG(INT^2) (1:5)	0.3237	0.0032	0.0789	0.1130	9077 (2233: 6844)	1.74 (1.47-2.05)

Table 6: 1-N Caliper Matching for GBM

Matching Method(P:NP)	ASD-MAX	ASD-MIN	ASD-MEAN	ASD-STD	Number of Matching	Matched HR with 95% CI
GBM(Greedy)	0.0787	0.0008	0.0288	0.0200	5626(2813:2813)	1.39 (1.17-1.64)
GBM (1:1)	0.0686	0.0000	0.0256	0.0212	5690(2845: 2845)	1.31 (1.10-1.56)
GBM (1:2)	0.1638	0.0117	0.0512	0.0367	7118(2697: 4421)	1.46 (1.24-1.71)
GBM (1:3)	0.2930	0.0016	0.0595	0.0740	8431(2650:5781)	1.42 (1.22-1.67)
GBM (1:4)	0.3717	0.0017	0.0715	0.0988	9600(2642:6958)	1.48 (1.27-1.72)
GBM (1:5)	0.4098	0.0008	0.0793	0.1114	10647(2603:8044)	1.50 (1.29-1.75)

4. DISSUSSION

This article aimed to provide practical guidance for researchers and practitioners on how to utilize propensity score method when estimating causal treatment effects of two treatment conditions. In estimating the multiple treatment propensity score, a powerful machine learning method, GBM, was used to obtain robust propensity score with better balance properties than a simple parametric model (namely the multinomial logistic) did.

As shown in our example in case study, use of matching can improve imbalances when interest lies in comparing more than two treatment programs, this allows researchers to make more robust inferences when estimating treatment effects. The 2 treatment groups survival of all the 3 methods are significantly different from each other for both before and after matching. By comparing the significant change for ASD before and after PS matching, we get the importance of dealing with balancing procedure in the multivariate models. In addition to that, we also find some difference between the 3 methods: The value of ADS after matching for GBM and Logistic Regression method with polynomial and interaction became relatively small contrast to the ordinary Logistic Regression. Which means we could apply both of the 2 methods as accuracy model for our future application and usage.

For the propensity score weighting method, sometimes the inability to achieve balance is particularly likely when the number of pretreatment variables is very large relative to the overall sample size. As we saw in our example, the 2 treatment groups are disparate, weighting can be very inefficient because most of the weight is applied to very few cases and most of the sample receives very little weight, making precise estimation of the causal effects difficult. But there're some difference of the 2 methods of weighting: The ATEs are more likely to be of interest compared with ATTs if every treatment potentially might be offered to every member of the population. Conversely, if the research question focuses on the effectiveness of one treatment program, then the ATT would be of interest because it

measures the relative effectiveness of programs t_0 and t_1 on the population receiving program t_1. So, in our case study, we use ATT because it measures the relative effectiveness of programs Proton and Non-Proton on the population receiving proton.

The methods discussed in this article do have their limitations as well. For instance, unmeasured confounder may still bias the results. In particular, the methods only remove confounding by observed variables, but no unknown or unmeasured confounders. If there are unmeasured variables that predict outcomes and differ among treatment groups, then the estimates can be biased as well. This limitation, however, is not specific to the 3 methods we present; indeed, all causal modeling strategies that use observational study data must contend with this limitation in one way or another.

There also exist limitations and Risks of Propensity Score Application. Future research should more carefully explore various estimation methods for obtaining propensity score when there are more than two treatment conditions. From the results given above, GBM has outperformed the use of multinomial logistic regression when we have tried to balance more than two treatment groups on pretreatment characteristics.

Another limitation of propensity analysis method in this article is: there're loss sample sizes after matching procedure. When doing matching procedure of the case study, there's a very large loss of sample size for comparison group, since the limited number of subjects in treatment group. Although we apply 1-N caliper matching to reduce the bias selection, the limitation of total sample size may still leave defect.

Bibliography

Austin, P. (2008). A critical appraisal of propensity-score matching in the medical literature between 1996 and 2003. *Stat Med*, 2037-2049.

Austin, P. (2011). Optimal caliper widths for propensity-score matching when estimating differences in means and differences in proportions in observational studies. *Pharm Stat*, 150-161.

Austin, P. C. (2011). An Introduction to Propensity Score Methods for Reducing the Effects of Confounding in Observational Studies. *Multivariate Behavioral Research*, 399-424.

Bilimoria KY, S. A. (2008). The National Cancer Data Base: A powerful initiative to improve cancer care in the United States. *Ann Surg Oncol*, 683-690.

Breiman, L. F. (1984). *Classification and Regression Trees.* Belmont, CA: Wadsworth.

Cochran, W., & Rubin, D. (1973). Controlling bias in observational studies: a review. *Sankhyā: Indian J Stat, Ser A*, 417-446.

Efron, & Tibshirani. (1993). An Introduction to the Bootstrap. *Chapman & Hall/CRC*.

Efstathiou JA, G. P. (2013). Proton beam therapy and localised prostate cancer: current status and controversies. *Br J Can- cer*, 1225-1230.

ES, W., PE, A., RG, F., & al., e. (2014). Proton beam ther- apy for localized prostate cancer 101: basics, controversies, and facts. *Rev Urol.*, 67-75.

G.W, I. (2004). Nonparametric estimation of average treatment effects under exogeneity: A review. *The Review of Economics and Statistics*, 4-29.

Gail, M., Krickeberg, K., Samet, J., Tsiatis, A., & Wong, W. (2002). *Statistics for Biology and Health.* Springer.

Greg Ridgeway, D. M. (2017). Toolkit for Weighting and Analysis of Nonequivalent Groups: A tutorial for the twang package.

Hirano K, I. G. (2003). Efficient estimation of average treatment effects using the estimated propensity score. . *Econometrica.*, 1161-1189.

Kewei Ming, P. R. (2001). A Note on Optimal Matching With Variable Controls Using the Assignment Algorithm. *Journal of Computational and Graphical Statistics*, 455-463.

LS, P. (2001). Reducing bias in a propensity score matched-pair sample using greedy matching techniques. *Paper 214-26 in Proceedings of the Twenty-Sixth Annual SAS Users Group International ConferenceCary, NCSAS Institute, Inc.*

Lunceford, J. K., & Davidian, M. (2004). Stratification and weighting via the propensity score in estimation of causal treatment effects: a comparative study. *Statistics in Medicine*.

McCaffrey DF, R. G. (2004). Propensity Score Estimation with Boosted Regression for Evaluating Causal Effects in Observational Studies. *Psychological Methods*, 403-25.

Nguyen, P., Gu, X., & Lipsitz, S. (2001). Cost implications of the rapid adoption of newer technologies for treating prostate cancer. *J Clin Oncol*, 1517-1524.

Normand, S. L. (2001). Validating recommendations for coronary angiography following an acute myocardial infarction in the elderly: A matched analysis using propensity scores. *Journal of Clinical Epidemiology*, 387–398.

P., H. (1986). Statistics and causal inference. Journal of the American Statistical Association. *Journal of the American Statistical Association*, 945-960.

Parsons, L. S. (2001). Reducing Bias in a Propensity Score Matched-Pair Sample Using Greedy Matching Techniques. *Paper 214-26 Proceedings of the Twenty-Sixth Annual SAS Users Group International ConferenceCary, NCSAS Institute, Inc.*

Paul, R. R., & Donald, B. R. (1983). The Central Role of the Propensity Score in Observational Studies for Causal Effects. *Biometrika*, 41-55.

RALPH B. D'AGOSTINO, J. (1998). Propensity score methods for bias reduction in the comparison of a treatment to a non-randomized control group. *STATISTICS IN MEDICINE*, 2265-2281.

Reshma, J. (2014). Considerations for Observational Research using Large Datasets in Radiation Oncology. *American Society of Radiation Oncology*, 11-24.

Robins JM, H. M. (2000). Marginal structural models and causal inference in epidemiology. *Epidemiology*, 550-560.

Rosenbaum, P. (2002). *Observational studies (2nd ed.).* New York, NY: Springer Verlag.

Shah A, P. J. (2013). Physician evaluation of internet health information on proton therapy for prostate cancer. *Int J Radiat Oncol Biol Phys.*, 173-177.

Stuart, E. A. (2010). Matching methods for causal inference: A review and a look forward. *Statistical Science*, 1-21.

Stuart, E. A., & Rubin, D. B. (2008). Best practices in quasi-experimental designs: Matching methods for causal inferences. *Best practices in quantitative methods*, 155-176.

Westreich, D., Lessler, J., & Funk, M. J. (2010). Propensity score estimation: machine learning and classification methods as alternatives to logistic regression. *Journal of Clinical Epidemiology*, 826-833.

Wisenbaugh, E., Andrews, P., & Ferrigni, R. (2014). Proton beam therapy for localized prostate cancer 101: basics, controversies, and facts. *Rev Urol*, 67-75.

Wooldridge, J. (2002). *Econometric of Analysis of Cross Section and Panel Data.* Cambridge: Mit Press.

Yuan, L., Dana, N., & Joseph, L. (2013). Propensity Score Matching for Multiple Treatment Comparisons in Observational Studies. *The 59th World Statistics Congress proceeding*.