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:

_____ _____

Luer Zhong Date

Evaluation of Propensity Score Matching Techniques on Overall Survival using the National Cancer Data Base

By

Luer Zhong
MSPH

Biostatistics and Bioinformatics Department

Jeffrey M. Switchenko, PhD
Committee Chair

Yuan Liu, PhD
Committee Member

Evaluation of Propensity Score Matching Techniques on Overall Survival using the National Cancer Data Base

By

Luer Zhong

B.A., Tianjin University, 2017

Thesis Committee Chair: Jeffrey M. Switchenko, PhD

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 Science in Public Health
in
Biostatistics and Bioinformatics Department
2019

Abstract

Evaluation of Propensity Score Matching Techniques on Overall Survival using the National Cancer Data Base

By Luer Zhong

Background: Observational studies are often used to mimic randomized controlled trial. Propensity scores can help facilitate since they can balance the distribution of baseline covariates of the treated and the untreated through matching. However, it is not clear whether a 1:1 match can be improved by increasing the number of controls (N) matched to cases. In this analysis, we calculated propensity score, and matched cases with controls using the greedy matching algorithm. Furthermore, we determined the differences between one to one and one to N greedy matching, along with different matching digits, and determined which matching performed better for two National Cancer Data Base (NCDB) files.

Methods: We calculated the propensity score by using the logistic regression model, and performed 1 to 1, …, 1 to 5 greedy matching across HPV status, with 5-to-1 digit and 5-to-2 digit matching, on both larynx and hypopharynx cancer datasets from NCDB. Overall survival was the clinical outcome, and match rate and standardized difference were utilized to determine which approach performed better. For the survival outcome, Kaplan-Meier survival curves, stratified log-rank tests and hazard ratios with 95% confidence intervals from the Cox proportional hazard model were reported.

Results: The number of matched HPV positive patients for 5-to-2 digit matching is smaller than that of 5-to-1 digit matching. Widths of the hazard ratio confidence interval for 5-to-1 digit matching are generally narrower than 5-to-2 digit matching. There are almost no standardized differences that are greater than 0.1 after N is bigger than 2, except for 5-to-2 digit matching on larynx cancer stage 1&2.

Conclusion: This paper concludes that as the matching ratio of the case and control changes from 1 to 1, to 1 to 5, the variable balancing is better. Better variable balancing and higher match rates exist when using 5-to-1 digit matching instead of 5-to-2 digit matching. If a dataset has the capacity to allow 1 to 3, 1 to 4, or 1 to 5 matching, it is recommended to increase the matching ratio to achieve better balance across baseline characteristics.

Keywords: Observational study, propensity score, greedy matching, survival outcome, balance check

Evaluation of Propensity Score Matching Techniques on Overall Survival using the National Cancer Data Base

By

Luer Zhong

B.A., Tianjin University, 2017

Thesis Committee Chair: Jeffrey M. Switchenko, Doctor

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 Science in Public Health
in
Biostatistics and Bioinformatics Department
2019

TABLE OF CONTENTS

INTRODUCTION

Randomized controlled trials (RCTs) are considered the gold standard approach to assess the effects of treatments, interventions and exposures on outcomes. Randomization makes sure that the treatment assignment is not confounded with baseline covariates, so the effects of treatments, interventions and exposures on outcomes can be estimated directly by comparing the treated group and the untreated group ***[1].*** The average treatment effect (ATE) is the average effect of moving an entire population from the untreated group to the treated group, and the average effect of treatment for the treated (ATT) is the average effect of those who will receive treatment eventually. For RCTs, ATE is equal to ATT ***[2].*** For an observational study, there are often systematic differences in baseline covariates between the treated group and untreated group, because the selection of treatment is influenced by subject characteristics ***[2].*** Thus, ATE is not equal to ATT for observational studies, and the effects on outcomes cannot be estimated directly by comparing the treated group and the untreated group. But in reality, it is not always possible to conduct a randomized controlled trial due to many practical reasons. Recently, observational studies are more and more often used to mimic randomized controlled trials. The propensity score can be used to estimate the average treatment effect, so it can allow one to mimic a randomized controlled trial through an observational study.

The definition of the propensity score is the probability of treatment assignment conditional on observed baseline covariates ***[3].*** The propensity score balances the distribution of baseline covariates of the treated and the untreated. Thus, the distribution of observed baseline covariates for the treated and the untreated will be similar with similar propensity scores. For randomized controlled trials, the propensity score is known and designed, but for observational studies, the propensity score is not known and can be estimated from the data. There are many ways to estimate the propensity score, such as logistic regression, random forests**,** and boosting ***[5].*** Logistic regression is the method that is most commonly used, and the estimated score is

derived from the fitted regression model. There are several ways to use propensity score to estimate the effect of treatments on outcomes, including propensity score matching, stratification on propensity score, inverse probability of treatment weighting using the propensity score, and covariate adjusting using the propensity score ***[3][6].***

Propensity score matching is a useful method to eliminate bias when estimating the effect of treatments on outcomes. It works by constructing pairs of a treated subject and an untreated subject which share similar propensity scores. After forming the matched sample, the effect of treatment can be directly estimated by comparing the outcomes of the treated and the untreated subjects within the sample. Thus, propensity score matching can be used to estimate the average effect of treatment for the treated (ATT) ***[8]***. How one can estimate the ATT depends on the type of outcomes. For continuous outcomes, the ATT can be estimated by the difference of the mean outcomes for the treated and the untreated subjects within the sample ***[3].*** For dichotomous outcomes, the ATT can be estimated by the difference of the proportion of subjects experiencing the event in the treated and the untreated subjects within the sample, or the relative risk ***[3][9][10].*** To assess the difference of the matched pairs, one can use a paired t test for continuous outcomes and McNemar's test for dichotomous outcomes. Commonly used matching algorithms include greedy matching and optimal matching. In greedy matching, one treated subject is matched to an untreated subject which has the closest propensity score. Once the match is formed, the untreated subject will not be considered again ***[2]***. Types of greedy matching include nearest neighbor matching, which allows each treated subject to match with one untreated subject sharing the most similar propensity score ***[13].*** In this matching method, one treated subject has only one untreated match, and all treated subjects will be matched. Another is the nearest matching within a specified caliper distance ***[13].*** This method has a restriction on the absolute difference of propensity scores between the treated and untreated subjects, where the absolute difference must be below some threshold. These two kinds of greedy matching are both one to one matching, meaning one treated subject only has one matched untreated subject. A

greedy matching algorithm can also perform one to N matching, where one treated subject can be matched with more than one (N) untreated subjects. The algorithm makes the best match first, and then makes the second best match, and so on. If a treated subject does not have N matched untreated subjects, it will be removed, but the untreated subjects can still be matched with another treated subject ***[14].*** Also, the number of matching digits can make a difference as well. Matching digit means that the cases are matched to controls on a specific digit of the propensity score, and for those that are not matched, cases are matched to controls on lower digits. In general, higher matching digits picks up fewer but better matched pairs, while lower matching digits picks up more matches.

The propensity score matching can be applied to survival analysis. Kaplan-Meier survival curves can be estimated for both treated and untreated groups within the matched sample, which allows one to directly compare the survival outcomes for the treated and untreated groups. To test the equality of the survival curves within the matched sample, one can use the stratified log-rank test, because subjects in the sample are not independent to each other. One can then build a Cox proportional hazard model to estimate the relative change in the hazard of the outcome ***[15]***. Propensity score is a balancing score, so the distribution of all measured baseline covariates is expected to be the same between the treated and the untreated groups. Within the sample which is matched according to propensity score, if systematic differences still exist between the treated and the untreated groups, there could be many reasons such as the propensity score model is not adequate enough or the distribution overlap of a covariate between treated and untreated is poor. The standardized difference can be used to compare the means between the two groups in order to check covariate balance, which is an important step before we achieve a valid conclusion about the association between the treatment and the outcome.

In this paper, we are going to analyze the propensity score for an observational study from The National Cancer Data Base, 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 1500 commission on cancer-accredited facilities, and the data represents approximately 70% of newly diagnosed cancer cases nationwide and 34 million historical records. We will focus on larynx and hypopharynx NCDB data.

The purpose of this paper is to determine the differences between one to one greedy matching and one to N greedy matching, along with different matching digits, and conclude which propensity score matching performs better for two data sets: larynx cancer data set and hypopharynx cancer data set. Here, N could be 2, 3, 4 and above. We will compare rate of match, and the standardized differences between one to one matching and one to N matching. It is predicted that one to one matching can yield a better match rate, but one to N matching will yield more power to detect true differences in outcome due to larger sample sizes at the cost of worse balancing. The goal is to compare the propensity score matching approaches after evaluating the sacrifices and gains for each matching algorithm.

METHODS

Data Description

1. Data sets

Head and Neck cancer is a group of cancers, which start in the mouth, nose, throat, larynx, sinuses, or salivary glands. About 75% of head and neck cancer is caused by alcohol and tobacco, and the diagnosis is confirmed by tissue biopsy. The treatment of head and neck cancer includes a combination of surgery, radiation therapy, chemotherapy and targeted therapy ***[17].*** In 2015, more than 5.5 million people were affected by head and neck cancer globally (larynx 1.4 million, mouth 2.4 million, throat 1.7 million), and it has caused more than 379,000 deaths ***[18].*** In the United States, approximately 1% of the population are affected at some point in their life, and males are affected twice as often as females. The most common age of diagnosis is between 55 and 65 ***[7].*** Thus, the analysis of head and neck cancer is very important. We mainly focused on larynx cancer and hypopharynx cancer of head and neck cancers. Laryngeal cancer starts from the larynx, and it may occur on the vocal folds or on tissues above and below the true cords. Hypopharynx cancer includes the pyriform sinuses, the postcricoid area and the posterior pharyngeal wall.

Human papillomavirus (HPV) is involved in at least 25% of head and neck cancers, and HPV positive head and neck cancer patients have distinct molecular, clinical and demographic entity from HPV negative. HPV positive status is correlated with a significant superior outcome, indicating that such tumors should have a distinct management approach. Thus, HPV status has been used as prognostic biomarker for head and neck cancers. We used HPV status as the standard of grouping the cases and the controls, and we considered HPV positive individuals as the case group and HPV negative individuals as the control group. In this case, we could compare the relationship between the HPV status and the overall survival (OS).

In this paper, we used Head and Neck Cancer data sets from the National Cancer Data Base (NCDB) to analyze the impact of HPV status on patients' OS through propensity score approaches. NCDB is a clinical oncology database sourced from hospital registry data collected in more than 1,500 commission on cancer-accredited facilities, and the data represents approximately 70% of newly diagnosed cancer cases nationwide and 34 million historical records. It is jointly sponsored by the American College of Surgeons and the American Cancer Society. Information of data contained in NCDB includes basic demographics, treatment, recurrence, comorbidity status and survival.

2. Study population

For the larynx cancer data set, it consisted of 105,593 patients who had larynx cancer recorded in NCDB from 2004 to 2014. Before analyzing the dataset, we performed data inclusion and exclusion. First, we included all non-missing HPV status patients, invasive behavior, non-distant metastatic cases, non-missing vital status, year of diagnosis 2010 or later, histology code 8052 8070 8071 8072 8073 8074 8075 8076 8078 8083 8084, and sequence number 0 or 1. Then, we excluded metastatic patients, cases missing survival time, missing or zero overall clinical stage, and class of case equal to 0. Finally, 4,835 patients remained in the study dataset. HPV status generally was missing or not collected prior to 2010.

For hypopharynx cancer, it consisted of 22,705 patients. We did the same data inclusion and exclusion, and 1,085 patients remained in the study dataset. The different numbers of patients of larynx and hypopharynx cancer allowed us to analyze the impact of 1:1 vs. 1:N on hazard ratio/confidence interval estimates, and with respect to diagnostic checks such as standardized differences and percent of observations matched on different scales of sample size.

3. Variable selection

Before calculating the propensity score, we needed to decide which covariates were in the propensity score model. There were many possible selections including: all measured baseline covariates, all baseline covariates that are associated with treatment assignment, all covariates that affect the outcome, and all covariates that affect both treatment assignment and the outcome ***[2].*** For the study of larynx cancer, there were 124 variables in the original data set, and we selected 8 variables in the final propensity score model. 7 categorical variables: primary site, sex, Charlson-Deyo comorbidity score, receipt of chemotherapy, receipt of radiation, receipt of surgery, overall clinical stage group; and 1 numeric variable: age at diagnosis. Also, we stratified on early and late stage of cancer, considering stage 1 and 2 as early stage, and stage 3 and 4 as late stage. For the study of hypopharynx cancer, the covariate selection was the same as for larynx cancer, except that we did not include the variable primary site and we did not stratify on stage. The analysis for hypopharynx cancer was not stratified on stage due to the database's small sample size without stratification. For both cancers, we used HPV status as the outcome in the propensity score model. HPV positive was defined as: HPV positive for specified high risk type(s) other than types 16 or 18; HPV positive for high-risk type 18 without positive results for high-risk type 16 or positivity of high-risk type 16 unknown; HPV positive for high-risk type 16 without positive results for high-risk type 18 or positivity of high-risk type 18 unknown; HPV positive for high-risk types 16 AND 18; HPV positive for high-risk type(s), NOS, high-risk type(s) not stated. HPV negative was defined as HPV negative for high-risk and low-risk types; HPV negative for high-risk types with no mention of low-risk types; Negative, NOS; HPV positive for low-risk types only. We used PUF_CASE_ID as the patient id.

Statistical Analysis

Statistical analysis was performed using SAS 9.4 (SAS Institute Inc., Cary, NC), and we used SAS macros developed by Dr. Liu in the Department of Biostatistics and Bioinformatics, Emory University to calculate the propensity score, do survival analysis and perform one to one greedy matching ***[4].*** The significance level we chose was 0.05. Descriptive statistics for each

variable were reported. The objective of the analysis was to identify the relationship between HPV status and overall survival, while accounting for the above covariates. We used SAS macros which can perform 1 to 1 greedy matching as well as 1 to N (2,3,4, …) greedy matching on different digits, and the algorithm of the matching macros can be displayed as below (Figure 3) ***[14].***

For the matching digit, we used 5-to-1 digit matching and 5-to-2 digit matching. 5-to-1 digit matching meant that the cases were first matched to controls on 5 digits of the propensity score, and for these that were not matched, cases were matched to controls on 4 digits, until cases were matched on 1 digit. For 5-to-2 digit matching, it was the same as 5-to-1 digit matching except that cases were eventually matched on a minimum of 2 digits ***[11].***

We compared survival outcomes, match rate (percent of cases which are matched), and standardized difference to assess which combination of N and digits performs best for propensity score matching across datasets of different sample sizes and different ratios of cases to controls. For the survival outcome, Kaplan-Meier survival curves, stratified log-rank tests and hazard ratios with 95% confidence intervals were used. To generate hazard ratios, we fitted Cox proportional hazard model, which is a regression model commonly used for investigating the association between the survival time of patients and predictor variables. A general format of Cox proportional hazard model can be written as: $\mathrm{h(t)} = h_0(\mathrm{t}) \times \exp(b_1x_1 + b_2x_2 + \cdots + b_px_p)$. The values of $\exp(b_i)$ are called hazard ratios, and a hazard ratio above 1 indicates a covariate that is positively associated with the event probability, and thus negatively associated with the length of survival.

To assess the difference of the matched pairs, we used paired t-tests for continuous outcomes and McNemar's tests for dichotomous outcomes. For the standardized difference of continuous covariates, it is defined as:

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

where $\bar{x}_{treated}$ and $\bar{x}_{untreated}$ denote the sample mean of the covariate in the treated and the untreated groups, and $s^2_{treated}$ and $s^2_{untreated}$ denote the sample variance of the covariate in the treated and the untreated groups. For the standardized difference of dichotomous covariates, it is defined as:

$$d = \frac{(\bar{p}_{treated} - \bar{p}_{untreated})}{\sqrt{\frac{\bar{p}_{treated}(1 - \bar{p}_{treated}) + \bar{p}_{untreated}(1 - \bar{p}_{untreated})}{2}}}$$

where $\bar{p}_{treated}$ and $\bar{p}_{untreated}$ denote the prevalence or the mean of the dichotomous covariate in the treated and the untreated groups. It is commonly considered that a standardized difference which is less than 0.1 indicates a negligible difference in the mean or prevalence of a covariate between the two groups ***[16].***

RESULTS

Descriptive Statistics

1. Larynx Cancer

As described in the Methods, first we analyzed the descriptive statistics for the larynx cancer dataset, and the result is in Table 1.1.

There were 4,304 HPV negative patients (89%) and 531 HPV positive patients (11%), which we considered as the control and case groups, respectively. There were five primary sites of larynx cancer: Glottis (43.5%), Supraglottis (40.7%), Subglottis (1.4%), Larynx overlapping (3.5%), and Larynx NOS (10.8%). In the dataset, 1094 (22.6%) of all patients were female and 3,741 (77.4%) were male. There were 3,440 (71.1%) patients that had a Charlson-Deyo score of 0, meaning no comorbidity, and 1,395 (28.8%) patients had Charlson-Deyo score of 1 or higher, meaning comorbidity existed. 2,636 (55.9%) patients did not have chemotherapy and 2,078 (44.1%) patients had chemotherapy, and 1,022 (21.3%) patients did not have radiation and 3,785 (78.7%) patients had radiation. There were 3,034 (62.9%) patients who did not have surgery at the primary site of larynx cancer, and 1,792 (37.1%) patients had surgery at the primary site of larynx cancer. For larynx cancer, we stratified clinical stages into: stage 1&2, which included 2,179 (45.1%) patients, and stage 3&4, which included 2,656 (54.9%) patients. The only numeric variable we considered into the propensity score model is age, which had a mean of 62, and range of 19 to 90.

2. Hypopharynx Cancer

Next, we analyzed the descriptive statistics for hypopharynx cancer dataset, and the result is as below in Table 1.2.

There were 893 HPV negative patients (82.3%) and 192 HPV positive patients (17.7%), which we considered as the control and case groups. There were six primary sites of hypopharynx

cancer: Pyriform (48.5%), Postcricoid (2%), Hypopharyngeal aspect of aryepiglottic fold (7.5%), Posterior wall of hypopharynx (6.5%), Overlapping lesion of hypopharynx (5.1%) and Hypopharynx NOS (30.5%). In the dataset, 201 (18.5%) of all patients were female and 884 (81.5%) were male. There were 795 (73.3%) patients who had a Charlson-Deyo score of 0, meaning no comorbidity, and 290 (26.7%) patients had a Charlson-Deyo score of 1 or higher, meaning comorbidity existed. 280 (26.3%) patients did not have chemotherapy and 785 (73.7%) patients had chemotherapy, and 166 (15.4%) patients did not have radiation and 914 (84.6%) patients had radiation. There were 267 (24.6%) patients who had surgery at the primary site of hypopharynx cancer, and 817 (75.4%) patients who did not have surgery at the primary site of hypopharynx cancer. For hypopharynx cancer, we did not stratify clinical stages into two. The only numeric variable we considered into the propensity score model is age, which had a mean of 62, and range of 25 to 90.

Kaplan-Meier Analysis

Before calculating propensity score and greedy matching, we compared the relationship between HPV status and overall survival. We used a Kaplan-Meier plot to plot the survival curves for HPV positive and HPV negative patients. Figure 1.1 is the KM survival curves for larynx cancer, along with summary statistic tables. The median survival for HPV negative patients was 5.4, for HPV positive patients was 5.8, HPV positive patients had a higher survival curve than HPV negative patients, and the log-rank p-value was 0.0253, indicating that the survival curves within the matched sample were significantly different. Figure 1.2 is the KM survival curves for hypopharynx cancer, along with summary statistic tables. The median survival for HPV negative patients was 2.8, for HPV positive patients was NA, HPV positive patients had a higher survival curve than HPV negative patients, and the log-rank p-value was <0.0001, indicating that the survival curves within the matched sample were significantly different.

Propensity Score Calculation

In this paper, we used logistic regression model to estimate the HPV-specific propensity score. For larynx cancer, we stratified clinical stages into two: stage1&2 and stage3&4, and we separately calculated the propensity score for both stage groups. Figure 2.1 is the propensity score distribution for stage1&2, and Figure 2.2 is the propensity score distribution for stage3&4. For hypopharynx cancer, we did not stratify across clinical stage, so Figure 2.3 is the propensity score distribution for all stages of hypopharynx cancer. All the distributions of propensity scores of HPV positive patients and HPV negative patients are similar.

Overall Survival Analysis

After calculating the propensity score, the next step was to perform greedy matching. We were interested in comparing the matching rate of cases, the hazard ratio with a 95% confidence interval, standardized differences of one to one matching and one to N matching, along with different matching digits, and then evaluating the best matching approach. We also compared the effect of the same matching approach on two different datasets – larynx and hypopharynx cancer datasets to see the impact of different sample sizes and initial ratio of cases to controls within the data.

1. Larynx Cancer

First, we performed 5-to-1 digit greedy matching with N changed from 1 to 5 on the larynx cancer dataset stage 1&2, and we performed 5-to-2 digit greedy matching with N changed from 1 to 5 on the larynx cancer dataset stage 1&2. The results are shown in Table 2.1. From the result we can see that, all log-rank p-values were > 0.05. We used the log-rank test to test the equality of the survival curves within the matched sample. For stage 1&2, 1 to 1 greedy matching, 1 to 2 greedy matching, 1 to 3 greedy matching, 1 to 4 greedy matching and 1 to 5 greedy matching for 5-to-1 digit and 5-to-2 digit matching all did not have statistically significantly different survival curves within each matched sample. Since 5-to-2 digit matching was stricter, the number of matched HPV positive patients for each N was smaller than 5-to-1

digit matching. All hazard ratios were around 1, indicating that HPV status had no significant effect on the survival time. The widths of the hazard ratios for 5-to-1 digit matching were generally narrower than 5-to-2 digit matching, indicating 5-to-1 digit matching increased the precision of the hazard ratio estimate.

Second, we performed 5-to-1 digit greedy matching with N changed from 1 to 5 on larynx cancer dataset stage 3&4, and we performed 5-to-2 digit greedy matching with N changed from 1 to 5 on larynx cancer dataset stage 3&4. The results are shown in Table 2.2. From the result, we can see that all log-rank p-values were < 0.05, indicating for stage 3&4, 1 to 1 greedy matching, 1 to 2 greedy matching, 1 to 3 greedy matching, 1 to 4 greedy matching and 1 to 5 greedy matching for 5-to-1 digit and 5-to-2 digit matching all had statistically significantly different survival curves within each matched sample. Likely, for larynx cancer patients in stages 3 and 4, the difference of survival outcomes between HPV positive and HPV negative was more evident than stage 1 and 2. Since 5-to-2 digits matching was stricter, the number of matched HPV positive patients for each N was smaller than 5-to-1 digits matching. All hazard ratios were below 1, indicating that HPV status had a reduced effect on the hazard. The widths of hazard ratio confidence interval for 5-to-1 digit matching were generally narrower than 5-to-2 digit matching, but the difference was very small.

2. Hypopharynx Cancer

Lastly, we performed 5-to-1 digit greedy matching with N changed from 1 to 5 on hypopharynx cancer, and we performed 5-to-2 digit greedy matching with N changed from 1 to 5 on hypopharynx cancer. The results are shown in Table 2.3. From the result we can see that all log-rank p-values were < 0.05, indicating that 1 to 1 greedy matching, 1 to 2 greedy matching, 1 to 3 greedy matching, 1 to 4 greedy matching and 1 to 5 greedy matching for 5-to-1 digit and 5-to-2 digit matching all had statistically significantly different survival curves within each matched sample. Since 5-to-2 digit matching was stricter, the number of matched HPV positive patients

for each N was smaller than 5-to-1 digit matching, except when N was 5, which was slightly greater than 5-to-1 digit matching. All hazard ratios were between 0.5 and 0.6, and the widths of hazard ratio confidence interval for 5-to-1 digit matching were generally narrower than 5-to-2 digit matching, indicating 5-to-1 digit matching increased the precision of the hazard ratio estimate. However, the width of hazard ratio confidence interval did not decrease as N increased, meaning the matching did not yield more precise results as N increased.

Balance Check

Another important factor to consider was the standardized differences for balance checking, in order to see if the distributions of all measured baseline covariates were the same between the HPV positive patients and HPV negative patients. The distribution of all measured baseline covariates is expected to be sufficiently similar if standardized difference is < 0.1.

1. Larynx Cancer – 5-to-1 digit matching

For larynx cancer, stage 1&2, 5-to-1 digit matching, the results for balance check are shown in Table 3.1.

For 1 to 1 matching of larynx cancer in stage 1&2, negligible differences in the mean or prevalence of all covariates between the two groups existed except for the covariate Charlson-Deyo score and chemotherapy, suggesting these two measured baseline covariates were not the same between the HPV positive patients and HPV negative patients after matching. For 1 to 2 matching of larynx cancer in stage 1&2, negligible differences in the mean or prevalence of all covariates between the two groups existed except for covariate sex, suggesting this measured baseline covariate was not the same between the HPV positive patients and HPV negative patients. For 1 to 3 matching and 1 to 4 matching of larynx cancer in stage 1&2, negligible differences in the mean or prevalence of all covariates between the two groups existed. For 1 to 5 matching of larynx cancer in stage 1&2, negligible differences in the mean or prevalence of all covariates between the two groups existed except for glottis in the covariate primary site,

suggesting this measured baseline covariate was not the same between the HPV positive patients and HPV negative patients.

From the balance check tables for Larynx cancer stage 1&2 for 5-to-1 digit matching, there were almost no standardized differences that were greater than 0.1 after N was bigger than 2. Generally, with N increased, the number of covariates which were not the same at baseline between HPV positive patients and HPV negative patients decreased. Thus, the greater N is, the more the distribution of baseline covariates is balanced.

For larynx cancer, stage 3&4, 5-to-1 digit matching, the results for balance check is shown in Table 3.2

For 1 to 1 matching of larynx cancer in stage 3&4, negligible differences in the mean or prevalence of all covariates between the two groups existed except for supraglottis in primary site, level 3 and 4A in clinical stage group, and age at diagnosis. For 1 to 2 matching of larynx cancer in stage 3&4, negligible differences in the mean or prevalence of all covariates between the two groups existed except for supraglottis in primary site, Charlson-Deyo score, and age at diagnosis. For 1 to 3 matching, 1 to 4 matching, and 1 to 5 matching of larynx cancer in stage 3&4, negligible differences in the mean or prevalence of all covariates between the two groups existed.

From the balance check tables for Larynx cancer stage 3&4, 5-to-1 digit matching, there were no standardized differences that were greater than 0.1 after N was bigger than 2. With N increased, the number of covariates which were not the same at baseline between HPV positive patients and HPV negative patients decreased. Thus, the greater N is, the more the distribution of baseline covariates is balanced.

2. Larynx Cancer – 5-to-2 digit matching

Next, we performed 5-to-2 digit matching for stage 1&2 patients, and the result is shown in Table 3.3.

For 1 to 1 matching of larynx cancer in stage 1&2, 5-to-2 digit matching, negligible differences in the mean or prevalence of all covariates between the two groups existed except for Charlson-Deyo score, and chemotherapy. For 1 to 2 matching of larynx cancer in stage 1&2, negligible differences in the mean or prevalence of all covariates between the two groups existed except for Charlson-Deyo score. For 1 to 3 matching of larynx cancer in stage 1&2, negligible differences in the mean or prevalence of all covariates between the two groups existed except for glottis in primary site and Charlson-Deyo score. For 1 to 4 matching of larynx cancer in stage 1&2, negligible differences in the mean or prevalence of all covariates between the two groups existed except for glottis in primary site. For 1 to 5 matching of larynx cancer in stage 1&2, negligible differences in the mean or prevalence of all covariates between the two groups existed except for radiation.

From the results we can see that, there were covariates which had standardized differences greater than 0.1 when N equaled to 1, 2, 3, 4, 5, which was different from 5-to-1 digit matching, where almost all covariates had standardized differences less than 0.1 when N was greater than or equal to 3. So, we can conclude that 5-to-1 digit matching matched better than 5-to-2 digit matching on larynx cancer stage 1&2.

We also did 5-to-2 digit matching on larynx cancer stage 3&4 to see if the results agreed with stage 1&2, and the result is shown in Table 3.4.

For 1 to 1 matching of larynx cancer in stage 3&4, negligible differences in the mean or prevalence of all covariates between the two groups existed except for supraglottis in primary site, level 3 and 4A in clinical stage group, and age at diagnosis. For 1 to 2 matching of larynx cancer in stage 3&4, negligible differences in the mean or prevalence of all covariates between the two groups existed except for supraglottis in primary site, and Charlson-Deyo score. For 1 to 3 matching, 1 to 4 matching, and 1 to 5 matching of larynx cancer in stage 3&4, negligible differences in the mean or prevalence of all covariates between the two groups existed.

From the results, for larynx cancer stage 3&4, 5-to-2 digit matching, there were no standardized differences that were greater than 0.1 after N was bigger than 2. When N increases, the number of covariates which were not the same at baseline between HPV positive patients and HPV negative patients decreased, which was different from stage 1&2. Thus, we could conclude that, 5-to-2 digit matching did not perform as well as 5-to-1 digit matching, but the difference was not substantial.

3. Hypopharynx Cancer – 5-to-1 digit matching

We did a balance check for both 5-to-1 digit and 5-to-2 digit matching on the larynx cancer data set (4,835), which was the larger data set. We also performed a balance check for 5-to-1 digit and 5-to-2 digit matching on the hypopharynx cancer data set (1,085), which was relatively smaller.

For hypopharynx cancer, 5-to-1 digit matching, the result is shown in Table 3.5.

For 1 to 1 matching of hypopharynx cancer, negligible differences in the mean or prevalence of all covariates between the two groups existed except for Charlson-Deyo score, chemotherapy, radiation and stage 1 in clinical stage group. For 1 to 2 matching of hypopharynx cancer, negligible differences in the mean or prevalence of all covariates between the two groups existed except for stage 4 in clinical stage group. For 1 to 3 matching, 1 to 4 matching, and 1 to 5 matching of hypopharynx cancer, negligible differences in the mean or prevalence of all covariates between the two groups existed.

From the balance check tables for hypopharynx cancer, 5-to-1 digit matching, there were no standardized differences that were greater than 0.1 after N was bigger than 2. With N increased, the number of covariates which were not the same at baseline between HPV positive patients and HPV negative patients decreased. Thus, the greater N is, the more the distribution of baseline covariates is balanced.

4. Hypopharynx Cancer – 5-to-2 digit matching

Next, we did 5-to-2 digit matching on hypopharynx cancer to see if the results agree with 5-to-1 digit matching, and the result is shown in Table 3.6.

For 1 to 1 matching of hypopharynx cancer, negligible differences in the mean or prevalence of all covariates between the two groups existed except for covariates Charlson-Deyo score, chemotherapy, radiation and stage 1 in clinical stage group. For 1 to 2 matching of hypopharynx cancer, negligible differences in the mean or prevalence of all covariates between the two groups existed except for level 4 in clinical stage group. For 1 to 3 matching, 1 to 4 matching, and 1 to 5 matching of hypopharynx cancer, negligible differences in the mean or prevalence of all covariates between the two groups existed.

From the balance check tables for hypopharynx cancer, 5-to-2 digit matching, there were no standardized differences that were greater than 0.1 after N was bigger than 2. With N increased, the number of covariates which were not the same at the baseline between HPV positive patients and HPV negative patients decreased. Thus, the greater N is, the more the distribution of baseline covariates is balanced.

For hypopharynx cancer, the performance of 5-to-1 digit and 5-to-2 digit matching were similar, and the standardized differences for all covariates were less than 0.1 when N was greater than 2. On the other hand, for larynx cancer, not all the standardized differences for all covariates were less than 0.1 when N was greater than 2, where we performed 5-to-2 digit matching on stage 1&2. Even though there were few exceptional circumstances, we can still conclude that, when N equals 1 or 2, the distribution of measured baseline covariates were not all the same between the HPV positive patients and HPV negative patients in the matched sample, but when N is greater than or equal than 3, the distribution of measured baseline covariates were similar between the HPV positive patients and HPV negative patients in the matched samples.

DISCUSSION

The purpose of this paper is to determine the differences between one to one greedy matching and one to N greedy matching, along with different matching digits, and conclude which propensity score matching performs better for two data sets from NCDB: larynx cancer data set and hypopharynx cancer data set. To do so, first we calculated the propensity score by using the logistic regression model. For the matching algorithm, we used greedy matching, in which one treated subject is matched to an untreated subject which has the closest propensity score. We performed 1 to 1, 1 to 2, 1 to 3, 1 to 4, 1 to 5 greedy matching, with 5-to-1 digit and 5-to-2 digit, on two data sets. From the results, we can see that the number of matched HPV positive patients for each N for 5-to-2 digit matching is smaller than that of 5-to-1 digit matching, so we may conclude that 5-to-2 digit matching is stricter than 5-to-1 digit matching. Widths of the hazard ratio confidence interval for 5-to-1 digit matching were generally narrower than 5-to-2 digit matching, indicating that 5-to-1 digit matching produced more precise hazard ratio estimates. For larynx cancer stage 1&2, survival curves of the matched samples were similar; however, for stage 3&4 and hypopharynx cancer, the survival curves consistently were different between the matched samples. This may be because of the distinguishing differences of survival outcomes between HPV positive and HPV negative patients in larynx cancer stage 3&4 and hypopharynx cancer. There are almost no standardized differences that are greater than 0.1 after N is bigger than 2, except for 5-to-2 digit matching on larynx cancer stage 1&2. Generally speaking, when N increases, the number of covariates which are not the same at baseline between HPV positive patients and HPV negative patients decreases, and 5-to-1 digit matching performs better than 5-to-2 digit matching. Although we found minimal differences in the hazard ratio effect size, the width of the 95% confidence interval, and the percent of case matches across these datasets as N was increased from 1 to 5, we found that the variable balancing was far better. Consistently, we found better variable balancing and, of course, higher HPV positive match rates

when using 5-to-1 digit matching instead of 5-to-2 digit matching. If a dataset has the capacity to allow 1 to 3, 1 to 4, or 1 to 5 matching, it is recommended to increase N from 1 to achieve better balance across baseline characteristics included in the propensity model.

This paper compares the performance of greedy matching in different scales. The paper does the matching on different case and control ratios from 1 to 1, to 1 to 5, which allows us to see the impact of different N. Also, the paper does the matching on different matching digits, which allows us to see the difference between two kinds of matched samples: the one with better but fewer matched pairs, and the other one with more matched pairs. This paper analyzes the impact of the size of the data set as well.

There are limitations of this study as well. We only match propensity scores as high as 5 digits. When it comes to higher digits, there will be an empty matched sample sub-set corresponding to a specific digit. This may be caused by the size of the data, and there is not enough data to do higher digit matching. In addition, there are 124 variables in the original data sets, but we only choose 8 of them. It is possible that we missed some important variables that could generate different propensity scores.

For further exploration, there are other approaches that could be tried. However, the results may not apply to other datasets; thus, a simulation study should be considered to draw definitive conclusions. In this paper, we used logistic regression to calculate the propensity score, and there are other methods such as random forest and boosting. Also, we used greedy matching to perform the propensity score matching, and researchers can also try optimal matching to see if there are interesting results. Additionally, researchers can continue increasing N to see if there are better matching ratios. Lastly, researchers can change the propensity score matching digits to a higher number, which could lead to better matched samples.

APPENDIX

Table 1.1 – Descriptive Statistics – Larynx Cancer

Variable	Level	N = 4835	%
HPV status	Negative	4304	89.0
	Positive	531	11.0
Primary Site	C320 - Glottis	2103	43.5
	C321 - Supraglottis	1970	40.7
	C322 - Subglottis	68	1.4
	C328 - Larynx: Overlapping	170	3.5
	C329 - Larynx NOS	524	10.8
Facility Type	Community cancer program/Integrated network cancer program	928	19.6
	Comprehensive community cancer program	1707	36.0
	Academic/Research program	2103	44.4
	Missing	97	-

Variable	Level	N = 4835	%
Facility Location	Northeast	1157	24.4
	South	1851	39.1
	Midwest	1104	23.3
	West	626	13.2
	Missing	97	-
Sex	Male	3741	77.4
	Female	1094	22.6
Race	White	3947	81.6
	Black	709	14.7
	Others/Unknown	179	3.7
Insurance status	Not Insured/Unknown	417	8.6
	Private	1654	34.2
	Medicaid/Medicare/Other Government	2764	57.2
Median Income Quartiles 2000	Not Available	133	-
	< $30,000	819	17.4
	$30,000 - $35,999	1006	21.4
	$36,000 - $45,999	1330	28.3
	$46,000 +	1547	32.9

Variable	Level	N = 4835	%
Charlson-Deyo Score	0	3440	71.1
	1+	1395	28.9
Year of Diagnosis	2010	544	11.3
	2011	1109	22.9
	2012	1503	31.1
	2013	1679	34.7
AJCC Clinical Stage Group	1	1406	29.1
	2	773	16.0
	3	1089	22.5
	4	69	1.4
	4A	1378	28.5
	4B	117	2.4
	4C	3	0.1
Surgical margins	Negative	1087	79.1
	Positive	287	20.9
	Missing	3461	-
Extracapsular extension	No	638	85.6
(path)	Yes	107	14.4
	Missing	4090	-

Variable	Level	N = 4835	%
Chemotherapy	No	2636	55.9
	Yes	2078	44.1
	Missing	121	-
Radiation	No	1022	21.3
	Yes	3785	78.7
	Missing	28	-
Surgery at Primary Site	No	3034	62.9
	Yes	1792	37.1
	Missing	9	-
Grade	Well Differentiated	568	14.7
	Moderately Differentiated	2392	61.7
	Poorly Differentiated/Undifferentiated	915	23.6
	Missing	960	-

Variable	Level	N = 4835	%
Age at Diagnosis	Mean	62.44	-
	Median	62	-
	Minimum	19	-
	Maximum	90	-
	Std Dev	11.40	-
	Missing	0	-
Lymph node size (cm)	Mean	3.24	-
	Median	2	-
	Minimum	0.10	-
	Maximum	98	-
	Std Dev	6.91	-
	Missing	3633	-

Table 1.2 – Descriptive Statistics – Hypopharynx Cancer

Variable	Level	N = 1085	%
HPV status	Negative	893	82.3
	Positive	192	17.7

Variable	Level	N = 1085	%
Primary Site	C129 - Pyriform sinus	526	48.5
	C130 - Postcricoid region	22	2.0
	C131 - Hypopharyngeal aspect of aryepiglottic fold	81	7.5
	C132 - Posterior wall of hypopharynx	70	6.5
	C138 - Overlapping lesion of hypopharynx	55	5.1
	C139 - Hypopharynx, NOS (laryngopharynx)	331	30.5
Facility Type	Community cancer program/Integrated network cancer program	194	18.0
	Comprehensive community cancer program	388	36.1
	Academic/Research program	493	45.9
	Missing	10	-

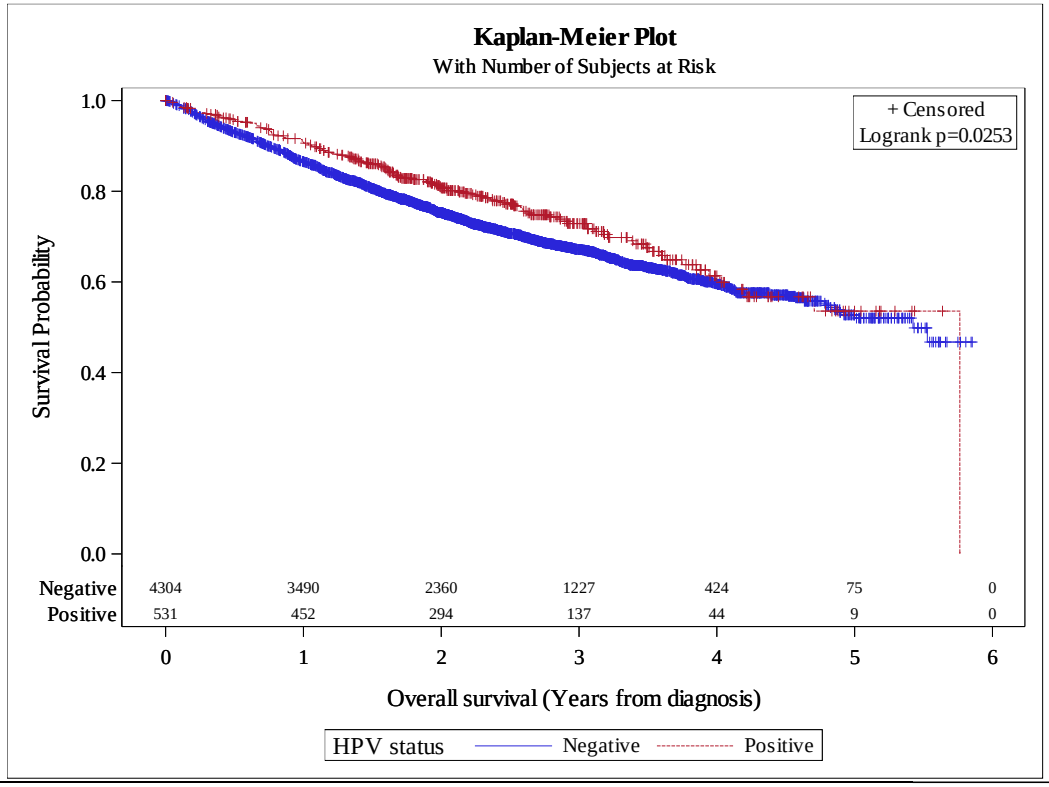
Variable	Level	N = 1085	%
Facility Location	Northeast	267	24.8
	South	424	39.4
	Midwest	231	21.5
	West	153	14.2
	Missing	10	-
Sex	Male	884	81.5
	Female	201	18.5
Race	White	890	82.0
	Black	166	15.3
	Others/Unknown	29	2.7
Insurance status	Not Insured/Unknown	86	7.9
	Private	376	34.7
	Medicaid/Medicare/Other Government	623	57.4
Median Income Quartiles 2000	Not Available	25	-
	< $30,000	184	17.4
	$30,000 - $35,999	197	18.6
	$36,000 - $45,999	296	27.9
	$46,000 +	383	36.1

Variable	Level	N = 1085	%
Charlson-Deyo Score	0	795	73.3
	1+	290	26.7
Year of Diagnosis	2010	109	10.0
	2011	268	24.7
	2012	315	29.0
	2013	393	36.2
AJCC Clinical Stage Group	1	60	5.5
	2	105	9.7
	3	223	20.6
	4	28	2.6
	4A	555	51.2
	4B	113	10.4
	4C	1	0.1
Surgical margins	Negative	164	72.9
	Positive	61	27.1
	Missing	860	-
Extracapsular extension	No	140	72.9
(path)	Yes	52	27.1
	Missing	893	-

Variable	Level	N = 1085	%
Chemotherapy	No	280	26.3
	Yes	785	73.7
	Missing	20	-
Radiation	No	166	15.4
	Yes	914	84.6
	Missing	5	-
Surgery at Primary Site	No	817	75.4
	Yes	267	24.6
	Missing	1	-
Grade	Well Differentiated	38	4.6
	Moderately Differentiated	442	53.8
	Poorly Differentiated/Undifferentiated	342	41.6
	Missing	263	-

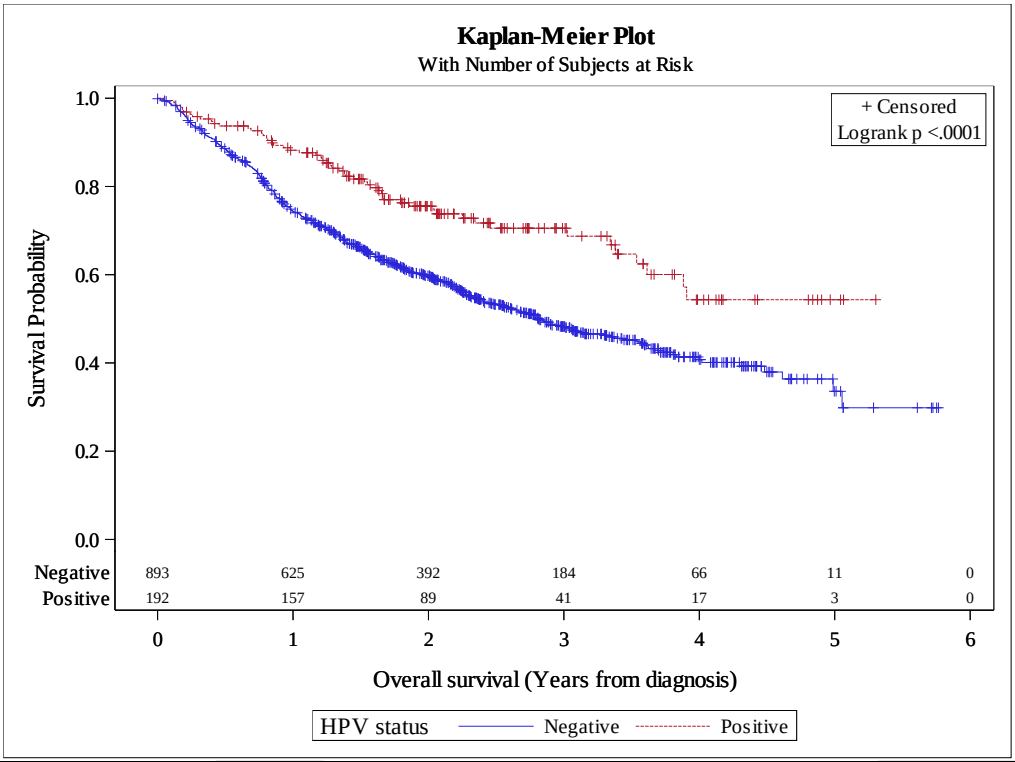
Variable	Level	N = 1085	%
Age at Diagnosis	Mean	61.72	-
	Median	61	-
	Minimum	25	-
	Maximum	90	-
	Std Dev	10.48	-
	Missing	0	-
Lymph node size (cm)	Mean	3.51	-
	Median	2.50	-
	Minimum	0.10	-
	Maximum	98	-
	Std Dev	6.32	-
	Missing	492	-

Figure 1.1 - KM survival curves for Larynx cancer - HPV



HPV status	No. of Subject	Event	Censored	Median Survival (95% CI)	1 Yr Survival	2 Yr Survival	5 Yr Survival
Negative	4304	1304 (30%)	3000 (70%)	5.4 (4.9, NA)	86.5% (85.5%, 87.5%)	75.3% (73.9%, 76.6%)	52.0% (48.3%, 55.6%)
Positive	531	132 (25%)	399 (75%)	5.8 (4.2, 5.8)	90.8% (88.0%, 93.0%)	80.7% (76.8%, 84.0%)	53.6% (43.3%, 62.8%)

Figure 1.2 - KM survival curves for Hypopharynx cancer - HPV



HPV status	No. of Subject	Event	Censored	Median Survival (95% CI)	1 Yr Survival	2 Yr Survival	5 Yr Survival
Negative	893	422 (47%)	471 (53%)	2.8 (2.4, 3.3)	74.4% (71.3%, 77.2%)	59.7% (56.3%, 63.0%)	33.6% (26.1%, 41.2%)
Positive	192	54 (28%)	138 (72%)	NA (3.6, NA)	88.2% (82.6%, 92.1%)	75.5% (68.3%, 81.4%)	54.3% (41.4%, 65.6%)

Figure 2.1 – Distribution of Propensity Score for Larynx Cancer in Stage1&2

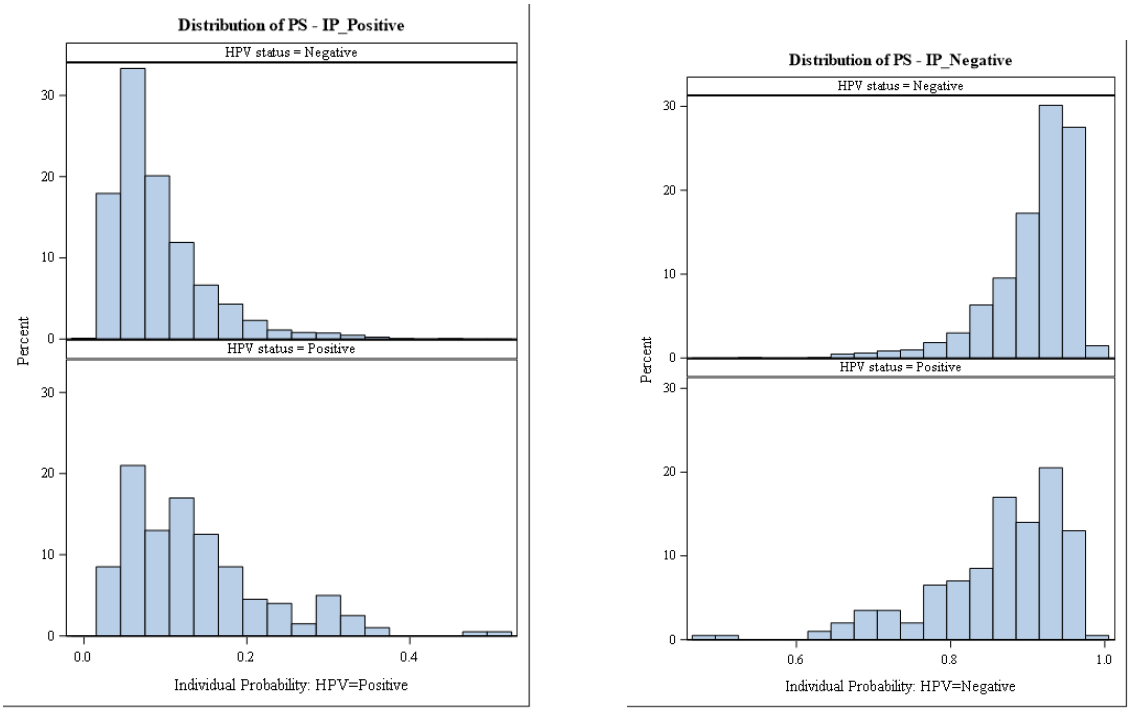


Figure 2.2 – Distribution of Propensity Score for Larynx Cancer in Stage3&4

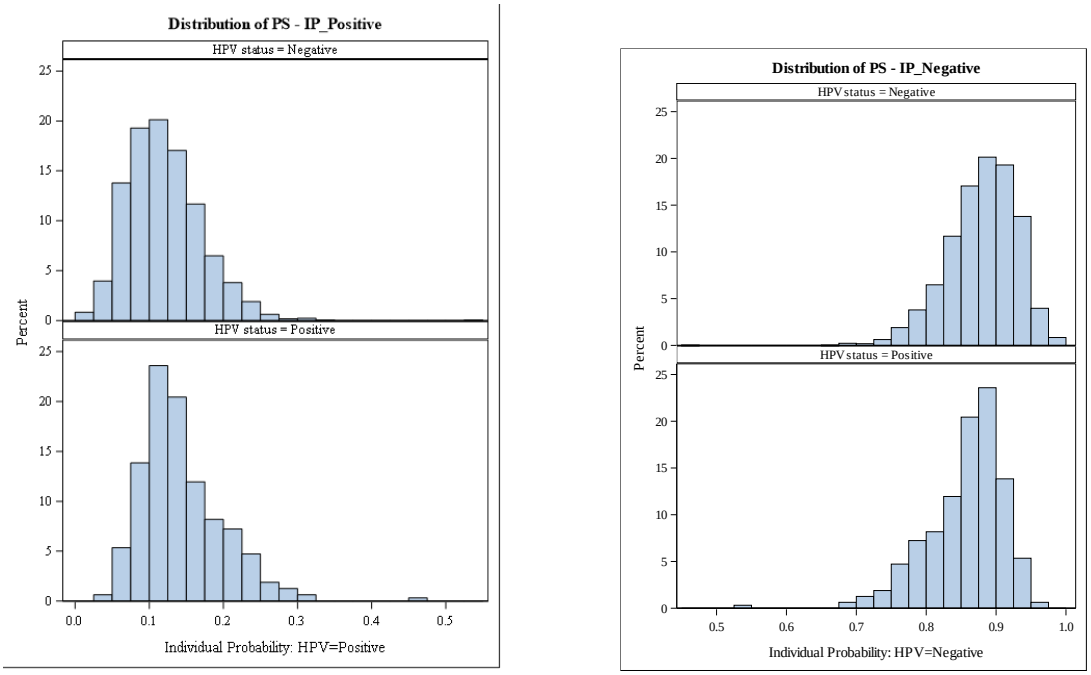


Figure 2.3 – Distribution of Propensity Score for Hypopharynx Cancer in All Stages

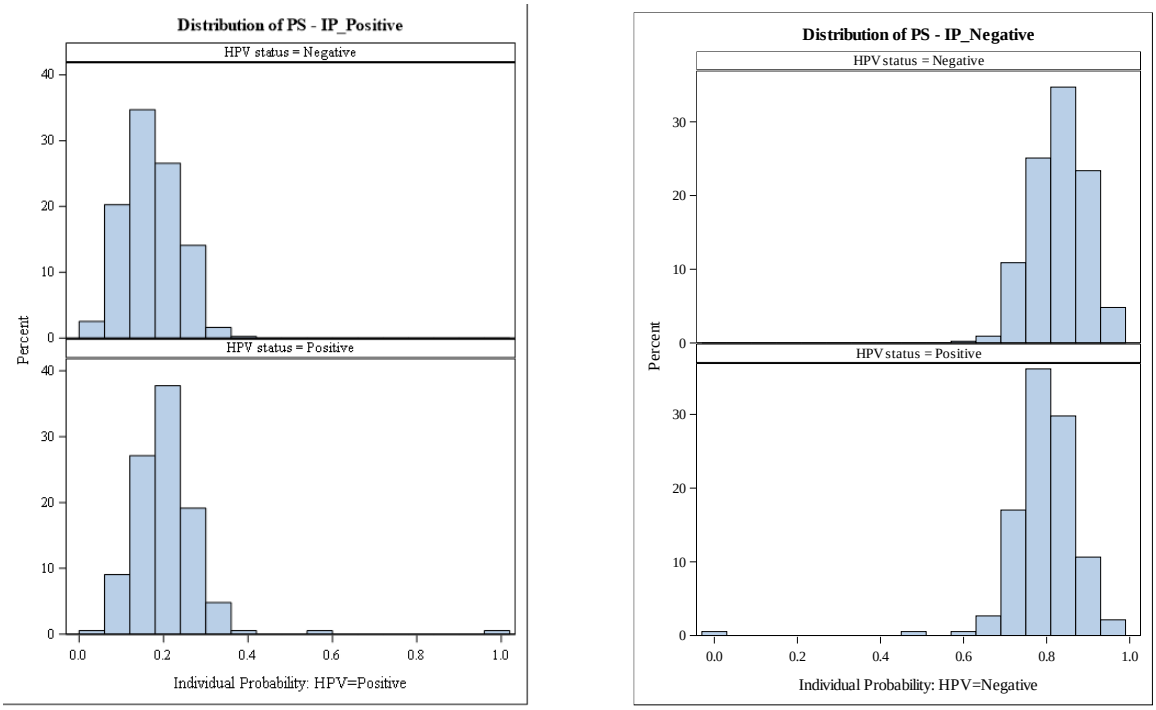


Figure 3 – Matching Algorithm of One to N Greedy Matching

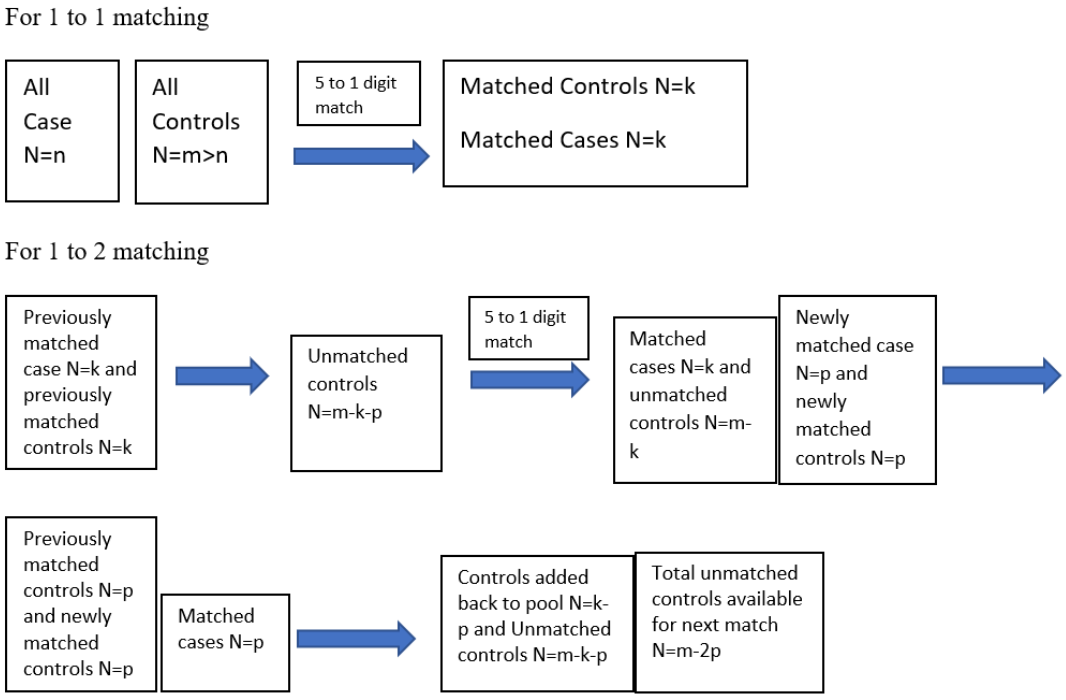


Table 2.1 – Overall Survival – Propensity Score – Larynx Cancer – Stage 1&2

5 to 1 digits	N	HPV Positive	HPV Negative	Hazard Ratio (95% Confidence Interval)	CI Width	HR P-value	Log-rank P-value
	1	199	199	1.08 (0.68-1.72)	1.04	0.747	0.773
	2	196	392	0.84 (0.55-1.28)	0.73	0.413	0.417
	3	187	561	0.89 (0.58-1.36)	0.78	0.583	0.569
	4	187	748	0.89 (0.59-1.35)	0.76	0.586	0.570
	5	159	795	1.08 (0.73-1.60)	0.87	0.693	0.697
5 to 2 digits	**N**	**HPV Positive**	**HPV Negative**	**Hazard Ratio (95% Confidence Interval)**	**CI Width**	**HR P-value**	**Log-rank P-value**
	1	193	193	1.08 (0.68-1.72)	1.04	0.752	0.776
	2	179	358	0.89 (0.58-1.38)	0.8	0.615	0.613
	3	169	507	1.03 (0.67-1.58)	0.91	0.888	0.886
	4	166	664	1.05 (0.70-1.58)	0.88	0.809	0.809
	5	153	765	1.02 (0.67-1.56)	0.89	0.918	0.917

Table 2.2 – Overall Survival – Propensity Score – Larynx Cancer – Stage 3&4

5 to 1 digits	N	HPV Positive	HPV Negative	Hazard Ratio (95% Confidence Interval)	CI Width	HR P-value	Log-rank P-value
	1	317	317	0.65 (0.50-0.84)	0.34	<.001	0.001
	2	317	634	0.69 (0.55-0.86)	0.31	0.001	0.002

	3	313	939	0.75 (0.60-0.93)	0.33	0.008	0.010
	4	313	1252	0.72 (0.62-0.94)	0.32	0.011	0.013
	5	313	1565	0.76 (0.62-0.93)	0.31	0.007	0.010
5 to 2 digits	**N**	**HPV Positive**	**HPV Negative**	**Hazard Ratio (95% Confidence Interval)**	**CI Width**	**HR P-value**	**Log-rank P-value**
	1	317	317	0.65 (0.50-0.84)	0.34	<.001	0.001
	2	307	614	0.72 (0.57-0.90)	0.33	0.005	0.006
	3	295	885	0.77 (0.62-0.95)	0.33	0.016	0.023
	4	292	1168	0.78 (0.63-0.97)	0.34	0.023	0.028
	5	278	1390	0.75 (0.60-0.93)	0.33	0.010	0.011

Table 2.3 – Overall Survival – Propensity Score – Hypopharynx Cancer

5 to 1 digits	**N**	**HPV Positive**	**HPV Negative**	**Hazard Ratio (95% Confidence Interval)**	**CI Width**	**HR P-value**	**Log-rank P-value**
	1	187	187	0.52 (0.37-0.73)	0.36	<.001	<.001
	2	187	374	0.57 (0.42-0.78)	0.36	<.001	<.001
	3	187	561	0.55 (0.41-0.75)	0.34	<.001	<.001
	4	163	652	0.61 (0.46-0.82)	0.36	0.001	0.001
	5	106	530	0.67 (0.49-0.93)	0.44	0.017	0.026
5 to 2 digits	**N**	**HPV Positive**	**HPV Negative**	**Hazard Ratio (95% Confidence Interval)**	**CI Width**	**HR P-value**	**Log-rank P-value**

	1	187	187	0.52 (0.37-0.73)	0.36	<.001	<.001
	2	182	364	0.56 (0.40-0.77)	0.37	<.001	<.001
	3	143	429	0.53 (0.38-0.75)	0.37	<.001	<.001
	4	133	532	0.59 (0.43-0.81)	0.38	0.001	0.002
	5	111	555	0.59 (0.41-0.85)	0.44	0.004	0.003

Table 3.1 – Balance Check – 5 to 1 Digits – Larynx Stage I-II

Covariates	**Level**	**1 to 1**	**1 to 2**	**1 to 3**	**1 to 4**	**1 to 5**
Primary site	Glottis	0.041	0.026	0.047	0.068	**0.103**
	Supraglottis	0.042	0.005	0.026	0.054	0.087
	Other	0.000	0.044	0.045	0.034	0.039
Sex	Male	0.089	**0.109**	0.078	0.074	0.066
	Female	0.089	**0.109**	0.078	0.074	0.066
Charlson-Deyo Score	0	**0.122**	0.099	0.081	0.044	0.018
	1+	**0.122**	0.099	0.081	0.044	0.018
Chemotherapy	No	**0.111**	0.044	0.044	0.057	0.046
	Yes	**0.111**	0.044	0.044	0.057	0.046
Radiation	No	0.011	0.011	0.051	0.050	0.093
	Yes	0.011	0.011	0.051	0.050	0.093
Surgery at Primary Site	No	0.050	0.021	0.022	0.032	0.038
	Yes	0.050	0.021	0.022	0.032	0.038
AJCC Clinical Stage Group	1	0.010	0.046	0.036	0.019	0.003

	2	0.010	0.046	0.036	0.019	0.003
Age at Diagnosis		0.024	0.053	0.034	0.020	0.066

Table 3.2 – Balance Check – 5 to 1 Digits – Larynx Stage III-IV

Covariates	**Level**	**1 to 1**	**1 to 2**	**1 to 3**	**1 to 4**	**1 to 5**
Primary site	Glottis	0.042	0.073	0.082	0.093	0.075
	Supraglottis	**0.101**	**0.108**	0.081	0.063	0.042
	Other	0.086	0.064	0.022	0.010	0.020
Sex	Male	0.071	0.071	0.053	0.056	0.055
	Female	0.071	0.071	0.053	0.056	0.055
Charlson-Deyo Score	0	0.094	**0.101**	0.060	0.012	0.013
	1+	0.094	**0.101**	0.060	0.012	0.013
Chemotherapy	No	0.036	0.004	0.010	0.039	0.035
	Yes	0.036	0.004	0.010	0.039	0.035
Radiation	No	0.026	0.039	0.035	0.018	0.018
	Yes	0.026	0.039	0.035	0.018	0.018
Surgery at Primary Site	No	0.000	0.028	0.032	0.031	0.004
	Yes	0.000	0.028	0.032	0.031	0.004
AJCC Clinical Stage Group	3	**0.115**	0.077	0.009	0.005	0.028
	4	0.021	0.021	0.028	0.026	0.004
	4A	**0.136**	0.097	0.030	0.019	0.017
	4B	0.045	0.037	0.030	0.038	0.027
Age at Diagnosis		**0.106**	**0.101**	0.079	0.070	0.057

Table 3.3 – Balance Check – 5 to 2 Digits – Larynx Stage I-II

Covariates	**Level**	**1 to 1**	**1 to 2**	**1 to 3**	**1 to 4**	**1 to 5**
Primary site	Glottis	0.053	0.080	**0.110**	**0.119**	0.083
	Supraglottis	0.054	0.060	0.090	0.099	0.056
	Other	0.000	0.047	0.048	0.048	0.056
Sex	Male	0.094	0.099	0.071	0.054	0.042
	Female	0.094	0.099	0.071	0.054	0.042
Charlson-Deyo Score	0	**0.114**	**0.136**	**0.108**	0.064	0.039
	1+	**0.114**	**0.136**	**0.108**	0.064	0.039
Chemotherapy	No	**0.113**	0.076	0.083	0.062	0.038
	Yes	**0.113**	0.076	0.083	0.062	0.038
Radiation	No	0.011	0.031	0.084	0.081	**0.124**
	Yes	0.011	0.031	0.084	0.081	**0.124**
Surgery at Primary Site	No	0.052	0.011	0.032	0.006	0.040
	Yes	0.052	0.011	0.032	0.006	0.040
AJCC Clinical Stage Group	1	0.021	0.073	0.028	0.012	0.000
	2	0.021	0.073	0.028	0.012	0.000
Age at Diagnosis		0.009	0.016	0.040	0.058	0.058

Table 3.4 – Balance Check – 5 to 2 Digits – Larynx Stage III-IV

Covariates	**Level**	**1 to 1**	**1 to 2**	**1 to 3**	**1 to 4**	**1 to 5**

Primary site	Glottis	0.042	0.057	0.077	0.093	0.053
	Supraglottis	**0.101**	**0.100**	0.088	0.081	0.046
	Other	0.086	0.070	0.035	0.011	0.007
Sex	Male	0.071	0.078	0.084	0.072	0.075
	Female	0.071	0.078	0.084	0.072	0.075
Charlson-Deyo Score	0	0.094	**0.100**	0.066	0.015	0.011
	1+	0.094	**0.100**	0.066	0.015	0.011
Chemotherapy	No	0.036	0.011	0.000	0.029	0.028
	Yes	0.036	0.011	0.000	0.029	0.028
Radiation	No	0.026	0.045	0.059	0.047	0.037
	Yes	0.026	0.045	0.059	0.047	0.037
Surgery at Primary Site	No	0.000	0.021	0.012	0.024	0.011
	Yes	0.000	0.021	0.012	0.024	0.011
AJCC Clinical Stage Group	3	**0.115**	0.069	0.038	0.011	0.024
	4	0.021	0.021	0.045	0.017	0.005
	4A	**0.136**	0.083	0.046	0.014	0.022
	4B	0.045	0.024	0.011	0.004	0.007
Age at Diagnosis		**0.106**	0.089	0.062	0.035	0.010

Table 3.5 – Balance Check – 5 to 1 Digits – Hypopharynx

Covariates	**Level**	**1 to 1**	**1 to 2**	**1 to 3**	**1 to 4**	**1 to 5**
Sex	Male	0.094	0.062	0.030	0.008	0.050
	Female	0.094	0.062	0.030	0.008	0.050

Charlson-Deyo Score	0	**0.160**	0.063	0.047	0.004	0.029
	1+	**0.160**	0.063	0.047	0.004	0.029
Chemotherapy	No	**0.149**	0.093	0.035	0.011	0.033
	Yes	**0.149**	0.093	0.035	0.011	0.033
Radiation	No	**0.104**	0.045	0.054	0.039	0.016
	Yes	**0.104**	0.045	0.054	0.039	0.016
Surgery at Primary Site	No	0.053	0.013	0.013	0.018	0.013
	Yes	0.053	0.013	0.013	0.018	0.013
AJCC Clinical Stage Group	1	**0.148**	0.032	0.020	0.024	0.032
	2	0.087	0.064	0.027	0.005	0.052
	3	0.013	0.007	0.017	0.030	0.014
	4	0.039	**0.105**	0.074	0.079	0.063
	4A	0.077	0.038	0.058	0.062	0.011
	4B	0.021	0.044	0.000	0.011	0.041
Age at Diagnosis		0.085	0.013	0.028	0.023	0.059

Table 3.6 – Balance Check – 5 to 2 Digits – Hypopharynx

Covariates	**Level**	**1 to 1**	**1 to 2**	**1 to 3**	**1 to 4**	**1 to 5**
Sex	Male	0.094	0.063	0.037	0.019	0.027
	Female	0.094	0.063	0.037	0.019	0.027
Charlson-Deyo Score	0	**0.160**	0.064	0.052	0.017	0.008
	1+	**0.160**	0.064	0.052	0.017	0.008
Chemotherapy	No	**0.149**	0.062	0.011	0.031	0.025

	Yes	**0.149**	0.062	0.011	0.031	0.025
Radiation	No	**0.104**	0.046	0.050	0.040	0.041
	Yes	**0.104**	0.046	0.050	0.040	0.041
Surgery at Primary Site	No	0.053	0.007	0.027	0.026	0.017
	Yes	0.053	0.007	0.027	0.026	0.017
AJCC Clinical Stage Group	1	**0.148**	0.032	0.023	0.026	0.008
	2	0.087	0.077	0.040	0.051	0.056
	3	0.013	0.020	0.052	0.000	0.013
	4	0.039	**0.106**	0.072	0.078	0.052
	4A	0.077	0.062	0.056	0.000	0.029
	4B	0.021	0.056	0.042	0.013	0.040
Age at Diagnosis		0.085	0.002	0.036	0.011	0.002

BIBLIOGRAPHY

[1] Sander Greenland, Judea Pearl, James M. Robins. (1999). Causal Diagrams for Epidemiologic Research. *In Epidemiology, Vol. 1, No. 10, pp. 37-48, January 1999.*

[2] Peter C. Austin. (2011). An Introduction to Propensity Score Methods for Reducing the Effects of Confounding in Observational Studies. *Multivariate Behavioral Research, 46:399–424, 2011Copyright © Taylor & Francis Group, LLC ISSN: 0027-3171 print/1532-7906 online DOI: 10.1080/00273171.2011.568786.*

[3] Paul R. Rosenbaum, Donald B. Rubin. (1983). The central role of the propensity score in Observational studies for causal effects. *Biometrika, Volume 70, Issue 1, April 1983, Pages 41–55.*

[4] Liu, Y., Nickleach, D, Zhang, C., Switchenko, J., Kowalski, J. (2018). Carrying out streamlined routine data analyses with reports for observational studies: introduction to a series of generic SAS® macros. *F1000Research: 7:1955. DOI: 10.12688/f1000research.16866.1. License: CC BY4.0.* https://f1000research.com/articles/7-1955

[5] Brian K. Lee, Justin Lessler, Elizabeth A. Stuart. (2009). Improving propensity score weighting using machine learning. *Statistics in Medicine, Volume29, Issue3, 10 February 2010, Pages 337-346.*

[6] Peter C. Austin, Muhammad M. Mamdani. (2005). A comparison of propensity score methods: a case-study estimating the effectiveness of post-AMI statin use.

[7] SEER Stat Fact Sheets: Oral Cavity and Pharynx Cancer. *SEER. April 2016. Archived from the original on 15 November 2016. Retrieved 29 September 2016.*

[8] Guido W. Imbens, Keisuke Hirano. (2004). *The Propensity Score with Continuous Treatments. Journal of the American Statistical Association, Vol. 99, No. 467 (September), pp. 854-866.*

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

[10] Peter C. Austin. (2011). Optimal caliper widths for propensity-score matching when Estimating differences in means and differences in proportions in observational studies. *Pharm Stat, 150-161.*

[11] Lori S. Parsons. Reducing Bias in a Propensity Score Matched-Pair Sample Using Greedy Matching Techniques.

[12] Xing Sam Gu, Paul R. Rosenbaum. (1993). Comparison of Multivariate Matching Methods: Structures, Distances, and Algorithms. *Journal of Computational and Graphical Statistics Volume 2, 1993 - Issue 4.*

[13] Paul R. Rosenbaum, Donald B. Rubin. (1985). Constructing a Control Group Using Multivariate Matched Sampling Methods That Incorporate the Propensity Score. *The American Statistician Volume 39, 1985 - Issue 1.*

[14] Lori S. Parsons. Performing a 1:N Case-Control Match on Propensity Score. *SUGI 29 Paper 165-29.*

[15] Peter C. Austin. (2013). The use of propensity score methods with survival or time-to-event outcomes: reporting measures of effect similar to those used in randomized experiments. *Statistics in Medicine, DOI: 10.1002/sim.5984.*

[16] Normand et al., (2001). Validating recommendations for coronary angiography following acute myocardial infarction in the elderly: A matched analysis using propensity scores. *Journal of Clinical Epidemiology Volume 54, Issue 4, April 2001, Pages 387-398.*

[17] "Head and Neck Cancers". NCI. March 29, 2017. Retrieved 17 September 2017.

[18] GBD 2015 Disease and Injury Incidence and Prevalence, Collaborators. (2016). "Global, regional, and national incidence, prevalence, and years lived with disability for 310 diseases and injuries, 1990-2015: a systematic analysis for the Global Burden of Disease Study 2015". *Lancet. 388 (10053): 1545–1602. doi:10.1016/S0140-6736(16)31678-6. PMC 5055577.*

PMID 27733282.