Assessment of Propensity Score Performance in Small Samples

By

Emily Peterson

Thesis

Submitted to the

Graduate School of Vanderbilt University

in partial fulfillment of the requirements

for the degree of

MASTER OF SCIENCE

in Biostatistics

August, 2015

Nashville, Tennessee

Approved

Tatsuki Koyama, PhD

Dan Ayers, M.S.

Nitin Jain, MD

ACKNOWLEDGEMENTS

This work would not have been possible without the support of the Vanderbilt University School of Medicine, Department of Biostatistics. I am especially indebted to Dr. Tatsuki Koyama, Dan Ayers, and Dr. Nitin Jain, who have supported me through this research and completion of this project.

I am grateful to all of those with whom I have had the pleasure to work during this and other related projects. Each member of my dissertation committee has provided me with extensive personal and professional guidance and taught me a great deal. I would especially like to thank Dr. Koyama, chairman of my committee, for being a great teacher and mentor.

Lastly, I would like to thank my family, whose love, support and guidance are with me in whatever I pursue. They have been inspriation through out my master's program, and will continue to guide me through this next chapter.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

Chapter 1.

Introduction

In observational studies, balance of baseline characteristics by randomization is not possible due to treatment selection, which is determined by subject characteristics. This is the case with rotator cuff tear therapies that include physical therapy or surgical treatments. Patient characteristics such as age, size of tear, gender, and amount of fatty deposits determine whether a patient will receive surgical treatment for their rotator cuff tear. As a result, baseline characteristics of surgery patients differ systematically from those patients that are assigned physcal therapy alone. In this case, one must account for these systematic differences in baseline characteristics between surgical and non-surgical subjects to remove any bias that is caused by these systematic differences. Historically, researchers have used regression adjustment of baseline covariates to account for differences between subjects that are treated with one intervention over another. However, propensity score models are being increasingly used to reduce the impact of treatment-selection bias and account for the systematic differences between treatment groups. The propensity score is defined as a subject's probability of receiving a specific-treatment which is conditioned on the observed baseline covariates. One question that may arise is if there is any gain in using propensity score rather that performing a regression with all covariates used to estimate the propensity score also include in the regression model. Rosenbaum and Rubin(1993) showed that covariate adjusted univariate regression yielded similar results to the multivariate regression when the variables included in the propensity score model were the same as those variables included in the multivariate regression. Therefore, the benefit of the use of propensity scores is one does not need to be concerned with choosing a subset of baseline covariates to be included in a regression analysis on outcome. Several methods exist that use propensity to balance differences in baseline covariates. Matching is a common method whereby control subjects are matched to treated subjects based on propensity scores. Stratification or subclassification consists of grouping subjects into strata defined by observed baseline covariates, then treated

and control subjects within the same strata are compared. Lastly, covariate adjustment regresses the outcome on an indicator variable for treatment and propensity score, which is the more common method for balancing baseline covariates. By this method, the effect of treatment is compared between treated and untreated groups with the same propensity score.[9] However, a disadvantage of covariate adjustment, as Rubin (1997) notes, is this method may be more sensitive to whether propensity score has been accurately estimated.[1][4][5][6]

Propensity scores are calculated using a logistic regression of treatment status on selected baseline variables. However, there has been little research on variable selection when confronted with the restriction of a small sample size. Variables to include in the propensity score model, are therefore, defined by their relationship with the outcome and by their relationship with the treatment. Variables can be categorized into four groups based on these two relationships: 1. true confounders are defined as variables that are strongly associated with both treatment and outcome, 2. variables that are only strongly associated with treatment, 3. variables that are only stronly associated with outcome, and lastly 4. variables that are strongly associated with neither.[5][6] The ideal method of model specification is driven by subject matter understanding and a detailed knowledge of how treatment is assigned. However, including variables only related to exposure(treatment), but that are not related to outcome, decreases the efficiency of an exposure effect according to Rubin (1997).[15] Brookhart et al. (2006) showed that there is a decrease in accuracy of estimated exposure effect and no decrease in bias when variables related to exposure and unrelated to outcome are included in the propensity score model. Therefore, including variables that are only related to treatment can be detrmimental. Other researchers have advocated for including baseline covaraites that are related to both treatment and outcome.[11] However, with the restriction of smaller sample sizes, including all baseline covariates related to treatment or outcome may not be plausible, therefore, it is necessary to determine a propensity score model that can be used in experiments that have smaller sample size restrictions. In addition, there has been no suggestion of a model that does not assume prior matter knowledge. In this case, if there is potential for variables to be incorrectly selected, this could add significant bias to estimated exposure effects.

The questions that remain to be answered from the previously mentioned studies are: How effective are these variable selection methods in small samples when modeling propensity scores? and what alternative is there if we cannot assume prior knowlegde of variable associations to outcome? To answer these questions, this study used a monte carlo simulation method where we assessed four propensity score models that included data reduction mechanisms to select a small set of baseline covariates. We assessed the performance of these methods using a sample size of $n = 40$ from data collected on rotator cuff repair among patients with diagnosed rotator cuff repairs.

The objectives of this study are as follows:

1. To assess performance of four propensity score models that use data reduction techniques to allow for inclusion of all baseline covariates.
2. To investigate the amount of false treatment effects caused when clinical knowlegde or prior assumptions of baseline variable association to outcome is incorrect.
3. To find a useful method of using propensity score in small samples and when prior knowlegde is not assumed.

Therefore,to assess the performance of different methods of variable selection in propensity score modeling when no definite prior knowledge i lacking, we used four different propensity score models. The first propensity score model includes 3 main effects, that we have pre-determined to be significantly related to outcome, and the first principal component. The second propensity score model includes 3 main effects, that we have pre-determined not to be significantly related to the outcome, and the first principal component. In doing this we can determine the bias that is a result of misspecifying the covariates to be included in the propensity score model when the number of covariates is restricted by sample size. Thirdly, we developed a propensity score model that includes the first four principal components, which accounts for total variance explained by four principal components when combining multiple variables. Lastly, we compared the previous three models to a propensity score model with maximum likelihood penalization. In this case, all covariates may be included in the model where the effects of each covariate are scaled to prevent the model from over-fitting. To compare performances of propensity score models, we calculated the type I error and power from each propensity score model out of $n = 1000$ simulations. From these comparisons, we were able to determine which propensity score model resulted in the least amount of false or incorrect treatment effects out of the 1,000 repetitions.

Chapter 2.

Background on rotator cuff repair

Rotator cuff tears account for 4.5 million physician visits and 75,000 surgeries annually. They are the leading cause of shoulder pain and disability. Most rotator cuff tears are degenerative and occur in older adults. Shoulder pain is responsible for 16 percent of all musculoskeletal complaints and has an incidence of 15 cases per 1,000 patients seen in primary care. Rotator cuff tear is the cause in 70 percent of all patients presenting with shoulder pain. Rotator cuff tears are treated by either non-operative treatment, which consists of rehabilitation therapies such as muscle strengthening, range of motion, and stretching and flexibility, or using a surgical open approach, arthoscopic assisted techinique, or as an arthroscopy only procedure.[17][18] Baseline characteristics of age, gender, number of tears, degree of fatty degeneration, size of tear, BMI, height, symptom duration, smoking status and cause of injury are some of the baseline covariates that determine whether a patient is assigned to non-surgical therapy or surgical treatment. The outcome being measured to assess shoulder and upper extremity function and disability is the Shoulder Pain and Disability Index (SPADI), which is measured for patients undergoing rehabilitation post surgery and for non-operative patients to assess their shoulder pain and mobility.*citewilliams*citemacdermid This project is a subsequent analysis to the current The Rotator Cuff Outcomes Workgroup (ROW) cohort study being performed by Vanderbilt University Medical Center Orthopaedics Department, headed by Dr. Jain Nitin, MD. The ROW cohort study recruited 387 patients who may or may not have had confirmed rotator cuff tears, of the 387 patients enrolled in the study, 164 patients with rotator cuff tears were followed for a period of 3 years, and were assigned to either non-operative rehabilitation or surgical intervention by the orthopaedic medical staff. Follow-up standardized questionnaires were completed at 3, 6, 12, 18, 24, and 36 months. The standardized questionnaires included SPADI as well as other pain and mobility self-administered surveys. For the purposes of this study, SPADI values measured at 1 year was considered the outcome variable that measured patient

progress associated with both interventions. Previous studies on rotator cuff repair outcomes used propensity scores to adjust for differences in baseline covariates between treatment groups, such as the study conducted by Joo Han et al. (2011). In a study of rotator cuff repair outcomes in patients with large-to-massive tear with pseudoparalysis, Joo Han et al. (2011), propensity score matching (1-to-1) was performed between pseudoparalytic $(n = 35)$ and nonpseudoparalytic groups $(n = 160)$. Variables that were used to match patients included age, gender, dominance, onset period, aggravation period, number of tendons involved, retraction, operation method, rows of repair, number of anchors, and fatty degeneration. Outcomes that were assessed to indicate repair were the American Shoulder and Elbow Surgeons survey (ASES) and UCLA shoulder rating score. The results showed that range of motion was improved in both groups after rotator cuff repair. All functional outcome scores had significantly improved by final follow-up (12 months from baseline) compared to their preoperative values. ASES and UCLA scores were significantly lower in the pesudoparalytic group. However, at the final followup, there was not a significant difference in functional measures between groups. [19]

Chapter 3.

Propensity score for causal effects

The propensity score is the conditional probability of assignment to a particular treatment given a vector of observed baseline covariates. Rubin and Rosenbaum (1983) stated that both large and small sample theory show that adjustment for scalar propensity score is sufficient to remove bias due to all observed covariates between subjects assigned to respective treatments. The propensity score is used in nonrandomized experiments such that direct comparison may be made also accounting for systematic differences between treatment groups. Therefore, the propensity score can be considered a balancing score $b(x)$, which is a function of the observed covariates x such that the conditional distribution of x given $b(x)$ is the same for both treated and control groups.[13]

Let the conditional probability of assignment to treatment one, given the covariates, be denoted by $e(x) = pr(z = 1|x)$.

where x_i represents baseline covariate values, z_i is the indicator variable for treatment assignment, $e(x)$ is the expected probability that a subject is assigned to treatment, and N is the total number of subjects.
It is assumed $pr(z_1...z_n|x_1...x_n) = \prod_{i=1}^{N} e(x_i)^{z_i}(1-e(x_i))^{1-z_i}$.

3.1. Large Sample Theory

Theorem 1: Treatment assignment and the observed covariates are conditionally indepdent given the propensity score.

$$x \perp z|e(x)$$

Theorem 2: Let $b(x)$ be a function of z. Then $b(x)$ is a balancing score.

$$x \perp z|b(x)$$

, if and only if b(x) is finer than e(x).

Theorem 3: If treatment assignment is strongly ignorable given x, then it is stronglt ignorable given any balancing score $b(x)$,

let r_1 be the response given subject had treatment 1

let r_0 be the response given subject had treatment 0

$$(r_1, r_0) \perp z|x$$

and

$$0 < pr(z = 1|x) < 1$$

for all x imply

$$(r_1, r_0) \perp z|b(x)$$

and

$$0 < pr(z = 1|b(x)) < 1$$

for all $b(x)$

Theorem 4: Suppose treatment assignment is strongly ignorable and $b(x)$ is a balancing score. Then the expected differences in observed responses to the two treatments at $b(x)$ is equal to the average treatment effect at $b(x)$, that is

$$E[r_i|b(x), z = 1] - E[r_0|b(x), z = 0] = E[r_i - r_0|b(x)]$$

Corollary 4.3: Covariance adjustment on balancing scores. Suppose treatment assignment is strongly ignoreable, so that in particular $E[r_i|z = l, b(x)] = E[r_i|b(x)]$ for balancing score $b(x)$. Further suppose the expectation of r_i given $b(x)$ is linear

$$E[r_i|z = l, b(x)] = \alpha_i + \beta_i b(x)(i = 0, 1)$$

Then the estimator

$$\hat{\alpha}_1 - \hat{\alpha}_0 + (\hat{\beta}_1 - \hat{\beta}_0)b(x)$$

is conditionally unbiased given $b(x_i)(i = 1,...,n)$ for the treatment effect at $b(x)$, namely $E[r_1 - r_0|b(x)]$, if $\hat{\alpha}_i$ and $\hat{\beta}_i$ are conditionally unbiased estimators of $\hat{\alpha}_i$ and $\hat{\beta}_i$, such as least squares estimators.

3.2. Small Sample Theory

Estimate $e(x)$ by $\hat{e}(a) = prop(z = 1|x = a)$.

Theorem 5: Suppose $0 < e(\hat{a}) < 1$. Then $prop(z = 0, x = a|e(\hat{x}) = e(\hat{a})) = prop(z = 0|e(\hat{x})) = e(\hat{x}))prop(x = a|e(\hat{x}) = e(\hat{a}))$

Corollary 5.1. Suppose the N units are a random sample from an infinite population, and suppose x takes on only finitely many values in the population and at each such value $0 < e(x) < 1$. Then with probability 1 as $N \to \infty$, subclassification on $e(\hat{x})$ produces sample balance.

In practice, except when x takes on only a few values, $e(\hat{a})$ will be either zero or one for most values of a.

The propensity score can be modelled using an appropriate logit model or discriminant score.

$$e(x) = pr(z = 1|x) = \frac{pr(z = 1)pr(x|z = 1)}{pr(z = 1)pr(x|z = 1) + pr(z = 0)pr(x|z = 0)}$$

Reference:Rosenbaum PR, Rubin DB. The central role of the propensity score in observational studies for causal effects. Biometrics.1983.80.41-55.

Chapter 4.

Variable Selection

There has not been any agreement in applied research as to which variables to include in a propensity score model. However, some of the suggested variable sets have included the following: all measured baseline covariates (known as the full propensity model), all baseline covariates that are associated with treatment assignment, all covariates that are associated with outcome, known as potential confounders, and all covariates that are associated with both treatment assignement and outcome (known as true confounders). The advantage to only including potential confounders or true confounders is the reduction in inefficiency cause by including variables only related to treatment. Rubin and Thomas(1996) first responded to the questions of how many variables should be included in a propensity score model, and whether to remove variables to achieve a more parsimonious model. The authors suggested that unless there is a concensus that a variable is unrelated to outcome or not a proper covariate, it is advisable to include it in the propensity score model. As stated by the authors,

" Excluding potentially relevant variables should be done only when the resultant matched samples are closely balanced with respect to these variables as weill typically occur when the treated and full control sample means of the excluded variables are exceptionally close or when the excluded variables are highly correlated with variables already in the propensity score model." citeRubinthomas

A study performed by Brookhart et al. (2006) assessed three monte carlo simulation studies: 1. propensity score was modeled with variables related to both outcome and exposure (a true confounder) (X_1), 2. propensity score was modeled with variables related to outcome only (X_2), and 3. propensity score was modeled with a variable related to the exposure but not the outcome (X_3). The covariates X_1, X_2, X_3 are independent standard normal random variables with mean 0 and unit variance.

PS Model 1: $Pr[A = 1|X] = \beta_0 + \beta_1 X_1$

PS Model 2: $Pr[A = 1|X] = \beta_0 + \beta_1 X_2$

PS Model 3: $Pr[A = 1|X] = \beta_0 + \beta_1 X_3$

PS Model 4: $Pr[A = 1|X] = \beta_0 + \beta_1 X_1 + \beta_2 X_2$

PS Model 5: $Pr[A = 1|X] = \beta_0 + \beta_1 X_1 + \beta_2 X_3$

PS Model 6: $Pr[A = 1|X] = \beta_0 + \beta_1 X_2 + \beta_3 X_3$

PS Model 7: $Pr[A = 1|X] = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3$

The authors adjusted for propensity score by including propensity score as a covariate in a multivariate outcome model. To evaluate the PS model, the authors measured variance, bias and mean-squared error of corresponding estimators. The simulated data consisted of 1000 repetitions for both $n = 500$ and $n = 2500$ sample sizes. The results of the simulation revealed that the optimal model was the one that included the confounder and variable only related to outcome (PS Model 4). It was found that variance of the estimated treatment effect increased when a variable only related to exposure (X_3) was added to the model. This was the case for both sample sizes. The results confirmed those found by Rubin and Thomas(1996).[1] One of the disadvantages of the above mentioned methods is the necessity to have knowledge of the associations between covariates and outcome. Without knowing covariate relationships to outcome, the mispecification of the propensity model is probable and therefore, yields possible large biases of covariate estimation.

4.1. Principal Component Analysis

In many situations observations are taken on a large number of correlated variables. In these situations it is may be advantageous to reduce the number of variables being studied without sacrificing information about the variables contained in the covariance matrix. Principal Component Analysis, developed by Hotelling(1933), makes an orthogonal transformation of the original variables to give them certain variance properties.

Let X be an $m \times 1$ random vector with mean μ and positive definite covariance matrix Σ.
Let $\lambda_1 \geq \lambda_2 \geq \lambda_3 \geq ...\lambda_m \ (> 0)$ be the latent roots of Σ and let $H = [h_1...h_m]$be an $m \times m$ orthogonal matrix such that

$$H'\Sigma H = \delta = diag(\lambda_1, ..., \lambda_m)$$

so that h_i is an eigenvector of Σ corresponding to the latent root λ_i

Let $U = H'X = (U_1...U_m)'$; then $Cov(U) = \delta$, so that $U_1, ..., U_m$ are all uncorrelated, and $Var(U_i) = \lambda_i$ where $i = 1, ..., m$. The components $U_1, ..., U_m$ are called principal components.

Reference: Muirhead RJ. Aspects of Statistical Theory. Wiley: New Yrok, NY.2005.

The first principal component is the first variables in the newly transformed set and is the normalized linear combination of the original variables with maximum variance; the second principal component is the normalized linear combination having maximum variance out of all linear combinations uncorrelated with the first principal component, and the subsequent principal components continue in the same fashion. Therefore, principal component analysis is concerned with explaining the variability in a vector variable by replacing it by a new variable with a smaller number of components with large variance. "The principal component analysis is fundamentally concerned with the eigenstructure of covariance matrices. The coefficients in the first principal component are the components of the normalized eigenvector corresponding to the largest latent root, and the variance of the first principal component is this largest root." [19] A common criticism of this method of data reduction is that it is not invariant under linear transformations of the variables. Therefore, when variables are not measured in the same units, it is recommended that principal components be extracted from the correlation matrix rather than the covariance matrix.(Hotelling, 1933) Based on this recommendation, principal components used for this analysis will be extracted from the correlation matrix as opposed to the covariance matrix.[19]

4.2. Penalized Maximum Likelihood

Penalized maximum likelihood estimation is a rigorous method to adjust for overfitting by applying shrinkage estimation directly into the model derivation, and is not done post-hoc. Overfitting is a result of smaller sample sizes with a large number of candidate predictors compared to the number of outcomes(events). PMLE is developed for logistic regression models and is a generalization of ridge regression methods. It maximizes the penalized log-likelihood , in which the model maximum log likelihood is adjusted (shrunk) by a penalty factor.

$$logL - o.5\lambda\Sigma(s_i, \beta_i)^2$$

where L is the maximum likelihood of the fitted model, and λ a so-called penalty factor, β the estimated regression coefficient for each predictor i in the model, and s_i is a scaling factor for each β_i to make $s_i\beta_i$ unitless. Therefore, the estimated regression coefficients and predictive accuracy measures are directly (during model fit) adjusted for overoptimism.[12]

Chapter 5.

Model Misspecification and Bias

In the article by Rosenbaum and Rubin (1993), they demonstrated that conditioning on propensity score yields unbiased estimates of the expected difference in observed responses between two treatments. Therefore, conditioning on propensity score gives unbiased estimates of the true treatment effect if the measure of treatment effect is the differences in means or proportions. However, Austin (2007) found that conditioning on propensity score resulted in biased estimation of conditional odds ratios, hazard ratios, and rate ratios, and in all scenarios treatment effects were biased towards the null. In contrast, traditional regression methods allowed unbiased estimation of the true conditional treatment effect when all variables associated with outcome were included in the model. [6][8] A meta-analysis performed by Sturmer et al. (2006), reviewed 211 studies that performed regression analyses using propensity score methods. Their findings concluded that there was no empirical evidence that propensity score analyses controlled for confounding more effectively than conventional outcome modeling. Cook and Goldman(1989) compared the performance of propensity score models using tests of significance under the null hypothesis, which was compared to tests of significance under the null for coventional outcome models. In their results, propensity scores produced valid results in most scenarios, but were biased whe there was strong treatment-confounder assocaitions.[15] As stated by the Sturmer et al. (2006),

"The use of propensity scores as the only analytic technique applied comes at the price of losing potentially useful information about predictors of outcome. It therefore seems desirable to use PS only if a reduction in bias or an improvement in efficiency can be achieved." (Sturmer et al., 2006)

In contrast, Drake (1993) found that the magnitude and direction of bias resulting from omitting and important confounder were similar when comparing the multivariate outcome modeling to treatment effect

estimated after controlling for propensity score. This suggests that propensity score may not be beneficial in reducing bias compared to conventional multivariate outcome models if variable selection is not restricted by sample size. However, misspecifying the propensity score model results in smaller biases than misspecifying the response model. This is particularly true when a quadratic term is omitted. Drake(1993) found, via a simulation study, that when the propensity score is generated according to a quadratic logistic model but estimated by a linear logistic model, one observed biases ranging from 4.7% to 17%. Omitting the quadratic term from the logistic regression for the response model produced biases randing from 57.5% to 15.8%. For continuous response models, the biases ranged from 10.7% to 108.7%. The author summarized by stating, the value of the propensity score lies primarily in guarding against model misspecification. It does not reduce bias due to omitted confounders, however, the propensity score is preferable when concerned about model misspecifications in the response model. A distinction must be made, however, that the propensity score model estimates avergae effects, and conventional outcome models esimtate individual effects.[10]

Chapter 6.

Balance Diagnostics

To assess the performance of propensity score models, goodness-of-fit diagnostics for the adequacy of the model have been developed in the context of propensity-score matching and stratification on the propensity score. Similarity between treated and untreated subjects of baseline covariates indicates that propensity score matching has yielded a matched sample. Other studies have suggested the comparison of non-parametric density functions of the propensity score in treated and untreated subjects.[8][9] Similarly, balance diagnostics have been proposed for stratification based on propensity score quintiles, in which Rosenbaum and Rubin (1997) used two-way ANOVA models to regress each measured baseline covariate on propensity score quintiles. Other authors have proposed side-by-side boxplots of the distribution of propensity score, by propensity score quintile, to assess the differences in propensity for treatment based on propensity score quintile.Several studies have described the use of standardized differences to compare the distribution of measured baseline covariates to assess whether matching on propensity score has resulted in a matched sample. However, it was Austin (2008) that developed Goodness-of-fit tests for covariate adjustment methods. The proposed diagnostic for covariate adjustment is the standardized difference defined as:

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

for continuous variable, and by

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

In the simulation study performed by Austin (2008), a propensity score model was fit using logitist regression, in which treatment assignment was regressed on 27 baseline covariates. The estimated propensity score

for treated subjects ranged from 0.0903 to 0.904, while the estimated propensity scores for for untreated subjects ranged from 0.0758 to 0.8927. Non-parametric estimates of the distribution of propensity scores are illustrated in Figure 1. In this case, the distribution of propensity score in untreated subjects had a heavier left tail, the support of the distribution of the propensity score was similar in treated and untreated subjects. In this analysis, covariate adjustment was used, and therefore, all subjects were retained. The figure below shows a higher propensity for treatment for the treated subjects compared with the untreated subjects as well as less variance in propensity for treatment. [6]

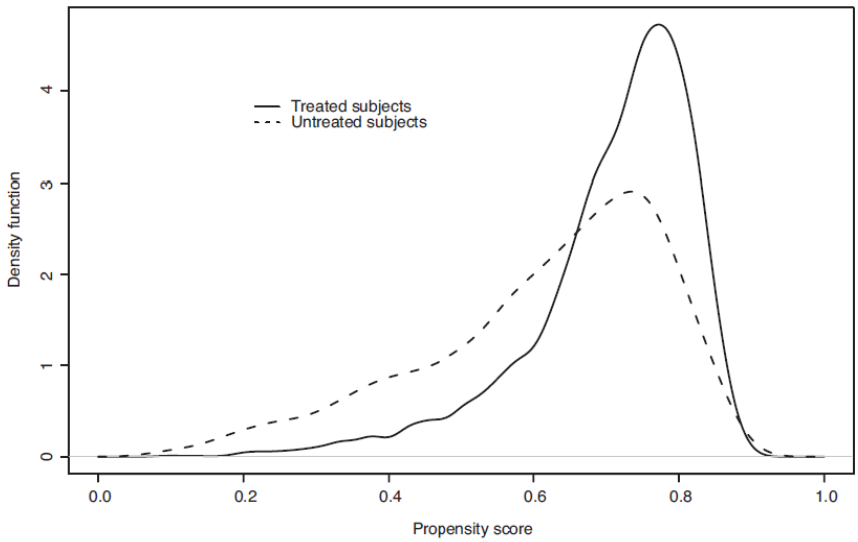

Figure 6..1: Distribution of the propensity score in treated and untreated subjects

The results of the simulation study showed that the crude standardized difference exceeded 0.1 for 22 of the 27 baseline covariates. It has been suggested in several articles that a standardized difference greater than 0.1 indicates potential imbalance in covariates between treated and untreated subjects. [6] Advantages of using the standardized difference metric is that it is not influenced by sample size, unlike t-tests proposed by Rubin and Rosenbaum (1997). In addition, it allows for comparison of balance in baseline covariates measured in different units. The empirical sampling distribution of the standardized difference is estimated, in large samples, as a normal distribution with mean δ and variance $(n_T + n_C)/n_T n_C + \delta/2(n_T + n_C)$, where n_C and n_T denoted the number of treated and control subjects. Under the null $(\delta = 0)$, then the variance is equal to $(n_T + n_C)/n_T n_C$. If we further assume equal sample sizes between treated and untreated groups, the variance of the standardized difference is $2/n$. Thus, 95% of the samples drawm from a population with a true $\delta = 0$ will lie within $\pm 1.96\sqrt{2/n}$. Austin (2008) states that

"Observing balance on means of main effecs alone does not ensure that bias has been eliminated from

estimating the treatment effect. Similarly, comparing the mean of interactions allowed one to determine whether the propensity score model has been incorrectly specified. This finding suggests that balance diagnostics needs to incorporate methods for comparing the distribution of measured baseline covariates between treated and untreated subjects."(Austin, 2008)

Therefore, Austin's (2008) recommendation is to make modifications to the propensity score model with the objective of making the estimated standardize differences lie within 2.5th and 97.5th percentiles of the empirical sampling distribution of standardized differences. The goal in modification is to have estimated standardized differences lie below a threshold consistent with the propensity score model that correctly balances baseline covariates. In smaller samples, it may be necessary to adjust for baseline covariates that cannot yield standardized differences below the threshold. [6][2]

Chapter 7.

Monte Carlo Simulations

Data Source

We used Monte Carlo simulations to examine the performance of different methods for propensity score modelling in a sample of $n = 40$ patients. Detailed clinical data were collected on a sample of 164 patients with rotator cuff tears that were seen at Bringham Women's Hospital, Mass General Hospital and Vanderbilt Medical Center. This sub-population was extracted from a larger population of patients that were seen for shoulder pain. These data were collected under the Rotator-Cuff Outcomes Workgroup (ROW) study, an ongoing initiative intended to improve the prognosis of treatment for rotator cuff tears. Furthermore, data were collected on attributes of patient tears, demographic characteristics, work, history, injury, lifestyle, pain measures, disability measures, functional measures, and mental health measures using MRI medical records, and standardized self-reported survey measures. Rotator cuff tear therapies include 2 methods of treatment: operative interventions and physical therapy sessions. Patients who receive surgical treatments for rotator cuff repairs, normally, are assigned physical therapy as a supplemental treatment to surgery. However, for the simplicity of this simulation, patients were separated into non-operative vs. operative treatment assignments. The covariates that were measured are listed below:

LongSize: size of rotator cuff tear measured longitudinally

TransSize: size of rotator cuff tear measured transversely

TearSize: size of rotator cuff tear

MHI: mental health score

Age: Age at baseline

Height: Height at baseline

MonthsOn: Number of months on study, ranges from 1 to 12

BMI: BMI index measure at baseline, ranges from

SympDur: Symptom Duration measured in months

MuscAtrophy: Binary indicator for presence of muscular atrophy

Cortisone: Binary indicator for use of cortisone injections

MRIBicep: Binary indicator for presence of tendonitis in Biceps

Fat Dep: Dichotomized variable for degree of Fatty Deposits in muscle (0 means degree below 2, 1 means degree of 2 or above)

Meds: Binary Indicator of use of medications

Gender: Binary Indicator for gender

Smoker: Binary Indicator for smoking status (0 means never, and 1 means former or current)

Comorbidities: Binary Indicator for presence of any comorbidities

NumberTears: Binary Indicator for whether cause was due to injury Number of tears, ranging from 0 to 3

Multiple Imputation of missing baseline values

For the $n = 164$ subjects, missing values of baseline covariates were imputed using multiple imputation via predictive mean matching (ref Little RJA, Rubin DB: Statistical Analysis with Missing Data. New York: Wiley, 1987) to avoid casewise deletion of patient records missing any covariate. First missing values were initialized using a random sample of the observed values. Then a bootstrap sample was taken, and for each variable with missing data, a flexible additive model was fit with that variable as the outcome, yielding a fitted value for every patient record. Each missing value was imputed using predictive mean matching, in which missing values are imputed with the observed patient value closest to its valuepredicted by the imputation model. This process was repeated 15 times, each time initialized using the previous step's results. The original patient data, with missing values replaced by the final imputed values, were used to fit the primary analysis models. To account for variability associated with the imputation procedure, coefficient standard error estimates were adjusted by multiplying by a constant function of the within and between imputation variances. Once data were complete using multiple imputation, we randomly generated a sample of $n = 5,000$ by resampling from the original 164 participants.

Showing a true null treatment effect

Using the sample size of $n = 5,000$, we checked the true associations between treatment effect and SPADI in several ways. In a regression of SPADI on treatment alone, we expect to see a significant treatment effect due to not adjusting for any of the significant baseline covariates. In the second regressional analysis, we regress SPADI on all baseline covariates that are associated with SPADI, and expect to see a non-significant treatment

effect due to adjusting for significant covariates. In contrast, the third regressional analysis regresses SPADI on baseline covariates only associated with treatment, and in this case, we would expect to see a significant treatment effect. Based on previous literature by Rubin (1987) and Austin (2006), variables only associated with treatment do not add any benefit or information to regressional analysis of the outcome, and may reduce efficiency and accuracy. Lastly, a regressional analysis is performed which regresses SPADI on some of the associated baseline covariates, without significant interactions included in the model. In this case, we expect to see a null treatment effect due to the significant baseline covariates for which we adjusted. In the larger sample size of $n = 5,000$, these expectations must be met to show that in larger samples the true effect can be found when we have adjusted for other significant baseline covariates.

1.$SPADI \sim trt$

2.$SPADI \sim trt + LongSize + AgeOn \times (TearSize + BMI) + Cortisone + FatDeps + Smoker$

3.$SPADI \sim trt + MHI + AgeOn \times TearSize + BMI + FatDeps + Gender + NumTears + Comorbidities$

4.$SPADI \sim trt + AgeOn \times TearSize + FatDeps$

Simulation Methods For each sample size assignment, we performed $B = 1000$ repetitions of each simulation. The treatment effect was pre-determined to either have a null treatment effect= 0 and a non-null effect treatment effect= 1 in the pre-specified relationship with SPADI. The propensity scores were generated by four methods described below:

1. Propensity scores generated using three main effects deemed clinically
2. important by subject knowledge, and the first principal component of all remaining baseline variables. The three cliniaclly significant effects are correctly specified and have a significant association with SPADI.

3. Propensity scores generated using three main effects deemed clinically important by subject knowledge, and the first principal component of all remaining baseline variables. The three cliniaclly significant effects are incorrectly specified and do not have a significant association with SPADI.

4. Propensity scores generated using four principal components that were modeled using all baseline covariates.

5. Propensity scores generated using penalized maximum likelihood estimation including all baseline covariates.

To assess the performance of the propensity scores in balancing baseline covariates. The distribution of propensity scores between treatments was assessed graphically. The propensity score subclassication assigned 50% of the sample to physical therapy, and 50% to surgery based on their propensity score. Patients with higher propensity scores were assigned to the surgery group. Therefore, to diagnose the performance of propensity scores, it would be expected that the surgery group would have higher propensity for surgery compared to the physical therapy group, and it would also be expected that there were no significant differences in baseline covariates based on the propensity score subclassification. The graph below shows a higher propensity for surgery in the surgery group, which is an indication that the propensity score has correctly modeled higher propensity scores for the surgery group in contrast to the physical therapy group.

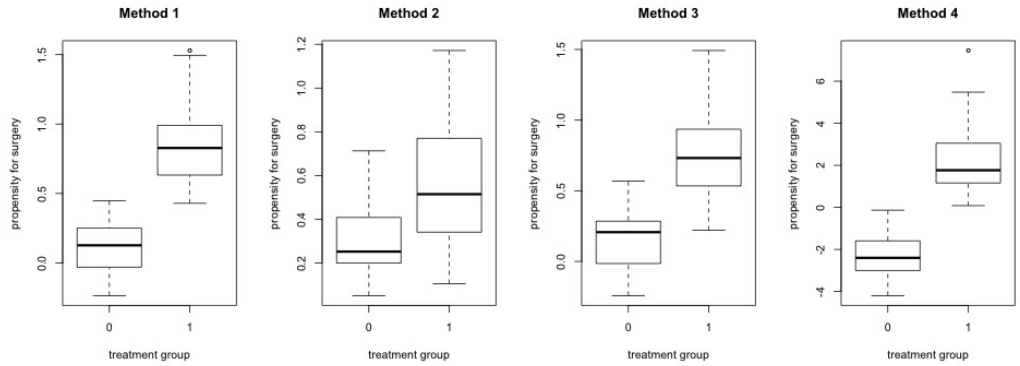

Figure 7..1: Distribution of the propensity score in surgery vs. physical therapy patients

Following the generation of the four different propensity scores. A linear regression analysis is performed, where SPADI is regressed on treatment assignment plus the respective propensity score. From each regression, the estimated treatment effect and p-value are extracted. Therefore, from the $B = 1000$ repetitions, there are 1000 associated p-values for treatment effect for each propensity score method. Performance of the propensity score model was defined as the ability to correctly identify a significant or non-significant treatment effect. Therefore, we measured the proportion of times each propensity score method incorrectly idenitfied a significant treatment effect when a null treatment effect had been assigned. Conversely, we also measured the proportion of times the propensity score method had yielded a non-significant treatment effect when a non-null treatment effect had been assigned.

Chapter 8.

Results:

Refer to Table 1, Table 2, and Table3 for summaries

8.1. Treatment and Outcome associated with the same baseline covariates

In the case where both treatment and outcome are associated with the three pre-specified main effects of (1) age at baseline, (2) degree of fatty deposits, and (3) tear size, we applied the four methods of variable selection and assessed the performance of the different propensity scores to yield a correct main effect for treatment. Three scenarios were tested, one for which we set treatment to have a null effect, secondly, we set treatment to have a coefficient of 1, and lastly, we set treatment to have a smaller main effect with a coefficient of 0.5.

The results showed that when treatment and outcome were both associated with the three main effects of (1)age at baseline, (2) degree of fatty deposits, and (3) tear size, then PS Method 1 correctly yielded null treatment effects 100% of the 1000 repeated simulations. This is due to the fact that PS Method 1 models propensity score with the same three main effects that are associated with both SPADI and treatment in addition to the first principal component. However, PS Method2 did not yield a correct treatment effect for any of the 1000 simulations, which signifies that when the propensity score method includes three main effects that are not associated with either outcome or treatment, the treatment was effect is not correctly identified as non-significant for any of the simulations. PS Method 3 did not improve from PS Method 2. Out of 1000 simulations, 976 simulations yielded an incorrect significant treatment effect. PS Method 3 uses four princi-

pal components. Therefore, using principal components instead of main effects did not benefit the propensity score model in this case. Lastly, PS Method 4, propensity score with penalized maximum likelihood yielded only eight significant treatment effects out of 1000 simulations. Therefore, PS Method 5 correctly yielded non-significant treatment effects a majority of the time.

The results also showed that when treatment and outcome were both associated with the three main effects of (1)age at baseline, (2) degree of fatty deposits, and (3) tear size, all four propensity score methods were able to correctly identify a significant treatment effect for almost 100% of the simulations. PS Method 1, PS Method 2, and PS Method 3, identified significant treatment effects for all 1000 simulations. PS Method 4 identified a significant treatment effect for 993 out of 1000 simulations. Therefore, when we have a non-null treatment effect of 1, all four methods were able to perform well even if the propensity score model does not include main effects that are related to SPADI or treatment.

In the case where both treatment and outcome are associated with the three pre-specified main effects of (1) long size, (2) muscular atrophy, and (3) bicep tendonitis, we applied the four methods of variable selection and assessed the performance of the different propensity scores to yield a correct main effect for treatment. Two scenarios were tested, one for which we set treatment to have a null effect, secondly, and we set treatment to have a coefficient of 1.

The results also showed that when treatment and outcome were both associated with the three main effects of (1)long size, (2) muscular atrophy, and (3) bicep tenditinis, then PS Method 1 incorrectly yielded non-null treatment effects 100% of the 1000 repeated simulations. This is due to the fact that PS Method 1 models propensity score with three main effects that are not associated with both SPADI and treatment in addition to the first principal component. However, PS Method2 did yield a correct treatment effect for a 100% of the 1000 simulations, which signifies that when the propensity score method includes three main effects that are associated with either outcome or treatment, the treatment was effect is correctly identified as non-significant for a majority of the simulations. PS Method 3 performed as badly as PS Method 1 in that it identified a 100% of the treatment effects as significant .Therefore, using principal components instead of main effects did not benefit the propensity score model in this case. Lastly, PS Method 4, propensity score with penalized maximum likelihood yielded only 12 significant treatment effects out of 1000 simulations.

The results above are evidence to suggest that when the three clinically main effects, included in the propensity score model, were correctly specified, meaning they are truly associated with SPADI, the propensity score model that included the correct set of variables performed the best. However, when the three clinically main effects were not the same as those associated with SPADI, the propensity score model did poorly. In both cases, the penalized maximum likelihood also performed optimally. Lastly, the propensity score of four principal components did not perform well in any case.

8.2. Treatment and Outcome with no baseline covariates in common

In the case where treatment is associated with the three pre-specified main effects of (1) age at baseline, (2) degree of fatty deposits, and (3) tear size, and SPADI is associated with three different effects of (1) long size, (2) muscular atrophy, and (3) bicep tendonitis, we applied the four methods of variable selection and assessed the performance of the different propensity scores to yield a correct main effect for treatment. Two scenarios were tested, one for which we set treatment to have a null effect, secondly, and we set treatment to have a coefficient of 1.

We expect that the propensity score method that includes main effects that are also associated with SPADI (PS Method 2) to have the lowest number of incorrect treatment effects. We would also expect that PS Method 1 to have a higher number of incorrectly identified significant treatment effects compared with PS Method 2 due to the fact that PS Method 1 does not include main effects associated with SPADI, but does include those associated with treatment. The results showed that when treatment and outcome were associated with three separate and different main effects, then PS Method 1 incorrectly yielded non-null treatment effects for 30 out of the 1000 repeated simulations. However, PS Method2 did yielded a correct treatment effect for a 100% of the 1000 simulations, which signifies that when the propensity score method includes three main effects that are associated with either outcome, it performs better than the propensity score model that does not include these main effects. PS Method 3 yielded incorrect significant treatment effects for 28 out of the 1000 simulations. Lastly, PS Method 4 yielded incorrect significant treatment effects for 52 out of the 1000 simulations. PS Method 4 performed significantly worse that the other three methods.

The results showed that when treatment was associated with (1) age at baseline, degree of fatty deposits, and tear size, and SPADI was associated with longsize, bicep tendonitis, and muscular atrophy, PS Method2

performed the best, yielding 978 significant treatment effects compared to 443 for PS Method 1, and 518 for PS Method 3. Similarly to its performance in the null case, PS Method 4 performed the worse with 12 correctly identified significant treatment effects out of 1000 simulations.

In the case where treatment is associated with the three pre-specified main effects of (1) long size, (2) muscular atrophy, and (3) bicep tendonitis, and SPADI is associated with three different effects of (1) at at baseline, (2) degree of fatty deposits, and (3) tear size, we applied the four methods of variable selection and assessed the performance of the different propensity scores to yield a correct main effect for treatment. Three scenarios were tested, one for which we set treatment to have a null effect, secondly, we set treatment to have a coefficient of 1.

We expect that the propensity score method that includes main effects that are also associated with SPADI (PS Method 1) to have the lowest number of incorrect treatment effects. We would also expect that PS Method 2 to have a higher number of incorrectly identified significant treatment effects compared with PS Method 2 due to the fact that PS Method1 does not include main effects associated with SPADI, but does include those associated with treatment. The results showed that when treatment and outcome were associated with three separate and different main effects, then PS Method 1 correctly yielded null treatment effects for 100% of the 1000 repeated simulations. However, PS Method2 did not yielded a correct treatment effect for a 22 of the 1000 simulations, which signifies that when the propensity score method includes three main effects that are not associated with outcome, it does not perform as well as the propensity score model that includes variables associated with outcome. PS Method 3 yielded incorrect significant treatment effects for 22 out of the 1000 simulations. Lastly, PS Method 4 yielded incorrect significant treatment effects for 6 out of the 1000 simulations. In this case, PS Method 4 performed better than both PS Method 2 and 3. This pattern is in contrast to the results stated above where PS Method 4 was the worst performing propensity score model.

The results showed that when treatment was associated with (1) long size, (2) bicep tendonitis, and (3) muscular atrophy, and when SPADI is associated with (1) age at baseline, (2) degree of fatty deposits, and (3) tear size, PS Method 1 yielded correct significant treatment effects for 997 out 1000 simulations, and PS Method 2, 3 and 4 yielded correct significant results for 386, 892, and 118 out of 1000 simulations, respectively.

The results from both the null and non-null treatment effects, in this case, demonstrate that contrary to our expections, both PS Method1 and PS Method 2 resulted in a low number of false significant treatment effects, regardless of the variables that were included in the propensity score model. Therefore, contrary to results previously found by Rubin and Austin, we have evidence to suggest that it is not only important to include

variables associated with outcome, but that variables associated with exposure are also useful in developing a precise propensity score model

8.3. Main effects associated with SPADI are a subset of main effects associated with treatment

In the case where treatment is associated with the three pre-specified main effects of (1) age at baseline, (2) degree of fatty deposits, and (3) tear size, (4) age at baseline, (5) degree of fatty deposits, and (6) tear size, and SPADI is associated with three different effects of age at baseline, tear size, and degree of fatty deposits, which is a subset of those associated with treatment, we applied the four methods of variable selection and assessed the performance of the different propensity scores to yield a correct main effect for treatment. Three scenarios were tested, one for which we set treatment to have a null effect, secondly, we set treatment to have a coefficient of 1.

We expect that the propensity score method that includes main effects that are also associated with SPADI (PS Method 1) to have the lowest number of incorrect treatment effects. We would also expect that PS Method 1 to have a higher number of incorrectly identified significant treatment effects compared with PS Method 2 due to the fact that PS Method1 does not include main effects associated with SPADI, but does include those associated with treatment. This will confirm previously literature that has stated main effects associated with treatment do not add benefit to propensity score models. The results showed that PS Method 1 correctly yielded non-null treatment effects for all of the 1000 repeated simulations. However, PS Method2 did not yield a correct treatment effect for a 857 of the 1000 simulations, which signifies that when the main effects associated with outcome are a subset of variables associated with treatment, it is necessary to have a propensity score model that includes the main effects associated with outcome. PS Method 2 used the three main effects associated with treatment, but not associated with outcome. Therefore, it performed substantially badly. In contrast PS Method 3, using four principal components, yielded 246 incorrect significant treatment effects, and PS Method 4 yielded on 5 incorrect significant treatment effects. Therefore, in this case, PS Method 4 performed almost as well as the propensity score model that included the correct main effects associated with outcome.

When this scenario included a non-null main effect for treatment, PS Method 1, 2, and 3, all yielded 100% significant treatment effects for the 1000 simulations. PS Method 4 only yielded 522 significant treatment effects out of the 1000 simulations.

In the case where treatment is associated with the three pre-specified main effects of (1) age at baseline, (2) degree of fatty deposits, and (3) tear size, (4) age at baseline, (5) degree of fatty deposits, and (6) tear size, and SPADI is associated with three different effects of long size, muscular atrophy, and bicep tendonitis, which is a subset of those associated with treatment, we applied the four methods of variable selection and assessed the performance of the different propensity scores to yield a correct main effect for treatment. Three scenarios were tested, one for which we set treatment to have a null effect, secondly, we set treatment to have a coefficient of 1.

We expect that the propensity score method that includes main effects that are also associated with SPADI (PS Method 2) to have the lowest number of incorrect treatment effects. We would also expect that PS Method 2 to have a higher number of incorrectly identified significant treatment effects compared with PS Method 1 due to the fact that PS Method 2 does not include main effects associated with SPADI, but does include those associated with treatment. Therefore, it will yield opposite results to those shown above. The results showed that PS Method 2 correctly yielded non-null treatment effects for all of the 1000 repeated simulations. However, PS Method 1 did not yield a correct treatment effect for a 881 of the 1000 simulations, which confirms our expected results that are opposite of what was shown in the previous simulation. PS Method 3 and 4 yielded 796 and 27 incorrect significant treatment effects, respectively. This pattern is similar to that shown previously, where PS Method 4 performed closely to the correct propensity score model.

When this scenario included a non-null main effect for treatment, PS Method 1, 2, and 3, all yielded 100% significant treatment effects for the 1000 simulations. PS Method 4 only yielded 522 significant treatment effects out of the 1000 simulations. These results are exactly those that were found in the prior simulation.

The results showed that when treatment was associated with (1) long size, (2) bicep tendonitis, and (3) muscular atrophy, and when SPADI is associated with (1) age at baseline, (2) degree of fatty deposits, and (3) tear size, PS Method 1 yielded correct significant treatment effects for 997 out 1000 simulations, and PS Method 2, 3 and 4 yielded correct significant results for 386, 892, and 78 out of 1000 simulations, respectively.

The results from both the null and non-null treatment effects and in both scenarios of associations, we demonstrated that when the outcome variable is associated with a subset of the variables associated with treatment, it is important to specify the correct main effects in the propensity score model. PS Method 3 performed badly under a null treatment effect, however, it performed well under the non-null treatment effect. In contrast to

PS Method 4, which performed significantly well under a null treatment effect, and extremely poorly under a non-null treatment effect. From this we see the penalized maximum likelihood will underestimated treatment effects, and therefore, yielded a higher number of insignificant results.

8.4. Main effects associated with treatment are a subset of main effects associated with outcome

In the case where treatment is associated with the three pre-specified main effects of (1) age at baseline, (2) degree of fatty deposits, and SPADI is associated with (1) age at baseline, (2) degree of fatty deposits, (3) tear size, (4) long size, (5) bicep tendonitis, and (6) muscular atrophy, it is evident that the main effects associated with treatment are a subset of effects asscoaited with SPADI. We applied the four methods of variable selection and assessed the performance of the different propensity scores to yield a correct main effect for treatment. Three scenarios were tested, one for which we set treatment to have a null effect, secondly, we set treatment to have a coefficient of 1.

We expect that PS Method 1 will perform the best out of the four propensity score methods due to the fact that the three main effects associated with treatment and SPADI are included in the propensity score model. We also expect that PS Method 2 will perform the worst out of the four models because the three main effects included are associated with SPADI but have no association to treatment. Lastly, we expect PS Method 4 to perform well relatively to PS Methods 2 and 3. The results showed that PS Method 1 performed the best with 10 out of 1000 simulations yielding incorrect significant treatment effects. PS Method 4 performed second best with 15 out of 1000 simulations yielding incorrect significant treatment effects. Lastly, PS Method 2 and 3 yielded 973 and 321 incorrect significant treatment effects out of 1000 simulations, respectively.

When this scenario included a non-null main effect for treatment, we expect PS Method 1 to yield the most number of true significant treatment effects out of 1000 simulations. We also expected to PS Method 4 to yield the highest number of false null treatment effects due to the penalized maximum likelihood shrinkage of treatment effects. However, the results showed that PS Method 1 performed badly with 385 non-null treatment effects out of 1000 simulations. PS Method 2 in contrast performed well with 100% of the 1000 simulations yielding non-null treatment effects. PS Method 3 performed second best with 974 non-null treatment effects and PS Method 4 performed significantly worse than the other three methods with only 43 non-null treatment effects out of 1000 simulations. Possible reasons for this discrepancy is that in PS Method 2, the three main

effects included in the model are not associated with treatment, but are associated with SPADI. Therefore, when treatment and propensity score is regressed on SPADI, propensity score does not add information, and treatment effects are more significantly associated with SPADI.

In the case where treatment is associated with the three pre-specified main effects of (1) long size, (2) bicep tendonitis, and (3) muscular atrophy, and SPADI is associated with the same three main effects and (4) age at baseline, (5) degree of fatty deposits, and (6) tear size, , we applied the four methods of variable selection and assessed the performance of the different propensity scores to yield a correct main effect for treatment. Three scenarios were tested, one for which we set treatment to have a null effect, secondly, we set treatment to have a coefficient of 1, and lastly, we set treatment to have a minimal main effect 0.5.

We expect the propensity score that includes the same three main effects as those associated with SPADI (PS Method 2) to yield the lowest number of false significant treatment effects. We would also expect PS Method 4 to yield a low number of false significant treatment effects due to the shrinkage of treatment effects by penalized maximum likelihood. The results confirmed these expectations with PS Method 2 and 4 both yielding 21 and 2 false significant effects out of 1000 simulations, respectively. What was not expected was that PS Method 4 would outperform PS Method 2. PS Method 1 and 3 yielded worse results with 956 and 847 false significant treatment effects, respectively.

In the case where treatment is associated with the three pre-specified main effects of (1) long size, (2) muscular atrophy, and (3) bicep tendonitis, we expect that the propensity score method that does not include these main effects to yield the highest number of significant treatment effects when treatment and propensity score is regressed on SPADI. We expect PS Method 4 to yield the lowest number of significant treatment results due to the shrinkage of treatment effects. The results showed that PS Method 1, which does not include the three main effects associated with SPADI, yielded the highest number of significant treatment effects. For both PS Method1 and 3 100% of the 1000 simulations yielded significant treatment effects. PS Method 2 yielded 304 significant treatment effects out of 1000 simulations, and PS Method 4 yielded 108 significant treatment effects out of 1000 simulations. Therefore, the results confirmed that PS Method 1 and 3 performed the best, and PS Method 4 performed the worst out of the four propensity score methods.

The results from the last two simulations demonstrate that when the variables associated with treatment are a subset of variables associated with outcome, there are a higher number of false significant effects when using the propensity score model that includes the variables not associated with treatment. This confirms our above results that suggest that variables associated with treatment and outcome are important in propensity score modeling. For this simulation, the propensity score model that did not include variables associated

with treatment performed significantly worse than the propensity score model that included the same set of variables.

8.5. Mixed main effects for both treatment and outcome

In the case where treatment is associated with the three pre-specified main effects of (1) age at baseline, (2) MHI, (3) muscular atrophy, (4) degree of fatty deposits, (5) tear size $\times$ age at baseline, (6) degree of fatty deposits $\times$ age at baseline, and SPADI is associated with (1) MHI, (2) tear size, (3) cortisone, (4) degree of fatty deposits, (5) Meds, (6) MHI $\times$ age at baseline, there are some common main effects between treatment and SPADI (age at baseline, MHI, degree of fatty deposits), and the other main effects differ. We applied the four methods of variable selection and assessed the performance of the different propensity scores to yield a correct main effect for treatment. Three scenarios were tested, one for which we set treatment to have a null effect, secondly, we set treatment to have a coefficient of 1.

In the case where we set the treatment effect to be a null effect, we expect PS Method 4 to yield the lowest number of non-significant treatment effects out of 1000 simulations. We expect this becuase there is a mixture of main effects between treatment and SPADI, and we expect the penalized maximum likelihood to result in the lowest number of incorrect treatment effects. We also expect PS Method 1 to yield a lower number of significant treatment effects that PS Method 2 because the propensity score model for PS Method 1 includes more main effects associated with treatment that PS Method 2. The results showed that PS Method 4 was the best model, which yielded 45 out of 1000 simulations with false significant treatment effects. PS Method 1 performed better than PS Method 2, with 111 incorrect treatment effects over 991 from PS Method 2. Lastly, PS Method 3 performed moderately with 188 false significant treatment effects out of 1000 simulations.

For the non-null treatment effect, we expect PS Method 1 to have a lower number of non-null treatment effects compared to PS Method 2 because the additional association between main effects and treatment in PS Method 1 results in the treatment effect being underestimated. However, PS Method2, there is no additional information added to the regression on SPADI by propensity score, when there is little association between treatment and propensity score. Therefore, treatment drives a majority of the association with SPADI. The results showed that PS Method 1 yielded 600 non-null results compared to PS Method 2 with 996 non-null results out of 1000 simulations. Therefore, our expectation of a better performed by PS Method 2 was confirmed. However, the worst performing model was PS Method 4 that yielded 384 non-null treatment effects.

PS Method 1 performed significantly better than PS Method 2 because there were a higher number of associated variables with outcome in PS Method 1. This confirms the hypothesis that variables related to outcome are essential to develop a useful propensity score. In addition, PS Method 4 performed better than the other three propensity score models. Therefore, using penalized maximum likelihood is a sufficient alternative when there is no prior knowledge of associated main effects.

Chapter 9.

Summary Findings

Table 9..1: Propensity Score Performance for the Null Treatment Effect

Null Treatment Effect for $n = 40$				
	PS Method 1 (set1)	PS Method 2 (set2)	PS Method 3 (4 PCs)	PS Method 4 (PMLE)
Trt and SPADI (set 1)	0	1000	976	8
Trt and SPADI (set 2)	1000	0	1000	12
Trt (set1) and SPADI (set2)	30	0	28	12
Trt (set2) and SPADI (set1)	0	22	22	6
Trt (set1, set 2) and SPADI (set1)	0	857	246	5
Trt (set1, set 2) and SPADI (set2)	881	0	796	27
Trt (set1) and SPADI (set1, set2)	10	973	321	15
Trt (set1) and SPADI (set1, set2)	111	991	188	45

The sets in the parantheses , following treatment and SPADI, refer to the set associated with each variable, respectively. Set 1 refers to the set of variables, which include age at baseline, tear size, and degree of fatty deposits. Set 2 refers to the set of variables, which include long size, muscular atrophy and bicep tendonitis.

Table 9..2: Propensity Score Performance For the non-null treatment Effect

Non-Null Treatment Effect for $n = 40$				
	PS Method 1 (set1)	PS Method 2 (set2)	PS Method 3 (4 PCs)	PS Method 4 (PMLE)
Trt and SPADI (set1)	1000	1000	1000	993
Trt and SPADI (set2)	1000	989	1000	874
Trt (set1) and SPADI (set2)	443	978	518	12
Trt (set2) and SPADI (set1)	997	386	892	118
Trt (set1 set 2) and SPADI (set1)	1000	1000	1000	522
Trt (set1, set2) and SPADI (set2)	1000	1000	1000	535
Trt (set1) and SPADI (set1, set2)	385	1000	974	43
Trt (set2) and SPADI (set1, set2)	1000	304	1000	108

The sets in the parantheses , following treatment and SPADI, refer to the set associated with each variable, respectively. Set 1 refers to the set of variables, which include age at baseline, tear size, and degree of fatty deposits. Set 2 refers to the set of variables, which include long size, muscular atrophy and bicep tendonitis. Without controlling for type I error, reviewing power of a statistical test is not meaningful.

Table 9..3: Propensity Score Performance for the Mixed Treatment Effect

Mixed Treatment and Outcome Effects for $n = 40$				
	PS Method 1 (set1)	PS Method 2 (set2)	PS Method 3 (4 PCs)	PS Method 4 (PMLE)
Trt and SPADI (set1, set2)(mixed)	111	991	188	45
Trt and SPADI (set1, set2) (mixed)	600	996	717	384

The sets in the parantheses , following treatment and SPADI, refer to the set associated with each variable, respectively. Set 1 refers to the set of variables, which include age at baseline, tear size, and degree of fatty deposits. Set 2 refers to the set of variables, which include long size, muscular atrophy and bicep tendonitis.

Chapter 10.

Conclusion

In the case where we know the associations of baseline variables to treatment and outcome, the propensity score model that includes variables associated with either treatment or outcome performed better than the other propensity score methods. Contrary to findings of previous research, only including variables associated with outcome resulted in a higher number of falsely idenitified treatment effects. However, to avoid misspecification of main effects, the penalized maximum likelihood propensity score model performed as well as the propensity score model with correctly specified main effects. Therefore, the recommendation based on our simulation results, is to use penalized maximum likelihood propensity score models in order to avoid yielding a higher number of incorrectly idenitified treatment effects. Lastly, future research may explore questions that have been raised by the results of this simulation. Some of the possible questions to be answered are: 1. how do changes in sample size affect the propensity score methods used in this simulation? Increases in sample size may yield better performances in alternative propensity score models. 2. What are the reasons why the principal component propensity score model (PS Method 3) performed significantly worse compared to the other three models? And lastly, future research should explore how the omission of quadratic terms may affect the performance of these four propensity score methods.

BIBLIOGRAPHY

[1] M. Alan Brookhart et al. Variable Selection for propensity score models. American Journal of Epidemiology. 2006. 163(12).1149-1156.

[2] Peter C. Austin and Muhammad M. Mamdani. A comparison of propensity score methods: A case-study estimating the effectiveness of post-AMI statin use.Statistics in Medicine. 2006. 25. 2084-2106.

[3] Peter C. Austin. The performance of different propensity score methods for estimating marginal odds ratios. Statistics in Medicine. 2007.26. 3078-3094.

bibitemAustin07 Peter C. Austin, Paul Grootendorst, Geoffrey M. Anderson. A comparison of the ability of different propensity score models to balance measured variables between treated and untreated subjects: A Monte Carlo Study. Statistics in Medicine. 2007. 26.734-753.

[4] Peter C. Austin, Paul Grootendors, tet al. Conditioning on the propensity score can result in biased estimation of common measures of treatment effect: A Monte Carlo study. Statistics in Medicine.2007.26.754-768.

[5] Peter C. Austin. Goodness-of-fit diagnostics for the propensity score model when estimating treatment effects using covariate adjustment with the propensity score. PHARMACOEPIDEMIOLOGY AND DRUG SAFETY. 2008. 17. 1202-1217.

[6] Peter C. Austin. Balance diagnostics for comparing the distribution of baseline covariates between treatment groups in propensity-score matched samples. Statistics in Medicine. 2009. 28. 3083-3107.

[7] Peter C. Austin. An Introduction to Propensity Score Methods for Reducing the Effects of Confounding in Observational Studies. Multivariate Behavioral Research. 2011. 46. 399-424.

[8] D'Agostino, Ralph B. Tutorial in Biostatistics: Propensity score methods for bias reduction in the comparison of a treatment to a non-randomized control group. Statistics in Medicine. 1998.17,2265-2281.

[9] Drake, Christiana. Effects of Misspecification of the Propensity Score on Estimators of Treatment Effect. Biometrics.1993. 49. 4. 1231-1236.

[10] Perkins SM, Tu W, Underhill MG, Zhou XH, Murray MD. The use of propensity scores in pharmacoepidemiologic research. Pharmacoepidemiology and Drug Safety. 2000; 9: 93-101.

[11] K.G.M.Moons, A. Rogier T. Dinders, E.W. Steyerburg, F.E.Harrell. Penalized maximum likelihood estimation to directly adjust diagnostic and prognostic prediction models for overoptimism: a clinical example. Journal of Clinical Epidemiology. 2004.57.1262-1270.

[12] Rosenbaum PR., Rubin DB. The central role of the propensity score in observational studies for causal effects. Biometrics.1983.70.41-55.

[13] Donald B. Rubin. Estimating Causal Effects from Large Data Sets Using Propensity Scores. Annals of Internal Medicine. 1997. 127.757-763.

[14] Til Sturmer et al. A review of the application of propensity score methods yielded increasing use, advantages in specific settings, but not substantially different estimates compared with conventional multivariable methods. Journal of Clinical Epidemiology. 2006. 59. 437-447.

[15] Rubin DB, Thomas N. Matching using estimated propensity score relating theory to practice. Biometrics. 1996; 52:249-264.

[16] Williams JR, Holleman DR Jr., Simel DL. Measuring Shoulder Function with Shoulder Pain and Disability Index. Journal of Rheumatology. 1995. 22(4). 727-732.

[17] MacDermid JC, Drosdowech D, Faber K. Responsiveness of self-report in patients recovering from rotator-cuff tear surgery. Journal of Shoulder and Elbow Surgery. 2006. 15(4).407-414.

[18] Muirhead RJ. Aspects of Statistical Theory. Wiley: New York, NY.2005.