The Pennsylvania State University
The Graduate School
Eberly College of Science

REINVESTIGATION OF CAUSAL EFFECTS OF RIGHT HEART CATHETERIZATION: A MATCHING APPROACH

A Thesis in
Statistics
by
Qiong Yang

Submitted in Partial Fulfillment
of the Requirements
for the Degree of

Master of Science

August 2014

The thesis of Qiong Yang was reviewed and approved* by the following:

Debashis Ghosh
Professor of Statistics
Thesis Advisor

Spiro E. Stefanou
Professor of Agricultural Economics

Aleksandra B. Slavkovic
Associate Professor of Statistics, Associate Head for Graduate Studies

*Signatures are on file in the Graduate School.

Abstract

Right Heart Catheterization (RHC) is a common procedure applied in critically ill patients. In US, over 1 million cases of RHC procedures are conducted annually at present. In particular, patients with low blood pressure, lung water, hearts abnormalities, kidney abnormalities are usually the targeted group to receive the procedure. Although there is no absolute clinical contradiction with the use of RHC, its effect has not been statistically validated using randomized controlled trials due to the lack of randomized data. Instead, studies on RHC have been using observational data to quantify its causal effect, which should be taken with caution because of the issues in estimating true causal effect in observational studies.

In my study, I reinvestigate the causal effect of RHC on subsequent survival using observational data and several multivariate matching schemes including nearest neighbor matching, optimal matching, full matching and genetic matching. Data for the study is from the Study to Understand Prognoses and Preferences for Outcomes and Risks of Treatments (SUPPORT), with a total of 5735 subjects.

Our study has three main findings. First, RHC increased the risk of dth30, after controlling for the observed confounders under each of the four matching schemes, which coincides with earlier studies using SUPPORT data. Second, the Monte Carlo simulation experiment conducted in our analysis suggests that our conclusion on the causal effect of RHC is robust because: 1) The bias, defined by the difference between true and estimated causal effect, is small with acceptable variance; 2) Our simulation results are robust to a variety of specifications on additivity and/or linearity in the relationship between confounders and the treatment indicator. However, for the case where strong nonlinearity is present, the model performance is not as good as in other scenarios. Third, this negative causal effect is sensitive to the possible hidden bias identified by sensitivity analysis. However, the performances of these four matching schemes are different confronting hidden bias. Full matching gives the most robust result compared to other three matching schemes.

Table of Contents

List of Figures

List of Tables

Chapter 1
Introduction

Right Heart Catheterization(RHC) is a common procedure applied in critically ill patients and has demonstrated its usefulness in many diagnostic applications. As a result of its wide applicability, great accuracy and convenience in measuring cardiac events, it has been serving as an important tool in modern day medicine.

Although there is no absolute clinical contraindication for use of RHC, it still needs to be implemented with caution for both clinical and statistical reasons. For the clinical reason, the risks associated with RHC are high when applying to patients with certain category of preexisting diseases including severe pulmonary hypertension and in the elderly.

For the statistical reason, the wide adoption of RHC in clinical practice has not been statistically validated in randomized controlled trials. As a surgical procedure, it is unethical to perform randomized trials to evaluate its effects and physicians would not submit enough patients for randomization. As a consequence for these two facts, there are no definitive guidelines for performing RHC or controlled clinical trials demonstrating its medical benefit and thus the value and necessity of performing routine RHC for coronary artery disease have been questioned.

Tremendous efforts from various disciplines are made to remove the possible effects of baseline differences among covariates between treatment and control units. Traditionally, methods are based on modeling the potential outcomes by making implicit assumptions on how potential outcomes are related to covariates. In particular, these efforts include difference in means, ANCOVA(main effects only or with interaction) and regression estimation. Another body of methods for discerning

causality is via matching based on propensity scores which balance the distributions of the covariates in the sense that the treated and untreated subjects with similar propensity scores have similar distributions for all the covariates [e.g., see (Rosenbaum and Rubin, 1983)].

One major difference between the two approaches is that the former approach makes some assumptions on the distributions of potential outcomes, while the latter makes assumptions mainly on distribution of the treatment indicator. Two common critiques for regression approach are:

- the extent of overlap among covariates between treated and control units is not addressed; and
- fine tuning (e.g., transformations of variables and variable selection) which is typically involved in the approach may potentially lead to some forms of researcher bias.

By contrast, matching addresses the problem of overlapping directly via the use of propensity scores. Interested readers may refer to Stuart (2010) where the relative merits for each approach are further discussed.

In this study, our main focus is matching. Intuitively, matching selects matched subsample of treated and untreated subjects whose covariate distributions are similar enough based on some distance metrics. Therefore the potential effects of selection bias will be minimized. The matched sample may resemble a randomized experiment. Further analysis is thus conducted based upon the post-matched sample instead of the original data. Details will be discussed in later chapters. We will use an empirical example to conduct several matching methods and the subsequent outcome analysis.

The rest of the thesis is organized as follows. In Chapter 2, we conduct a brief overview of the development of RHC and its statistical study in quantifying the causal effect. In Chapter 3, we introduce the notions of causal inference and matching as means to quantify causal relationship. In Chapter 4, we revisit the dataset on RHC and perform the matching. Outcome analysis based on post-matching data is conducted. Sensitivity analysis is used to check the possible influence of unobserved confounders. A simulation is also implemented to check the credibility of our empirical results. Chapter 5 presents the final comments.

Chapter 2
Literature Review

A tremendous amount of the literature has studied RHC from a variety of aspects, with two main emphases. One is on its methodology development and the other is on its clinical effects. For the purpose of my study, attention is limited to the area of its clinical effects which is usually quantitatively investigated using various statistical methods. Even when only concentrating on the study of RHC via statistical analysis, the proposed research questions differ from one another. The general principle that carries over among these studies is that evaluation of RHC procedure must be weighted against the accompanying risks, as pointed out by Hemmerling et al. (2013).

One branch of research studies the complications of RHC in a specific group of patients. The motivation behind such studies are bi-fold. For one reason, RHC serves usually as a routine procedure to patients with certain diseases including coronary angiographic and other noncardiac surgery. For the other reason, practitioners still have not reached a definite decision on whether to conduct routine RHC procedure among these patients due to data availability. Among them, Shanes et al. (1987) limited their attention to determine the effect of RHC procedure for patients with coronary arteriography using data collected from 1984 to 1986. Several pressure measures were taken to evaluate the performance of the routine RHC procedure. Their conclusion was that the yield obtained from routine RHC is not sufficient to warrant its use.

Polanczyk et al. (2001) evaluated the relationship between use of "perioperative RHC and postoperative cardiac complication rates in patients undergoing major noncardiac surgery" using data from 1989 to 1994 collected at Tertiary care teaching hospital in the United States. A multivariate logistic regression model

was employed with the main covariates of the model being the preoperative clinical variables which was considered to impact the decision to use RHC. Matching based on propensity scores and type of surgical procedure was used as well to compare the patients from treatment and control groups. The major conclusions drawn in this study supported that perioperative RHC was again not associated with improved postoperative outcomes and was associated with prolonged hospitalization even after adjusting for the potentially confounding variables. Bergersen et al. (2011) conducted a study to investigate a more complicated scenario where congenital heart disease patients were the focus. A method was developed to allow equitable comparisons of adverse event rates and its significance was proved.

Hoeper and Lee (2006) investigated the risks associated with RHC procedures in patients with pulmonary hypertension based on a "multicenter 5-year retrospective and 6-month prospective evaluation". The overall RHC procedure-related mortality turned out to be 0.055%. The main conclusion was that among patients receiving RHC procedures, those with pulmonary hypertension have the relatively low morbidity and mortality rates. Elliott et al. (2011) also studied the group of patients with pulmonary hypertension. The research question they proposed was whether the differences in outcomes were associated with the frequency of RHCs. Kaplan Meier estimates in survival analysis was used to answer the question. They had different research findings, though. They claimed that there was no difference in survival among patient receiving different frequencies of RHC procedures, despite the existence of wide variability in the frequencies of RHC.

Another group of research studied the death rates brought directly by applying RHC. Robert and Richard (1968) finished the first such study in which the death rate was examined, with a reported rate of 0.16%. As is argued by Morton et al. (1993), two factors further complicate the death rate from the use RHC. One is "the use of heparin, low-osmolar contrast media and better equipment as well as experience" which would decrease the death rate. The other is "the tendency to select older and sicker patients, especially those in the acute stages of coronary events" may increase the death rate. To accommodate these two factors, Morton et al. (1993) conducted a study on the rates of death within 24 hours after the procedure or later to check if they are causally related to the procedure in a catheterization lab from 1977 to 1991. They found the death rate was approximately 0.1%.

Natarajan et al. (2002) quantified the waiting times, morbidity and mortality

of patients waiting for RHC. Two main findings were provided via a multivariate analysis. First, patients waiting for RHC can be prevented from experiencing major adverse events, such as death, myocardial infarction and congestive heart failure. Second, the study also supported that patients at higher risk should get earlier access by investigating the effect of increased capacity and prioritization schemes.

A third group of the research focused more broadly on the performance of RHC procedure. In particular, they generally studied the survival after the RHC procedure which may not necessarily be due to the treatment of RHC procedure. In particular, the treatment effect of RHC is considered as "an amalgam of the effect of catheterization itself plus the therapies that the information gleaned from catheterization make possible"(Bhattacharya et al., 2012).

Connors et al. (1996) examined the association between the use of RHC during the first 24 hours of care in the intensive care unit (ICU) and subsequent survival, length of stay, intensity of care, and cost of care. The data came from five US teaching hospitals between 1989 and 1994, which is the same data set we analyze. The main outcome measures were more broader than the literature we discussed so far. These include survival time, cost of care, intensity of care, and length of stay in the ICU and hospital. They constructed a propensity score for RHC using multivariable logistic regression. Case-matching and multivariable regression modeling techniques were used to estimate the association of RHC with specific outcomes after adjusting for treatment selection using the propensity score. Sensitivity analysis was used to estimate the potential effect of an unidentified or missing covariate on the results. They concluded that RHC was associated with increased mortality and increased utilization of resources after adjustment for treatment selection bias; however, they were unsure why this happened.

Using another matching technique of propensity score weighting scheme, Hirano and Imbens (2001) reinvestigated the same dataset focusing only on the subsequent survival as the outcome measure. In particular, they combined regression adjustment and weighting based on propensity score while allowing for flexibility in specifying the regression function and propensity score. One advantage of this approach is its improved efficiency. They found very similar conclusion compared to Connors et al. (1996) in terms of the effect of RHC.

A follow up research by Bhattacharya et al. (2012) extended the study of Con-

nors et al. (1996) by assuming the existence of differences between patients receiving RHC and those who do not. This assumption was made possible via implementing both the instrumental variable bounds(Manski bounds) and the bounds which exploit mild nonparametric, structural assumptions in addition to an instrumental variable(Shaikh bounds and Vytlacil bounds). The justified instrument for RHC was indicators of weekday admission. The conclusions drawn based on these two different bounds are different. They did not find RHC procedure was associated with mortality decrease or increase with Manski bound, while they found it helped reducing mortality in the case where Shaikh and Vytlacil bounds were used.

In our study, we use the same dataset focusing on the effect of RHC procedure on mortality after 7 days of hospitalization. We make a comparison on a variety of propensity score matching schemes. In the next section, we review the methods that have been adopted in propensity score matching.

Chapter 3
Model and Methodology

3.1 Average Causal Effect Model Set Up

Let Y_i be the response variable for subject i, which is observed in the data. Let $D_i = 1$ if subject i receives treatment D, and $D_i = 0$ if the subject does not. Let X_i be the exogenous covariates(instruments) of subject i. Let Y_{i1} be the outcome variable if subject i receives treatment and Y_{i0} be the outcome variable if the subject does not. For all the subjects with $D_i = 1$, the response is recorded as $Y_{i1} = Y_i$, but Y_{i0} is not observed in the data; for all the subjects with $D_i = 0$, $Y_{i0} = Y_i$, but Y_{i1} is not observed in the data. So $Y_i = Y_{i1}D_i + Y_{i0}(1 - D_i)$. For every i, we only observe either Y_{i1} or Y_{i0}.

Our question is whether D_i causes a change in the mean value of Y_i? It will not be sufficient to only show a significant difference between $\bar{Y_1} = \frac{\sum_i D_i Y_i}{\sum_i D_i}$ and $\bar{Y_0} = \frac{\sum_i (1 - D_i) Y_i}{\sum_i (1 - D_i)}$ since we need to rule out alternative explanations including systematic differences between groups other than D_i. This is identified as the fundamental problem of causal inference, i.e., one of the two potential outcomes is not observed (Holland, 1986). Specifically, the confounders which are common among all observational studies will make the sample mean comparison controversial.

Confounders $X_i = (x_{i1}, x_{i2}, \cdots, x_{ip})$ are defined as pretreatment variables that may jointly influenece D_i and Y_i. Another misunderstanding in estimating the causal effect is to regress Y_i on X_i and D_i and interpret the coefficient for D_i as the effect of D_i on Y_i. Mathematically, this coefficient is

$$E(Y_i|D_i = 1, x_i) - E(Y_i|D_i = 0, x_i).$$

This approach is comparing two different groups of individuals while causality is about changes in response when different treatments are applied to the same individual.

The idea of potential outcome initiated in Rubin (1974) can tackle this fundamental problem based on a counterfactual account of the causal relation. The significance of this idea can be thought of as involving a deliberate and accurate extension of the conceptual equipment of randomized experiments to the analysis of observational data. This line of research yields a precise definition of a (treatment) effect and allows for the development of mathematical conditions under which estimates can or cannot be interpreted as causal effects.

Specifically, by making certain assumptions, however, it becomes possible to estimate the average causal effect for the population and for the treated group. These assumptions are:

(1) "Stable unit treatment value assumption" (known as "SUTVA" in Rubin (1980)) which guarantees that each individual has only one potential outcome under each exposure condition. This is necessary to ensure that the causal effect for each individual is stable. This stability assumption has two elements; the first one is that the exposure has the same effect on an individual regardless of how the individual came to be exposed and the second one is that the effect of the exposure on an individual is independent of the exposure of other individuals.

Violation of either aspect of SUTVA creates unstable estimates of the causal effect; "unstable" means that there is no unique potential outcome for each individual under each exposure condition. In general, the instability arises because there are multiple "versions of treatment". These different versions of treatment arise from ambiguities in the measurement or implementation of the treatment (violation of the first aspect of "SUTVA") or from effects of the treatments received by others (violation of the second aspect of "SUTVA")(Rosenbaum, 1980).

(2) "Unconfoundedness assumption", first coined by Rosenbaum and Rubin (1983), requires that the treatment is conditionally independent of the potential outcomes given the covariates. $D_i \perp (Y_{i0}, Y_{i1}) \mid X_i$. Since the data are uninformative about the distribution of the control outcome Y_{i0} for those with $D = 1$ and about the distribution of the treatment outcome Y_{i1} given receipt of the control treatment $D = 0$, the data cannot directly provide evidence on the validity of the unconfoundedness assumption of the control treatment. Nevertheless, there are

indirect ways of assessing the unconfoundedness assumption in some settings.

Specifically, two categories of approaches have been implemented to assess this assumption. The first category focuses on estimating the causal effect of the original treatment on a pseudo outcome, a variable that is known as a priori and is not affected by the treatment. In this approach the full set of pretreatment variables are divided into two parts, the pseudo outcome and the remaining covariates. The average causal effect of the treatment on the pseudo outcome can then be estimated. The unconfoundedness can be true if one does not have evidence that this treatment effect is significantly different from zero.

The second category focuses on estimating the causal effect of a pseudo treatment which is known not to have any effect on the true outcome. The implementability of this approach relies on the presence of multiple control groups initiated by (Rosenbaum, 1987). [1]

(3) "Overlap assumption" guarantees that the distributions of propensity scores or the distributions of set of observed covariates have common support between the treatment and control group. Mathematically, it is stated as $0 < P(D_i = 0) < 1$ and $0 < P(D_i = 1) < 1$. Such an assumption guarantees that each subject in the population could have been exposed to either treatment. Overlap is required for matching treatment and control cases. In the case where treatment observations having covariate values which do not overlap with control observations, it is unclear exactly what estimand is estimated since some treatment observations have been dropped along with some control observations, i.e., no comparable treatment and control matches are available. Notice that strongly ignorable assumption holds when both unconfoudedness and overlap assumptions are valid as described in (Rosenbaum and Rubin, 1983).

Under these assumptions, it is possible to make inference on average causal effect for the population (ACE) and for the treated subjects (TTE) where $ACE \equiv E(D_i) = E(Y_{i1}) - E(Y_{i0})$ and $TTE \equiv E(Y_{i1}|D_i = 1) - E(Y_{i0}|D_i = 1)$.

If all potential outcomes were seen, we would estimate the ACE by

$$\hat{ACE} = \frac{1}{n}\sum D_i = \frac{1}{n}\sum Y_{i1} - \frac{1}{n}\sum Y_{i0}$$

[1] Interested readers can refer to $http://www.bus.miami.edu/_assets/files/events/miami-unconf$ on the procedure of assessing the creditability of uncounfoundedness.

In an observational study, it is unlikely that D_i will be independent of Y_{i0} and Y_{i1}. It is crucial to have good pretreatment covariates to help understand and adjust for baseline differences between the groups.

An alternative to ACE for the whole population is the ACE for the treated (ACE_1), which measures how much the treatment helps or hurts the individual unit who actually received it. Therefore, it can be more relevant than ACE when discussing certain policy implications (Hirano and Imbens, 2001). Moreover, sometimes data do not contain enough information to estimate ACE reliably while they can be used to estimate ACE_1. Mathematically, it is defined as

$$ACE_1 = E(Y_{i1}|D_i = 1) - E(Y_{i0}|D_i = 1).$$

To make unbiased inferences about the ACE_1, one only needs a weaker version of the overlap assumption, $P(D = 1|x) < 1$ for all x, in addition to the assumption of no unmeasured confounders.

In practice, the choice of the causal quantity of interest depends on the research question, in particular, whether the interest is in estimating the treatment effect for the overall target population (i.e., treated and untreated units together) or the treatment effect for the treated units only. One of the problems of estimates of ACE and ACE_1 is that it may differ depending on the model you specify. (King and Zeng, 2006) "define model dependence at point x as the difference, or distance, between the predicted outcome values from any two plausible alternative models $\cdots$ By plausible alternative models, we mean models that fit the data reasonably well and, in particular, they fit about equally well around either the center of the data(such as a multivariate mean or median) or the center of a sufficiently large cluster of data nearest the counterfactual x of interest." Matching is one nonparametric preprocessing to ameliorate model dependence (Ho et al., 2007). They argued that in the preprocessed data set, the treatment variable is closer to being independent of the background covariates, which makes any subsequent parametric adjustment either irrelevant or less important.

In the following section we discuss a very specific class of statistical methods, called matching, for removing selection bias. These methods try to match treatment and control units on observed baseline characteristics X in order to create comparable groups just as randomization would have done. If treatment selection

is ignorable (i.e., all confounding covariates are measured) and if treatment and control groups are perfectly matched on observed covariates X, then potential outcomes are independent of treatment selection. Matching estimators are of course not alone in their aim of estimating causal treatment effects. [2]

3.2 Matching

Suppose initially M units are available, numbered as $i = 1, \cdots, M$. X_i is the covariate for the i-th unit and the treatment assignment for this unit is D_i. Suppose that unit i is assigned to treatment with probability $\Pi_i = P(D_i = 1 \mid X_i, Y_{i0}, Y_{i1})$ and to control with probability $1 - \Pi_i = P(D_i = 0 \mid X_i, Y_{i0}, Y_{i1})$ and also assume that assignments of distinct units are independent of each other. When the treatment mechanism is unconfounded, the propensities depend only on $X_i, \Pi_i = P(D_i = 1 \mid X_i)$. Rosenbaum and Rubin (1983) defined these probabilities as propensity scores, serving as a fundamental tool for the construction of matched sets.

Propensity scores have two properties: 1) balancing property, i.e., given the propensity score, the pre-treatment variables are balanced between treatment and control groups; 2) suppose that assignment to treatment is unconfounded given the pre-treatment variables X, then assignment to treatment is unconfounded given the propensity scores. The balancing property of the propensity score ensures that: 1) observations with the same propensity score have the same distribution of observable covariates and is independent of the treatment status; and 2) for a given propensity score, assignment to treatment is random and therefore treatment and control units are observationally identical on average.

To make inference on Π_i, a direct approach to construct matched sets is to find one or more control units that have similar covariates X for each treated unit, which is also the rational behind stratification. Subjects are usually grouped into strata on the basis of the covariate. From the M units, select $N \leq M$ subjects and group them into S non-overlapping strata with n_s units in stratum s. Only information on X and a table of random numbers is used for the selection of N

[2] There are also other methods like standard regression, analysis of covariance models, structural equation models [see for example, (Kaplan, 2005), (Pearl, 2009), (Steyer, 2005)] or Heckman selection models (Heckman, 1979) also try to identify causal effects. Since these methods have a different focus on causality and typically rely on stronger assumptions, particularly functional form and distribution assumptions, they are not discussed in my study.

subjects. The i-th unit in stratum s has treatment assignment D_{si} and covariate X_{si}. A stratification formed in this way is called a stratification on X. Exact stratification on X is practical only when X is of low dimension and its coordinates are discrete; otherwise, it will be difficult to locate many units with the same X.

Matching is a special form of stratification in which there are constraints on the number of observed treated and control units in each stratum. A matching on X is then a matched sample formed by placing some restrictions on S, m and $n = (n_1, \cdots, n_s)$ and picking a stratification that satisfies these restrictions based exclusively on the patterns of X. The rationale of matching is to construct matched sets or strata within which the observed covariates are balanced between the treated and control group and then compare outcomes between the treated and control group within matched strata. One can then obtain unbiased estimates of treatment effects. In particular, an exact matching on X is a matching in which X is the same for all units in each matched set(strata). This attempt to match each treated subject to a control with almost the same observed covariates will quickly proven to be impractical when there are many covariates; it is not possible to implement it even asymptotically if X contains continuous covariates. Other matching methods need to be used in such circumstance.

Two common approaches are used in the literature: propensity score matching and multivariate matching based on Mahalanobis distance. In both approaches, the targeted covariate balance tolerates close but inexact matches, instead of perfect match (Sekhon, 2010). They are designed to produce matched pairs or sets that balance observed covariates, so that the distributions of observed covariates are similar between treated and control groups based on a defined "distance". Both of these two matching methods are built on specific notions of distance between observations of pretreatment covariates. Matched pairs or sets are formed in such a way that subjects in the same matched set have the same probability of receiving the treatment. Within a stratum or matched set, subjects may have different values of D, but having the same propensity score.

As a summary, instead of stratifying or matching exactly on X, these two approaches imagine forming matched sets in which units in the same matched set have the same chance of receiving the treatment Π_i. Such method will yield a conditional distribution of treatment assignments D given X that is the same as a uniform randomized experiment. Theoretical arguments show that certain

distances are highly effective in producing comparable groups even when exact matches are not used (Rosenbaum and Rubin, 1985).

Propensity score matching usually involves matching each treated unit to the nearest control unit on the unidimensional metric of the propensity score vector. Because the propensity score is generally unknown, it must be estimated. When estimating the propensity score, two choices have to be made. The first one is about the model to be used for the estimation, and the second one is about the variables to be included in this model. Let's start by discussing the model choice. In principle, discrete choice model of any kinds can be theoretically feasible. One common way to estimate propensity scores is by fitting a logistic regression, $\log(\frac{\Pi_i}{1-\Pi_i}) = X_i'\gamma$ [3]. Preferences for logit model to other linear probability models stem from the well-known shortcomings of the latter including the unlikeliness of the functional form when the response variable is highly skewed and predictions that are outside the $[0, 1]$ bounds of probabilities (Smith, 1997). The case where there are more than two choices of treatments, the modeling choice is more complicated which is beyond the scope of my study. [4]

We proceed with the choices of covariates. For X, we want to consider any covariates that might play a role in the selection of treatments and/or might be associated with the outcome. Omitting important variables can seriously increase bias in resulting estimates[e.g., see Heckman et al. (1997) for a discussion]. Typically, we should consider any covariates that might play a role in the selection of treatments and/or might be associated with the outcome. However, sometimes questions may arise if it is better to include too many rather than too few variables. As discussed by Alex Bryson and Purdon (2002), the following two reasons are discussed about why over-parameterized models should be avoided. First, including extraneous variables in the model may exacerbate the support problem. Second, although the inclusion of non-significant variables will not yield biased or inconsistent estimates, it can increase their variances.

We now turn to multivariate matching. The most popular method of multivariate matching is based on Mahalanobis distance [for example, see (Cochran and Rubin, 1973) and (Rubin, 1980)]. The Mahalanobis distance between any two

[3]However, researchers found that slight mis-specification of the propensity score model can result in substantial bias of estimated treatment effect. For one such example, refer to Kang and Schafer (2007).

[4]Interesting readers may refer to Imbens (2000) and Lechner (2001).

column vectors is defined as:

$$d(x_i, x_j) = [(x_i - x_j)' S^{-1} (x_i - x_j)]^{1/2}$$

where S is the sample covariance matrix of x. If x consists of more than one continuous variable, multivariate matching estimates contain a bias term that does not asymptotically go to zero at rate $\sqrt{n}$.

Mahalanobis distance was originally developed to work with multivariate normal data, and for data of that type it works fine. With data that are not normal, the Mahalanobis distance can exhibit some rather odd behavior. If there are extreme outliers and/or a long-tailed distribution in one covariate, its standard deviation will be inflated, and the Mahalanobis distance will give little weight to that covariate in matching. The concern is especially obvious in the case of binary covariates: "the variance is largest for events that occur about half the time, and it is smallest for events with probabilities near zero and one. In consequence, the Mahalanobis distance gives greater weight to binary variables with probabilities near zero or one than to binary variables with probabilities closer to one half"(Rosenbaum, 2002).

As pointed out by Rubin (2001), among other researchers, propensity score matching and multivariate matching can be combined in a number of ways. The combination of these two together were shown to reduce more effectively the covariate balance and mean square error of the causal effect estimates (Rosenbaum and Rubin, 1985). The rational behind this improvement is that the propensity score is a balancing score only asymptotically. In finite samples, some covariate imbalances will remain by one matching method, which can be adjusted by another matching method.

Propensity score matching and multivariate matching (if the propensity score is estimated by logistic regression) are equal percent bias reducing (EPBR) if all of the covariates used have ellipsoidal distributions [5] [e.g., (Rubin, 1976) and (Rubin and Thomas, 1992)]. A matching method is EPBR for X when the percent reduction in the biases of each matching variable is the same. In the study by Iacus and Porro (2011), EPBR was critically reviewed, mainly focusing on the two strict assumptions on the distribution of the covariates that are almost never

[5] Distributions such as the normal or t or if the covariates are discriminant mixtures of proportional ellipsoidally symmetric distributions as proposed by Rubin and Stuart (2006).

known to hold in any observational studies. Instead, they proposed another class of matching methods called "Monotonic Imbalance Bounding Class" by generalizing and modifying the definition of EPBR. Coarsened exact matching is one example from this class.

Several algorithms which assign controls and treated units using the distance have been developed to conduct multivariate matching, including among others nearest neighbor matching, optimal matching, full matching, weighting and genetic matching. The most common implementation of both propensity score matching and multivariate matching is to apply one-to-one nearest neighbor greedy matching without replacement. This procedure matches each treated unit in some arbitrary sequence to the nearest control unit, using a chosen distance metric (Austin, 2009). In nearest neighbor matching(NN), the absolute difference between the estimated propensity scores for the control and treatment groups is minimized. The control and treatment subjects are randomly ordered. Then the first treated subject is selected along with a control subject with a propensity score closest in value to it. One critique of NN is the possible wastefulness of discarding a large number of unmatched observations. However, as is discussed in Stuart (2010), the power does not reduced dramatically. The excess cases do not have any comparable subjects in the other group, which makes the excess cases not crucial for drawing causal inference. Another critique is that NN does not generally minimize the total distance within pairs, and theoretical arguments and simple examples show that the NN's distance can be much larger than the minimum attainable (Rosenbaum, 1989).

Using network flow theory, optimal matching minimizes the average absolute distances across all matched pairs, not merely across the treatment and control groups [e.g., (Rosenbaum, 1989), (Bertsekas, 1991) and (Ho et al., 2011)]. Optimal matching, when combined with propensity score caliper, can increase the speed of the algorithms that find the optimal match. So in this sense, the calipers are desirable. The advantage of optimal matching over NN is that NN comes with no guarantee that it will find a tolerable match when it exists (Rosenbaum, 1989).

Both NN and optimal matching do not necessarily use all the units in the data. By contrast, full matching make full use of each individual unit by forming a series of matched sets and it is more flexible than NN matching. Specifically, each matched set has either 1 treated unit and multiple controlled units or 1 control unit

and multiple treated units (Stuart and Green, 2008). The flexibility comes from the fact that treated units who have many comparison units who are similar will be placed in the same set, whereas treated units with few similar comparison units will be placed in the set with relatively fewer comparison units. Hansen (2004) illustrates full matching in great details and points out that it is optimal in the sense that it minimizes propensity score distances and also succeed in removing bias due to observed covariates.

Genetic matching, developed in Alexis Diamond (2012) and Sekhon (2011), is a combination of propensity score matching and multivariate matching which automatically finds the set of matches that minimizes the discrepancy of the distributions of potential confounders between the treated and control groups. That is, covariate balance is maximized. It determines the weight of each covariate based on an evolutionary search algorithm. It is considered to be a limiting case of propensity score matching which uses nonparametric technique and does not depend on knowing or estimating the propensity scores. However, the performance is better if incorporated with a propensity score. The simulation study by Diamond and Sekhon (2013) argues that "the algorithm improves covariate balance, and that it may reduce bias if the selection on observables assumption holds".

We will use these matching methods to investigate our empirical example shortly and compare their performances.

Chapter 4
Empirical Investigation

4.1 Data Description

We investigate the data collected by the Study to Understand Prognoses and Preferences for Outcomes and Risks of Treatments (SUPPORT) which comprised a two-year prospective observational study (Phase I) followed by a two-year controlled clinical trial (Phase II). See Figure 4.1 taken from Bergner et al. (1995) for a detailed organization of the experiment.

Phase I included a prospective observational study that described the process

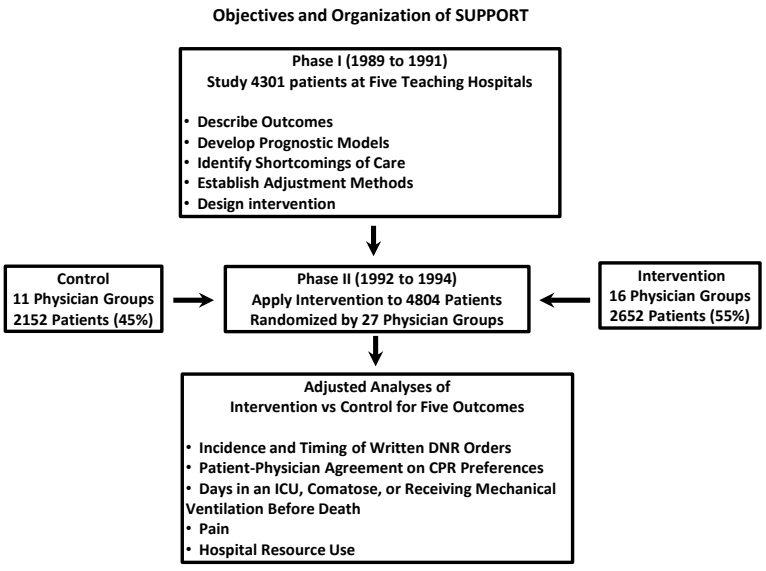

Figure 4.1. Experiment Organization

of decision making and patient outcomes. In Phase I, data from patients during 1989-1991 were collected with information on the care, treatment preferences, and patterns of decision-making among critically ill patients. It also functioned "as a preliminary step for devising an intervention strategy for improving critically-ill patients' care and for the construction of statistical models for predicting patient prognosis and functional status"[1].

Phase II was a cluster randomized controlled clinical trial to test the effect of the intervention. Enrollment, data collection, and interviewing were virtually identical during the two phases (Donald J. Murphy and Lynn, 1990). Here patients during 1992-1994 were exposed to an intervention which provided physicians with accurate predictive information on the future functional ability, survival probability within six months, and patients' preferences for end-of-life care. As part of the intervention, a skilled nurse was distributed to collect patient preferences information, provide prognoses, enhance understanding, enable palliative care, and facilitate advance planning. The intention of the intervention was to enhance efficient communication, help making earlier decisions to have orders against resuscitation, decrease undesirable time that patients went through (e.g., in the Intensive Care Unit, on a ventilator, and in a coma), help physician better understand patients' caring preferences, decrease patient pain, and efficiently utilize hospital resources. Phase II also collected information regarding the implementation of the intervention, such as patient-specific logs maintained by nurses assigned to patients as part of the intervention.

The patients were hospitalized with one of nine prespecified serious illnesses in one of five US teaching hospitals between 1992 and 1994 and were all followed up for a 6-month period. Qualified patients were in the advanced stages with the following nine disease categories: acute respiratory failure (ARF), chronic obstructive pulmonary disease (COPD), congestive heart failure (CHF), liver disease, coma, colon cancer, lung cancer, multiple organ system failure with malignancy (MOSF), and multiple organ system failure with sepsis. The disqualified patients include those: younger than 18 years, discharged or died within 48 hours of qualifying for the study, admitted with a scheduled discharge within 72 hours, who did not speak English, admitted to the psychiatric ward, who had acquired immunodeficiency syndrome, or were pregnant or sustained an acute burn, head, or

[1]The quoted description here comes from http://www.icpsr.umich.edu/icpsrweb/ICPSR/studies/2957.

other trauma[2] (Knaus et al., 1995). The newly admitted patients to hospital and ICU were identified and reviewed by the trained nurses who are familiar with SUPPORT. Potential adverse events, including 6-month mortality for intervention patients and changes in patient satisfaction with medical care were monitored and recorded by another independent committee.

The five hospitals include: Beth Isreal Hospital, Boston, Mass; Duke University Medical Center, Durham, NC; Metro-Health Medical Center, Cleveland, Ohio; StJoseph's Hospital, Marshfield, Wisconsin; and University of California Medical Center, Los Angeles. After admission into the hospital, patients, with the discretion of the physicians, chose whether or not to receive RHC, which is a direct measurement of cardiac function (Bergner et al., 1995).

To make the design as close to randomization as possible, the following efforts were made both to the design and analysis of the experiment. First, the assignment of patients to intervention and control groups (usual care) were based on the specialties of their attending physicians. These physicians were partitioned into five groups including internal medicine, pulmonology/medical ICU, oncology, surgery, and cardiology. Second, a cluster randomization scheme was implemented to assign the intervention randomly to 27 physician group-site combinations, while maintaining 50% to 60% of patients to be assigned as intervention status and that at least one intervention and one control physician specialty group to be at every five hospitals. Third, analysis were based on allocation to intervention (i.e, intention to treat), irrespective of whether a given patient received the intervention. Meanwhile, investigators had no information on the phase II results during the data collection process.

Using propensity scores, two studies by Connors et al. (1996) and Hirano and Imbens (2001) have analyzed whether RHC leads to better clinical outcomes respectively using different matching techniques to attenuate the possible effect of treatment selection bias. In particular, Connors et al. (1996) use "pairwise/case matching" where each individual is matched at most once, finding that the use of RHC shortens the survival length. Hirano and Imbens (2001) instead adopt weighting technique together with regression adjustment and draw the same conclusion. Our study is an extension to these two approaches in the sense that we conduct

[2]Such patients were included in the study unless they later developed acute respiratory failure or multiple organ system failure.

a more complete matching techniques and make comparisons. In addition, a simulation study is implemented for two purposes; one is to verify the performance the empirical results from the matching; the other is that we allow for different scenarios on how covariates impact the treatment selection, i.e., we distinguish whether the relationship between the treatment indicator and the covariates is linear and/or additive.

The dataset has 5735 individuals. The treatment variable is whether or not receiving RHC in the first 24 hours after entering the study. 2184 patients received the RHC treatment, leaving 3551 units not receiving the treatment. The binary outcome variable (dth30) is whether or not survive after 30 days of admission into hospital. Table 4.2 presents the summary statistics of the covariates used in our analysis where the first three columns characterize all the categorical variables and the remaining three columns list all the continuous covariates under study. Variables description is provided in Table 4.1.

From Table 4.2, we find that among the patients receiving RHC procedure, 830(38%) of them failed to survive after 30 days of admission into hospital. For patients not receiving RHC procedure, 1088(31%) of them failed to survive. For the whole population, the rate of patients who fail to survive within 30 days of admission is 34%. cat1 stands for the primary disease category the patients revealed during admission. For all patients Among patients, 42% of them were reported to have acute respiratory failure (ARF) which amounted to the largest portion of the patients population. The second largest group is patients with multiple organ system failure with sepsis (about 20%). Lung cancer and colon cancer were among the lowest two disease categories (less than 1%). For patients receiving RHC procedure, the largest two disease categories were ARF (42%) and multiple organ system failure with sepsis (32%). The smallest two categories were lung and colon cancer, almost negligible (less than 1%). For patients not receiving RHC procedure, the largest two disease categories were ARF (46%) and multiple organ system failure with sepsis (15%). The smallest two disease categories were the lung and colon cancer, again negligible. In the summary statistics table, we chose to report the smallest four disease categories as "others" since their own shares were rather small.

Let's now focus our attention on the structure of covariates. "ca" indicates patients' cancer status which has three categories: with cancer, no cancer, and

metastatic. Among all patients, the relative ratios of each category were 17%, 76%, 7%. Among patients receiving the RHC procedure, the relative ratios of each category were 15%, 79%, 6%. Among patients not receiving the RHC procedure, the relative ratios of each category were 20%, 75%, 5%. The reason to include this covariate is to control for the possible causes of death by cancer itself. Covariate "dnr1" is a binary indicator for the "do not-resuscitate" (DNR) status on day 1. Among all patients, 89% of the patients were not given a DNR order in day 1. This ratio was 93% and 86% for patients receiving RHC procedure and not receiving RHC procedure respectively.

"ninsclas" stands for patients' medical insurance type. Six types were reported: Medicaid, Medicare, Medicaid and Medicare, No insurance, Private, Private and Medicare. Among all patients, private insurance accounts for the largest portion of 30%, followed by Medicare 25%, Medicare and Private 22%. The portion of No insurance, Medicare and Medicaid were negligible (less than 10%). Among patients receiving RHC procedure, private and Medicare were again the largest two insurance options, with ratios of 34% and 23% respectively. The third largest insurance option is Medicare and Private with a portion of 22%. The portions of the other three options are similar, all around 9%. Among patients not receiving RHC procedure, private and Medicare were the most frequently used two insurance options, both around 27%. It is followed by the patients using the combination of these two insurances(21%). The least three insurance options for them is Medicaid (13%), Medicare and Medicaid (7%), no insurance (5%). Such information is consistent with our intuition about the insurance markets.

"resp" is a binary indicator for respiratory diagnosis. Among all patients, 64% of them did not get the diagnosis. The rate was 71% and 58% respectively for patients receiving and not receiving the RHC treatment. The difference of the ratios between these two groups of patients was remarkably different. "card" is the binary indicator for cardiovascular diagnosis. "neuro" is the binary indicator for neurological diagnosis. "gastr" is the binary indicator for gastrointestinal diagnosis. "renal" is the binary indicator for renal diagnosis. "meta" is the binary indicator for metabolic diagnosis. "hema" is the binary indicator for hematologic diagnosis. "seps" is the binary indicator for sepsis diagnosis. "trauma" is the binary indicator for trauma diagnosis. "ortho" is the binary indicator for orthopedic diagnosis.

"income" is a categorical covariate with four levels: under 11k, 11k-25k, 25k-

50k, above 50k. Patients with income category of "under 11k" accounts for more than half of the whole population(56%) . The ratios for the population receiving RHC and not receiving RHC are 52% and 58% respectively. Patients with the income category " above 50K" account for the least portion for all these three patients samples (all less than 10%). Please see Figure 4.2 for the pie chart. Race is a categorical covariate with three levels: black, white and others. The pie chart for three types of population is shown in Figure 4.3. One can find that the race distributions do not differ too much from one another. Sex is a binary covariate with male and female patients. The pie chart for sex for these three types of population is shown in Figure 4.4. In all three types, male patients were slightly more than female patients.

All the covariates in the 4th column of Table 4.2 ending with number 1 indicate a measurement done in day 1 of patients admission. These include: heart beat rate (hrt1), respiratory diagnosis (resp1), temperature (temp1), PaO2/FIO2 ratio (pafi1), albumin (alb1), hematologic diagnosis (hema1), bilirubin (bili1), creatinine (crea1), sodium (sod1), potassium (pot1), partial pressure of Carbon Dioxide in Arterial Blood (paco21), PH (ph1), weight (wtkilo1), APACHE score (aps1), glasgow coma score (scoma1), Mean blood pressure (meanbp1), white blood cells (wblc1).

All the covariates in the 4th column of Table 4.2 ending with "hx" represent comorbidities illness. These include:

- cardiohx: acute MI, peripheral vascular disease, severe cardiovascular symptoms (NYHA-Class III), very severe cardiovascular symptoms (NYHA-Class IV).
- chfhx: congestive heart failure.
- dementhx: dementia, dtroke or cerebral infarct, parkinsons disease.
- psychhx: psychiatric history, active psychosis or severe depression.
- chrpulhx: chronic pulmonary disease, severe pulmonary disease, very severe pulmonary disease.
- renalhx: chronic renal disease, chronic hemodialysis or peritoneal dialysis.
- liverhx: cirrhosis, hepatic failure.

- gibledhx: upper GI bleeding.
- malighx: solid tumor, metastatic disease, chronic Leukemia/Myeloma, acute Leukemia, lymphoma.
- immunhx: immunosupperssion, organ transplant, HIV positivity, diabetes mellitus without end organ damage, diabetes mellitus with end organ damage, connective tissue disease.
- transhx: transfer (> 24 Hours) from another hospital.
- amihx: definite myocardial infarction.

Table 4.1. Covariates definition

	Variable name	Variable Definition
Basics	Age	Age
	Sex	Sex
	Race	Race
	Edu	Years of education
	Income	Income
	Ninsclas	Medical insurance
	Cat1	Primary disease category
admission diagnosis:	Resp	Respiratory Diagnosis
	Card	Cardiovascular Diagnosis
	Neuro	Neurological Diagnosis
	Gastr	Gastrointestinal Diagnosis
	Renal	Renal Diagnosis
	Meta	Metabolic Diagnosis
	Hema	Hematologic Diagnosis
	Seps	Sepsis Diagnosis
	Trauma	Trauma Diagnosis
	Ortho	Orthopedic Diagnosis
	Das2d3pc	DASI (Duke Activity Status Index)
	Dnr1	DNR status on day1
	Ca	Cancer
	Surv2md1	Support model estimate of the prob. of surviving 2 months
	Aps1	APACHE score
	Scoma1	Glasgow Coma Score
	Wtkilo1	Weight
	Meanbp1	Mean blood pressure
	Pafi1	PaO2/FIO2 ratio
	Paco21	PaCo2
	Ph1	PH
	Sod1	Sodium
	Pot1	Potassium
	Crea1	Creatinine
	Bili1	Bilirubin
	Alb1	Albumin
comorbidities illness:	Cardiohx	Acute MI, Peripheral Vascular Disease, Severe Cardiovascular Symptoms (NYHA-Class III), Very Severe Cardiovascular Symptoms (NYHA-Class IV)
	Chfhx	Congestive Heart Failure
	Dementhx	Dementia, Stroke or Cerebral Infarct, Parkinsons Disease
	Psychhx	Psychiatric History, Active Psychosis or Severe Depression
	Chrpulhx	Chronic Pulmonary Disease, Severe Pulmonary Disease, Very Severe Pulmonary Disease
	Renalhx	Chronic Renal Disease, Chronic Hemodialysis or Peritoneal Dialysis
	Liverhx	Cirrhosis, Hepatic Failure
	Gibledhx	Upper GI Bleeding
	Malighx	Solid Tumor, Metastatic Disease, Chronic Leukemia/Myeloma, Acute Leukemia, Lymphoma
	Immunhx	Immunosupperssion, Organ Transplant, HIV Positivity, Diabetes Mellitus Without End Organ Damage, Diabetes Mellitus With End Organ Damage, Connective Tissue Disease
	Transhx	Transfer (> 24 Hours) from Another Hospital
	Amihx	Definite Myocardial Infarction

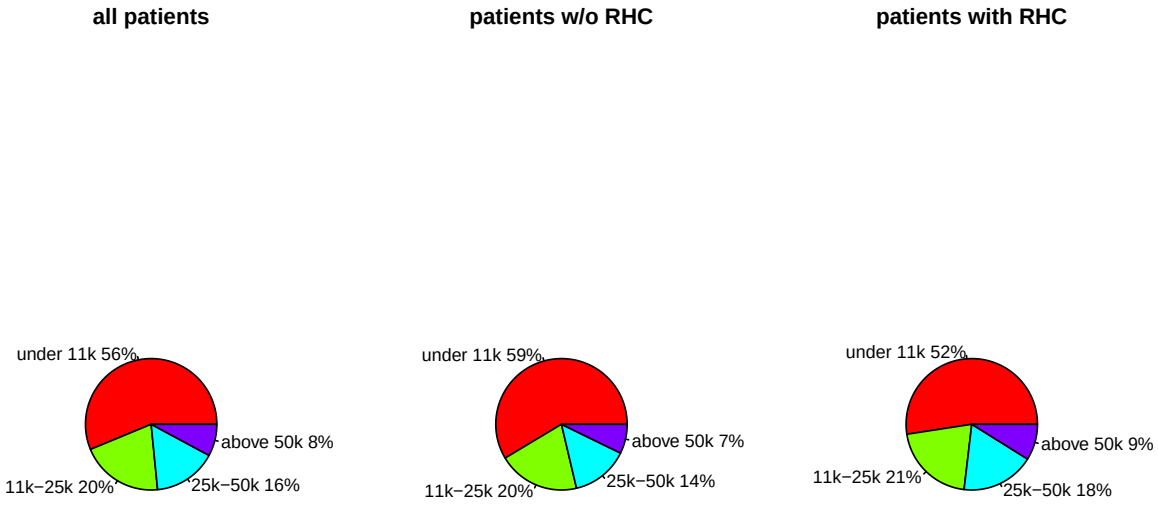

Figure 4.2. Pie chart of Income

The continuous variable age is the last covariate in the dataset. The mean value of age for all patients is 61.38, while it is 60.75 and 61.76 for patient receiving and not receiving RHC procedure. We also use the Fisher's F-test to check the equal variance of the two subpopulations (patient receiving and not receiving RHC procedure). The p-value turns out to be less than .05, a strong evidence against the null of equal variance assumption. The two sample t-test with unequal variances suggests a p value of .02. The age difference between these two subpopulation is therefore significant.

We now proceed to the implementation of matching methods.

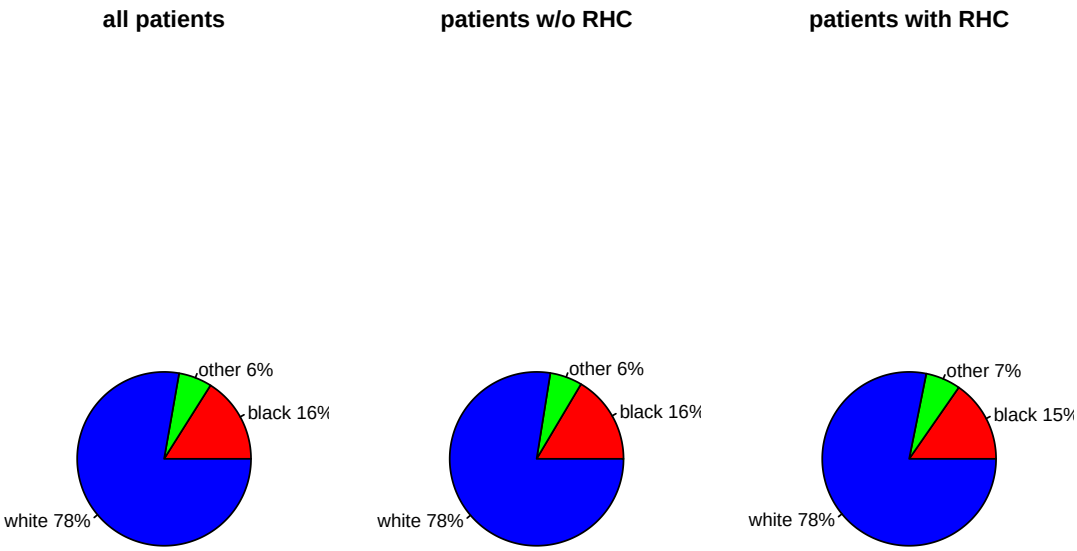

Figure 4.3. Pie chart of Race

4.2 Matching and Outcome Analysis

The following matching algorithms are implemented: nearest neighbor matching, full matching, optimal matching and genetic matching. All four algorithms can be directly implemented using the R package "MatchIt". For each algorithm, the same set of covariates is used both in estimating the propensity scores and the post-outcome analysis. For the post-outcome analysis, logistic regression of the response variable dth30 is implemented to see the additive increase in log-odds resulting from a one-unit increase in the covariates. Table 4.3 presents the results based on this regression. In each syntax, the default option is used. For example, in NN, logistic regression is used to compute the distance measure based on the

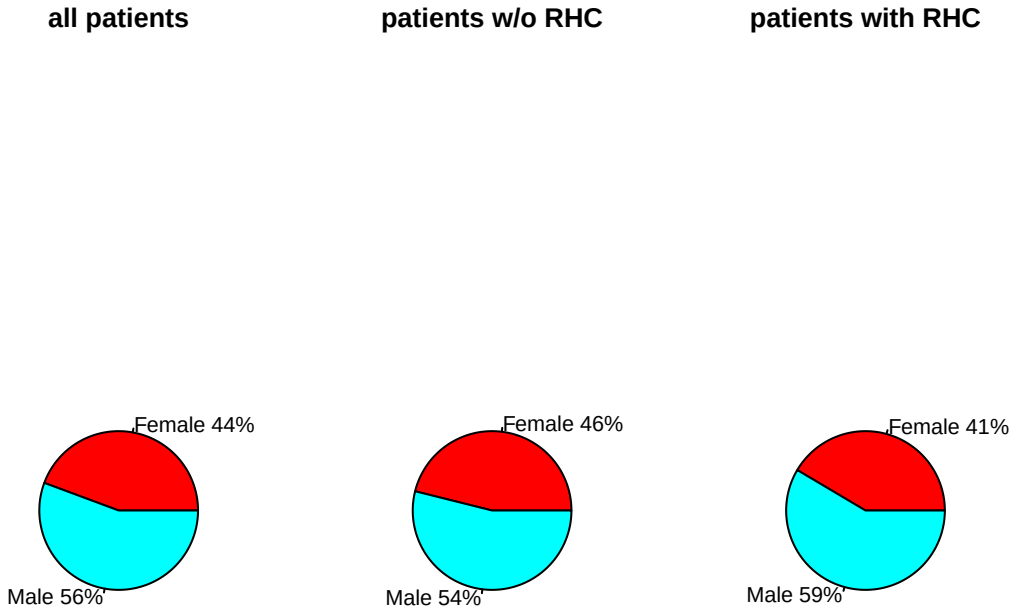

Figure 4.4. Pie chart of Sex

estimated propensity score. The number of best control matches for each individual in the treatment group is set to be 1, which is also the case in optimal matching. In full matching, no constraints are set on the ratio of treatment and control units within each matched set.

From Table 4.3, we find that RHC treatment within the first 24 hours of admission leads to lower possibility of surviving after 30 days of admission after accounting for baseline difference between patients receiving and not receiving the treatment. Although our analysis differs from the two studies by Connors et al.

(1996) and Hirano and Imbens (2001) in terms of the choices of either response variable or the potential confounders [3], the conclusion coincides with the findings in these two studies.

Using bootstrapping, the standard error of our estimate is also reported. The rationale behind bootstrapping is as follows. Our objective is to draw inference for the population based on samples. We know that sample varies from one to the other. Therefore, we want to quantify the magnitude of fluctuations of the samples. Bootstrapping is one method to quantify the magnitude. From Table 4.3, we find that the magnitude of standard errors is within a comparable range, with full matching yielding the smallest variance and genetic matching producing the largest variance.

4.3 Simulation

A Monte-Carlo simulation experiment is conducted mainly for two reasons. One is to check the performance and credibility of our estimators reported in table 4.3 by estimating the bias defined as the difference between the estimated and true average causal effect. The second purpose is to check whether our results are robust to flexible specifications of the relationships between the potential confounders and the treatment mechanism.

We follow a similar data generating process as Setoguchi et al. (2008) with a few modifications. D denotes whether receiving RHC treatment within the first 24 hours after admission, which is the binary exposure variable, with $p(D) = 0.5$. Ten covariates will be generated. Four of the covariates are correlated with both D and the outcome variable. Three of the covariates are associated with D only, while the remaining three covariates are associated with the outcome variable only. Among these covariates, six are constructed as discrete variables, while the rest are constructed as continuous variables.

Seven scenarios are considered based on the relationships among these covariates and D. In particular, these scenarios vary with the degree of linearity and/or

[3]Connors et al. (1996) examined length of stay, intensity of care and cost care as response variables in addition to dth30; while Hirano and Imbens (2001) used relatively different set of potential confounders since they implemented some variable selection techniques before the analysis.

Table 4.2. Characteristics of patients

Variable	Receiving RHC	Not receivng RHC	Variable	Receiving RHC	Not receivng RHC
dth30_yes	830	1088	hrt1	118.9(41.47)	112.9(40.94)
dth30_no	1354	2463	resp1	26.65(14.17)	28.98(13.94)
cat1_ARF	909	1581	temp1	37.59(1.82)	37.63(1.74)
cat1_MOSF w/Sepsis	700	527	pafi1	192.4(105)	240.63(116)
cat1_others	575	1443	alb1	2.97(0.92)	3.16(0.67)
cat1_COPD	58	399	hema1	30.5(7.41)	32.7(8.79)
cat1_Coma	95	341	bili1	2.7(5.32)	1.99(4.42)
cat1_CHF	209	247	crea1	2.47(2.05)	1.92(2.02)
cat1_MOSF w/Malignancy	158	241	sod1	136.3(7.60)	137(7.67)
ca_yes	334	638	pot1	4.05(1.01)	4.07(1.03)
ca_meta	123	261	paco21	36.79(10.97)	39.95(14.24)
dnr1_no	2029	3052	ph1	7.38(0.11)	7.39(0.10)
ninsclas_private	731	967	wtkilo1	72.36(27.72)	65.04(29.50)
ninsclas_medicare	511	947	cardiohx	0.2(0.4)	0.15(0.37)
ninsclas_others	452	1637	chfhx	0.16(0.40)	0.19(0.37)
resp_yes	632	1481	dementhx	0.11(0.25)	0.06(0.32)
card_yes	924	1007	psychhx	0.04(0.20)	0.08(0.27)
neuro_yes	118	575	chrpulhx	0.14(0.35)	0.21(0.41)
gastr_yes	420	522	renalhx	0.04(0.21)	0.04(0.20)
renal_yes	148	147	liverhx	0.06(0.24)	0.07(0.26)
meta_yes	93	172	gibledhx	0.02(0.15)	0.03(0.18)
hema_yes	115	239	malighx	0.2(0.40)	0.24(0.43)
seps_yes	516	515	immunhx	0.29(0.45)	0.25(0.43)
trauma_yes	34	18	transhx	0.14(0.35)	0.09(0.29)
income_$> 50K$	194	257	amihx	0.04(0.20)	0.02(0.16)
income_25k-50k	393	500	edu	11.86(3.15)	11.57(3.13)
income_11k-25k	452	713	surv2md1	0.56(0.19)	0.6(0.19)
ortho_yes	4	3	das2d3pc	20.7(5.03)	20.37(5.48)
sex_female	906	1637	aps1	60.74(20.27)	50.93(18.81)
race_black	335	585	scoma1	18.97(28.26)	22.25(31.37)
race_white	1707	2753	meanbp1	68.2(34.24)	84.87(38.87)
race_other	142	213	wblc1	16.27(12.54)	15.26(11.41)
			age	60.75(15.63)	61.76(17.28)

Table 4.3. Matching outcome analysis and bootstrapping standard error

	NN		Full		Optimal		Genetic	
	Coef	Se	Coef	Se	Coef	Se	Coef	Se
Swang1	-0.29	0.062	-0.33	0.054	-0.31	0.06	-0.41	0.09

additivity of modeled associations between the exposure D and the covariates. Following Setoguchi et al. (2008), these seven scenarios are:

1) Scenario A: a model with linearity and additivity;

2) Scenario B: a model with mild non-linearity, including two squared-terms covariates;

3) Scenario C: a model with moderate non-linearity, including three squared-term covariates;

4) Scenario D: a model with mild non-additivity, including four terms with products of two of the covariates;

5) Scenario E: a model with mild non-additivity and non-linearity;

6) Scenario F: a model with moderate non-additivity;

7) Scenario G: a model with moderate non-additivity and non-linearity.

Datasets were generated 1000 times for each of the seven simulation scenarios. The covariates are generated in two steps. Six of the covariates are first generated from normal distribution with 0 mean and unit variance. Correlations between some of the covariates are introduced with correlation coefficients varying from 0.2 to 0.9. These values present the magnitude of correlation coefficient before dichotomizing some of the covariates. The correlations will be attenuated after dichotomization.

In the simulation experiment, the effect of exposure is set to be constant with the coefficient of $D = 0.4$. The formulas of data generation functions and the coefficients of the formulas are the same as the ones used in Setoguchi et al. (2008).

Our focus of the simulation experiment is the bias and standard error of the estimators in the post-matched dataset, which reflect the true effect of exposure D on the response variable Y. Results are provided in Table 4.4, suggesting that the performances of the seven scenarios are similar to one another. The differences between true average causal effect and estimated average causal effect, namely the bias, are all small. The standard error are also acceptable. In nearest neighbor matching and full matching for Scenario F, the bias is larger than in any other

matching algorithms and scenarios. Nonadditivity specification may not be suitable for this kind of data analysis. These findings do provide solid evidence on the credibility and robustness of the performances of our estimators.

4.4 Sensitivity Analysis

Our results so far are based on the assumption that all possible confounders are identified and used to correct for the possible selection bias. Such assumption only holds in the randomized experiment setting, where each individual's true propensity score is known since an equal probability assignment mechanism is used to assign subjects to treatment or control. However, little thought suggests that each individual patient's estimated propensity scores are derived from observed covariates and there is always a concern about unmeasured confounders. Omitting relevant covariates may result in hidden bias that propensity scores cannot accommodate.

Sensitivity analysis is a specific statement quantifying the magnitude of the influence of hidden bias. Studies usually focus on quantifying the magnitude of hidden bias that needs to be present to alter the current conclusion on treatment effects. In our setting, we ask how strong the hidden bias has to be to alter our conclusion of negative effect of RHC on the survival status.

We follow the sensitivity analysis for the matching estimate developed by Rosenbaum (2002) and Rosenbaum (2005). The key parameter in a Rosenbaum style sensitivity analysis is the treatment odds ratio, known as Γ, a measure of the degree of departure from a study that is free of hidden bias, with each observed covariates being the same. Typically, a range of values for Γ is selected first. Then one does a sensitivity analysis on upper and lower bound and also the patterns of the p-values as Γ changes; each p-value is the consequence of a specific randomization test based on specific types of outcome variable, either discrete or continuous.

In our case, the binary outcome suggests that McNemar's test is appropriate. We report in Table 4.5 the upper bounds of the p-values for the McNemar's test for each of the four algorithms using, with some modifications [4], an R package "rbounds" developed by Keele (2010).

[4]In the case of full matching, the number of individual units differs from strata to strata. Hence, the computation of the McNemar's test statistic is different.

Table 4.5 demonstrates that for NN matching, when the odds of one patient receiving RHC treatment are 1.6 times higher because of different values on a missing covariate, although all observed covariates being identical, our inference on the response variable dth30 will be altered at 10% significance level. The thresholds for optimal matching, full matching and genetic matching are 1.2, 1.7 and 1.3 respectively. One general conclusion is that while it appears that RHC treatment has a negative effect on the survival within 30 days after admission, this finding is also sensitive to the hidden bias due to possibly unoberserved confounder. Among all the algorithms implemented, full matching is most robust to hidden bias.

As a further step, the above sensitivity analysis assumes "an extremely strong, near perfect, relationship" between the unobserved confounder and the outcome variable. In practice, such relationship may be implausible. Based on this consideration, we apply an amplification technique following the idea from Silber and Rosenbaum (2009) which allows both the relationship between the missing confounder and the treatment and also the relationship between the missing confounder and the outcome variable. The amplification will convert our above one dimensional analysis on Γ into a two or higher dimensions analysis.

Let Δ control the strength of the relationship between the missing confounder and dth30, while Λ control the relationship between the missing confounder and RHC. Following the decomposition formula derived in Silber and Rosenbaum (2009), $\Gamma = \dfrac{\Delta\Lambda + 1}{\Delta + \lambda}$, we generate Figure 4.5 depicting the corresponding amplification for each algorithm. To reach the same effect of Γ on our inference of causal effect, a set of all possible combinations of Λ and Δ is listed in Figure 4.5.

As demonstrated in Figuire 4.5, for the missing confounder to maintain the same strength as when missing confounder is only associated with the outcome variable, tradeoff is necessary between the two-direction associations; namely, if the association between the missing confounder and dth30 becomes stronger, the association between RHC and the missing variable has to be weaker. The degree of tradeoff varies among different schemes, with optimal matching requiring the largest tradeoff while full matching the smallest.

Table 4.4. Bias and standard errors of the simulation

	NN		Full		Optimal		Genetic	
	Bias	Std. dev	Bias	Std. dev	Bias	Std. dev	Bias	Std. dev
Scenario A	-0.005	0.17	-0.002	0.15	-0.005	0.17	-0.003	0.2
Scenario B	-0.003	0.18	0.001	0.16	0.002	0.18	0.0002	0.21
Scenario C	0.002	0.17	0.003	0.15	0.001	0.17	-0.002	0.19
Scenario D	0.0004	0.17	0.001	0.15	0.0003	0.17	0.005	0.19
Scenario E	0.0002	0.18	0.001	0.16	0.00004	0.18	0.005	0.21
Scenario F	0.163	0.001	0.153	0.001	0.001	0.17	0.002	0.19
Scenario G	0.0001	0.16	0.0003	0.15	0.001	0.16	-0.002	0.19

Table 4.5. Range of significance levels for hidden bias of various magnitudes

	P-value upper bounds			
Gamma	NN	Optimal	Full	Genetic
1	0	0	0.0312	0
1.1	0	0.00055	0.03943	0
1.2	0	0.02969	0.04828	0.00035
1.3	0.0001	0.26766	0.05769	0.01284
1.4	0.00043	0.70805	0.06754	0.12249
1.5	0.00928	0.9485	0.07778	0.43162
1.6	0.07431	0.99585	0.08825	0.7734
1.7	0.27626	0.99983	0.09895	0.94657
1.8	0.58068	1	0.10979	0.99227
1.9	0.83	1	0.12072	0.99927
2	0.95179	1	0.13169	0.99995

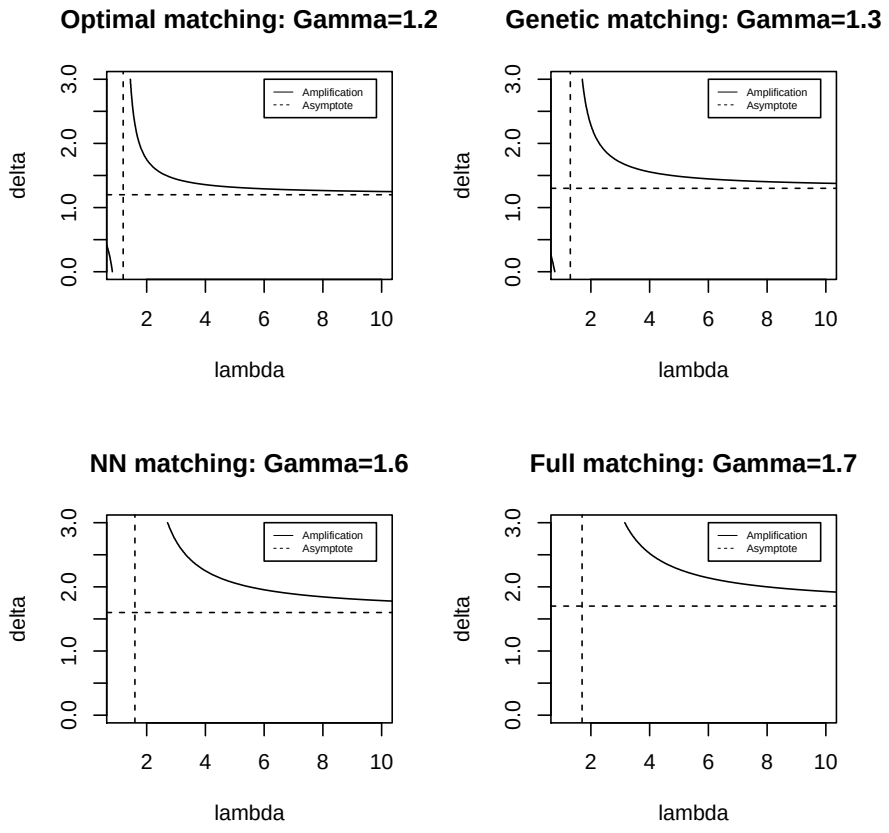

Figure 4.5. Amplification (Λ, Δ) of Γ

Chapter 5
Summary

In this study, we give a brief overview of the history of Right Heart Catheterization and a review of the methodological development of causal inference. The causal relationship between Right Heart Catheterization and adult patients' survival within 30 days after admission was reevaluated via four matching schemes. These four schemes draw the same conclusion that the risk of death within 30 days increases as a consequence of the treatment of Right Heart Catheterization within 24 hours of admission. Although different approaches are attempted, this finding coincides with the two previous studies mentioned earlier. Compared to these two studies, our perspective attempts to include more complete matching schemes so that we can have a comparison on the relative performance of each scheme.

Based on the conclusion on the causal effect of Right Heart Catheterization, we do believe these four schemes do not differ too much. Moreover, this negative causal effect is sensitive to the possible hidden bias identified by sensitivity analysis. Nevertheless, confronting possible hidden bias, the performances of these four schemes are different. Full matching is the most robust scheme to missing confounders, while the other three performs equally worse.

The Monte Carlo simulation experiment suggests that full matching performs better than other schemes in terms of the bias and also variance under the linear and additivity assumption. If the assumptions are relaxed, the bias and standard errors for these four matching schemes vary slightly. In particular, strong non-additivity may be a concern because it increases bias of the treatment effect estimator.

The following directions may be appropriate for future research. Within our current modeling framework, we may include more matching algorithms to make the

comparisons among them more complete. For example, we may consider subclassification, weighting, and coarsened exact matching. A more interesting extension to our current matching methods is to investigate the performances of scenarios where different matching methods are combined. For example, subclassification can be used in conjunction with nearest neighbor matching. On the other hand, since we do have the information regarding patients' actual survival time within the study period, we may switch our framework to survival analysis which contains more information on the survival length.

Appendix
R Code

```
###Read in data
rhc<-read.csv("rhc.csv",header=TRUE)
names(rhc)
dim(rhc)
###check missing data
newrhc<-na.omit(rhc)
dim(newrhc)
sum(is.na(dthdte)) ### number of missing data pts
sum(is.na(dschdte))
sum(is.na(urin1))
sum(is.na(adld3p))
sum(is.na(cat2))
#############################################################################
###################read in the new data#####################################
###################after creating two var and delete some var#########
#############################################################################
newdata<-read.csv("data.csv",header=TRUE)
dim(newdata)
attach(newdata)
#########################################################
##################Characteristics of the patients#####
#########################################################
newdata1 <- newdata[ which(newdata$swang1=='1'), ]
table(newdata1$sex)
table(newdata1$cat1)
table(newdata1$ca)
newdata2 <- newdata[ which(newdata$swang1=='0'), ]
table(newdata2$sex)
```

```
table(newdata2$cat1)
table(newdata2$ca)
#############################################################################
#####################matching#################################################
#############################################################################
#######1)nearest neighbour#################################################
require(MatchIt)
m.out1<-matchit(swang1~cat1+ca+cardiohx + chfhx + dementhx
+ psychhx+ chrpulhx +renalhx +liverhx + gibledhx + malighx
+ immunhx+ transhx+amihx+age+sex+edu+surv2md1+das2d3pc+aps1
+scoma1+meanbp1+pafi1+alb1+hema1+bili1+crea1
+sod1+pot1+paco21+ph1+wtkilo1+dnr1+ninsclas+resp
+card+neuro+gastr+renal+meta+hema+seps+trauma+ortho+race
+income, method="nearest", data=newdata, ratio=1)
m.out1
final_data1<-match.data(m.out1)

# ordinary mean comparison:

fit1<- glm(as.factor(dth30) ~ swang1, family="binomial",data=final_data1)

print(summary(fit1))

#######2).Full Matching#####################################
require(optmatch)
require(MatchIt)
m.out2<-matchit(swang1~cat1+ca+cardiohx + chfhx + dementhx +
psychhx+ chrpulhx +renalhx +liverhx + gibledhx +malighx +
immunhx+ transhx+amihx+age+sex+edu+surv2md1+das2d3pc+aps1+
scoma1+meanbp1+pafi1+alb1+hema1+bili1+crea1
+sod1+pot1+paco21+ph1+wtkilo1+dnr1+ninsclas+resp+card+neuro
+gastr+renal+meta+hema+seps+trauma+ortho+race+income,
method="full", data=newdata)
m.out2
final_data2<-match.data(m.out2)

#ordinary mean comparison:

fit2 <- glm(as.factor(dth30) ~ swang1, family="binomial",
data=final_data2)
```

```
print(summary(fit2))

##########3) optimal matching#########################################

m.out3<-matchit(swang1~cat1+ca+cardiohx + chfhx + dementhx + psychhx
+ chrpulhx +renalhx +liverhx + gibledhx +malighx + immunhx+ transhx
+amihx+age+sex+edu+surv2md1+das2d3pc+aps1+scoma1+meanbp1+pafi1+alb1
+hema1+bili1+crea1+sod1+pot1+paco21+ph1+wtkilo1+dnr1+ninsclas+resp
+card+neuro+gastr+renal+meta+hema+seps+trauma+ortho+race+income,
method="optimal", data=newdata)
m.out3
final_data3<-match.data(m.out3)

# ordinary mean comparison:

fit3 <- glm(as.factor(dth30) ~ swang1, family="binomial",
data=final_data3)

print(summary(fit3))


###4)Genetic matching###################

m.out4<-matchit(swang1~cat1+ca+cardiohx + chfhx + dementhx + psychhx
+ chrpulhx +renalhx +liverhx + gibledhx +malighx + immunhx+ transhx
+amihx+age+sex+edu+surv2md1+das2d3pc+aps1+scoma1+meanbp1+pafi1+
alb1+hema1+bili1+crea1+sod1+pot1+paco21+ph1+wtkilo1+dnr1+ninsclas+
resp+card+neuro+gastr+renal+meta+hema+seps+trauma+ortho+race+income,
method="genetic", data=newdata)
m.out4
final_data4<-match.data(m.out4)
# ordinary mean comparison:

fit4 <- glm(as.factor(dth30) ~ swang1, family="binomial",data=final_data4)

print(summary(fit4))

#####################################
####bootstrap to calculate variance##
#####################################
repe=1000
```

```
est1=rep(NA, repe)
standerror1=rep(NA, repe)

est2=rep(NA, repe)
standerror2=rep(NA, repe)

est3=rep(NA, repe)
standerror3=rep(NA, repe)

est4=rep(NA, repe)
standerror4=rep(NA, repe)


for (j in 1:repe){

bo=sample(1:dim(newdata)[1], replace=TRUE)
bootsample=newdata[bo,]

##NN
require(MatchIt)
m.out1<-matchit(swang1~cat1+ca+cardiohx + chfhx + dementhx + psychhx
+ chrpulhx +renalhx +liverhx + gibledhx +malighx + immunhx+ transhx
+amihx+age+sex+edu+surv2md1+das2d3pc+aps1+scoma1+meanbp1+pafi1+alb1
+hema1+bili1+crea1+sod1+pot1+paco21+ph1+wtkilo1+dnr1+ninsclas+resp
+card+neuro+gastr+renal+meta+hema+seps+trauma+ortho+race+income,
method="nearest", data=bootsample, ratio=1)

final_data1<-match.data(m.out1)

model1=glm(as.factor(dth30)~swang1, family="binomial", data=final_data1)

est1[j]=summary(model1)$coef[2,1]
standerror1[j]=summary(model1)$coef[2,2]

##FULL
require(optmatch)
require(MatchIt)
m.out2<-matchit(swang1~cat1+ca+cardiohx + chfhx + dementhx + psychhx
+ chrpulhx +renalhx +liverhx + gibledhx +malighx + immunhx+ transhx
+amihx+age+sex+edu+surv2md1+das2d3pc+aps1+scoma1+meanbp1+pafi1+alb1
```

```
+hema1+bili1+crea1+sod1+pot1+paco21+ph1+wtkilo1+dnr1+ninsclas+resp+
card+neuro+gastr+renal+meta+hema+seps+trauma+ortho+race+income,
method="full", data=bootsample)
m.out2

final_data2<-match.data(m.out2)

model2 <-glm(as.factor(dth30) ~ swang1, family="binomial",
data=final_data2)

est2[j]=summary(model2)$coef[2,1]
standerror2[j]=summary(model2)$coef[2,2]

##optimal

m.out3<-matchit(swang1~cat1+ca+cardiohx + chfhx + dementhx
+ psychhx+ chrpulhx +renalhx +liverhx + gibledhx +malighx +
immunhx+ transhx+amihx+age+sex+edu+surv2md1+das2d3pc+aps1+
scoma1+meanbp1+pafi1+alb1+hema1+bili1+crea1
+sod1+pot1+paco21+ph1+wtkilo1+dnr1+ninsclas+resp+card+
neuro+gastr+renal+meta+hema+seps+trauma+ortho+race+income,
method="optimal", data=bootsample)
m.out3

final_data3<-match.data(m.out3)

model3 <-glm(as.factor(dth30) ~ swang1, family="binomial",
data=final_data3)

est3[j]=summary(model3)$coef[2,1]
standerror3[j]=summary(model3)$coef[2,2]

##genetic

m.out4<-matchit(swang1~cat1+ca+cardiohx + chfhx + dementhx + psychhx
+ chrpulhx +renalhx +liverhx +gibledhx +malighx + immunhx+ transhx
+amihx+age+sex+edu+surv2md1+das2d3pc+aps1+scoma1+meanbp1+pafi1+alb1
+hema1+bili1+crea1+sod1+pot1+paco21+ph1+wtkilo1+dnr1+ninsclas+resp
+card+neuro+gastr+renal+meta+hema+seps+trauma+ortho+race+income,
method="genetic", data=bootsample)
m.out4
```

```
final_data4<-match.data(m.out4)

model4 <-glm(as.factor(dth30) ~ swang1, family="binomial",
data=final_data4)

##est4[j]=summary(model4)$coef[2,1]
##standerror4[j]=summary(model4)$coef[2,2]

}

sd(est1)
sd(est2)
sd(est3)
sd(est4)


#####################################################################
##############Monte-carlo Simulation##################################
#####################################################################
set.seed(053013)
n=1000
rep=5                    # number of data sets
scenario="A"
W=matrix(0,n,10)
beta=matrix(c(0.8,-0.25,0.6,-0.4,-0.8,-0.5,0.7,0,0,0),10,1)
alpha=matrix(c(0.3,-0.36,-0.73,-0.2,0.71,-0.19,0.26),7,1)

F.sample.cor=function(x, rho) {
              y <- (rho * (x - mean(x)))/sqrt(var(x)) +
                        sqrt(1 - rho^2) * rnorm(length(x))
              #cat("Sample corr = ", cor(x, y), "\n")
              return(y)
     }

BiasNN=rep(NA,rep)
standerrorNN=rep(NA,rep)

BiasF=rep(NA,rep)
standerrorF=rep(NA,rep)

BiasO=rep(NA,rep)
```

```
standerrorO=rep(NA,rep)

BiasC=rep(NA,rep)
standerrorC=rep(NA, rep)

standerrorG=rep(NA,rep)
BiasG=rep(NA,rep)

BiasS=rep(NA,rep)
standerrorS=rep(NA,rep)

for (i in 1:rep){
W[,1]=rnorm(n, mean=0, sd=1)
W[,2]=rnorm(n, mean=0, sd=1)
W[,3]= rnorm(n, mean=0, sd=1)
W[,4]=rnorm(n, mean=0, sd=1)
W[,5]=F.sample.cor(W[,1], 0.2)
W[,6]=F.sample.cor(W[,2], 0.9)
W[,7]=rnorm(n, mean=0, sd=1)
W[,8]=F.sample.cor(W[,3], 0.2)
W[,9]=F.sample.cor(W[,4], 0.9)
W[,10]=rnorm(n, mean=0, sd=1)

W[,1]=ifelse(W[,1]>mean(W[,1]),1,0)  # dichotomizing covariates
W[,3]=ifelse(W[,3]>mean(W[,3]),1,0)
W[,5]=ifelse(W[,5]>mean(W[,5]),1,0)
W[,6]=ifelse(W[,6]>mean(W[,6]),1,0)
W[,8]=ifelse(W[,8]>mean(W[,8]),1,0)
W[,9]=ifelse(W[,9]>mean(W[,9]),1,0)

if (scenario=="A"){
P1=1/(1+exp(1-W%*%beta))  # true propensity score
} else
if (scenario=="B"){
P1=1/(1+exp(-(W%*%beta+beta[2]*W[,2]^2)))
} else
if (scenario=="C"){
P1=1/(1+exp(-(W%*%beta+beta[2]*W[,2]^2+beta[4]*W[,4]^2+beta[7]*W[,7]^2)))
} else
if (scenario=="D"){
P1=1/(1+exp(-(W%*%beta+beta[1]*0.5*W[,1]*W[,3]+beta[2]*0.7*W[,2]*W[,4]
```

```
+beta[4]*0.5*W[,4]*W[,5]+beta[5]*0.5*W[,5]*W[,6])))
} else
if (scenario=="E"){
P1=1/(1+exp(-(W%*%beta+beta[2]*W[,2]^2+beta[1]*0.5*W[,1]*W[,3]+
beta[2]*0.7*W[,2]*W[,4]+beta[4]*0.5*W[,4]*W[,5]+beta[5]*0.5*W[,5]*W[,6])))
} else
if (scenario=="F"){
P1=1/(1+exp(-(W%*%beta+beta[1]*0.5*W[,1]*W[,3]+beta[2]*0.7*W[,2]*W[,4]
+beta[3]*0.5*W[,3]*W[,5]+beta[4]*0.7*W[,4]*W[,6]+beta[5]*0.5*W[,5]*W[,7]
+beta[1]*0.5*W[,1]*W[,6]+beta[2]*0.7*W[,2]*W[,3]
+beta[3]*0.5*W[,3]*W[,4]+beta[4]*0.5*W[,4]*W[,5]+beta[5]*0.5*W[,5]*W[,6])))
}else
{
P1=1/(1+exp(-(W%*%beta+beta[2]*W[,2]^2+beta[4]*W[,4]^2+beta[7]*W[,7]^2+
beta[1]*0.5*W[,1]*W[,3]+beta[2]*0.7*W[,2]*W[,4]+beta[3]*0.5*W[,3]*W[,5]
+beta[4]*0.7*W[,4]*W[,6]+beta[5]*0.5*W[,5]*W[,7]
+beta[1]*0.5*W[,1]*W[,6]+beta[2]*0.7*W[,2]*W[,3]+
beta[3]*0.5*W[,3]*W[,4]+beta[4]*0.5*W[,4]*W[,5]
+beta[5]*0.5*W[,5]*W[,6])))
}

T=as.numeric(runif(n)<P1)        #generate treatment and control groups
V=W[,c(-5,-6,-7)]
m=exp(V%*%alpha-0.4*T)/(1+exp(V%*%alpha-0.4*T))
Y=as.numeric(runif(n)<m)  #binary outcome based on covariates and group

Data<-data.frame(Y=Y,T=T,W=W)

#########################
##matching methods#####
#########################

###nearest neighbour matching

require(MatchIt)

###m.out1<-matchit(T~W, method="nearest", data=as.data.frame(Data),
###ratio=1,replace=TRUE)

###m.out1<-matchit(T~W, method="nearest", data=as.data.frame(Data),
###ratio=1)
```

```
m.out1<-matchit(T~W.1+W.2+W.3+W.4+W.5+W.6+W.7+W.8+W.9+W.10,
method="nearest", data=Data, ratio=1)
finaldataS1<-match.data(m.out1)

###modelNN<-glm(Y~T+V, data=finaldataS1, family="binomial")

modelNN<-glm(Y~T+W.1+W.2+W.3+W.4+W.8+W.9+W.10,
data=finaldataS1,family="binomial")
BiasNN[i]<-summary(modelNN)$coef[2,1]-(-0.4)

standerrorNN[i]=summary(modelNN)$coef[2,2]

###full matching

require(optmatch)
require(MatchIt)
####m.out2<-matchit(T~W, method="full", data=as.data.frame(Data))
m.out2<-matchit(T~W.1+W.2+W.3+W.4+W.5+W.6+W.7+W.8+W.9+W.10,
method="full", data=Data)
finaldataS2<-match.data(m.out2)
##modelF<-glm(Y~T+V, data=finaldataS2, family="binomial")


modelF<-glm(Y~T+W.1+W.2+W.3+W.4+W.8+W.9+W.10,
data=finaldataS2,family="binomial")

BiasF[i]<-summary(modelF)$coef[2,1]-(-0.4)

standerrorF[i]=summary(modelF)$coef[2,2]

###optimal matching

m.out3<-matchit(T~W.1+W.2+W.3+W.4+W.5+W.6+W.7+W.8+W.9+W.10,
method="optimal", data=Data)

finaldataS3<-match.data(m.out3)
##modelO<-glm(Y~T+V, data=finaldataS3, family="binomial")

modelO<-glm(Y~T+W.1+W.2+W.3+W.4+W.8+W.9+W.10,
data=finaldataS3,family="binomial")
```

```
BiasO[i]<-summary(modelO)$coef[2,1]-(-0.4)
standerrorO[i]=summary(modelO)$coef[2,2]

##Genetic matching
m.out5<-matchit(T~W.1+W.2+W.3+W.4+W.5+W.6+W.7+W.8+W.9+W.10,
method="genetic", data= Data)

finaldataS5<-match.data(m.out5)

modelG<-glm(Y~T+W.1+W.2+W.3+W.4+W.8+W.9+W.10,data=finaldataS5,
family="binomial")
BiasG[i]<-summary(modelG)$coef[2,1]-(-0.4)
standerrorG[i]=summary(modelG)$coef[2,2]
}
mean(BiasNN)
sd(BiasNN)
mean(BiasF)
sd(BiasF)
mean(BiasG)
sd(BiasG)
mean(BiasO)
sd(BiasO)

#########################################################################
####################Sensitivity analysis#################################
#########################################################################
require(Matching)
Y<-newdata$dth30
Tr<-newdata$swang1
X<-cbind(cat1,ca,cardiohx, chfhx, dementhx,psychhx,chrpulhx,renalhx,
liverhx,gibledhx,malighx,immunhx,transhx,amihx,age,sex,edu,surv2md1,
das2d3pc,aps1,scoma1,meanbp1,pafi1,alb1,hema1,bili1,crea1,sod1,
pot1,paco21,ph1,wtkilo1,dnr1,ninsclas,resp,card,neuro,gastr,renal,
meta,hema,seps,trauma,ortho,race,income)

gen1<-GenMatch(Tr=Tr,X=X,pop.size=50,data.type.int=FALSE, print=0,
 replace=FALSE)

###############################################
###sensitivity analysis for genetic matching###
###############################################
```

```
mgen1<-Match(Y=Y, Tr=Tr, X=X, Weight.matrix=gen1, replace=FALSE)
summary(mgen1)
require(rbounds)
###Function to calculate Rosenbaum bounds for binary data
binarysens(mgen1,Gamma=2, GammaInc=.1)

##Amplification
gamma <- 1.1
delta <- seq(0,3, by=.01)
lambda <- ((gamma*delta) - 1)/(delta - gamma)
lambda <- round(lambda, 2)

plot(lambda,delta, ylim=c(0,max(delta)), xlim=c(1,10), type="l")
abline(v=gamma, lty=2)
abline(h=gamma, lty=2)

#Amplification Set
cbind(delta,lambda)

gamma <- 1.6
delta <- seq(0,3, by=.01)
lambda <- ((gamma*delta) - 1)/(delta - gamma)
lambda <- round(lambda, 2)

plot(lambda,delta, ylim=c(0,max(delta)), xlim=c(1,10), type="l")
abline(v=gamma, lty=2)
abline(h=gamma, lty=2)

#Amplification Set
cbind(delta,lambda)


##################################################
###sensitivity analysis for NN matching#######
##################################################
mgen1<-Match(Y=Y, Tr=Tr, X=X, M=1,replace=FALSE)
summary(mgen1)
###Function to calculate Rosenbaum bounds for binary data
binarysens(mgen1,Gamma=2, GammaInc=.1)
```

```
#################################################
###sensitivity analysis for optimal matching###
#################################################
library(rbounds)

tr1.out.1 <- final_data3[final_data3$swang1==1, c("dth30" , "subclass")]
tr1.out.1 <- na.omit(tr1.out.1)
co1.out.1 <- final_data3[final_data3$swang1==0, c("dth30", "subclass")]
co1.out.1 <- na.omit(co1.out.1)
#Table of Matched Outcomes
out1.tab <- matrix(c(table(tr1.out.1$dth30, co1.out.1$dth30)), nrow = 2,
dimnames = list(Treated = c("die", "alive"), Control = c("die", "alive")))

d.pairs1 <- out1.tab[2]
d.pairs2 <- out1.tab[3]

binarysens(d.pairs1 , d.pairs2 , Gamma=2, GammaInc=.1)

#######################################################
###sensitivity analysis for full matching########
#######################################################

rdata <- read.csv("mydatafull.csv")
mdata <- as.matrix(rdata)
sIndex <- unique(mdata[, 4])

dCounts <- NULL
treatCounts <- NULL
for( i in 1: length(sIndex)){
        sNum <- sIndex[i]
        idx <- which(mdata[,4] == sNum)
        sData <- mdata[idx,]
        ctrlIsOne <- TRUE
        oneIndex <- which(sData[,2 ] == 0)
        if (length(oneIndex) != 1){
                oneIndex <- which( sData[,2 ] == 1)
                ctrlIsOne <- FALSE
        }
        oneRes <- sData[oneIndex, 5]
        neIndex <- which(sData[ -oneIndex, 5] != oneRes)
        dCounts <- c(dCounts, length(neIndex))
```

```
        if ( (ctrlIsOne && oneRes == 1) || ((!ctrlIsOne && oneRes == 0))){
                treatCounts <- c(treatCounts, 0)
        }else{
                treatCounts <- c(treatCounts, length(neIndex))
        }
}

dCounts
treatCounts

sens.analysis.mcnemar=function(D, Tobs,Gamma)
{

        p.positive=Gamma/(1+Gamma);
        p.negative=1/(1+Gamma);
        lowerbound=1-pbinom(Tobs-1,D,p.negative);
        upperbound=1-pbinom(Tobs-1,D,p.positive);

        list(lowerbound=lowerbound,upperbound=upperbound);

}

sr1 <- sens.analysis.mcnemar(sum(dCounts), sum(treatCounts), 1.1)
sr2 <- sens.analysis.mcnemar(sum(dCounts), sum(treatCounts), 1.2)
sr3 <- sens.analysis.mcnemar(sum(dCounts), sum(treatCounts), 1.3)

#Amplification
par(mfrow=c(2,2))

gamma <- 1.2
delta <- seq(0,3, by=.01)
lambda <- ((gamma*delta) - 1)/(delta - gamma)
lambda <- round(lambda, 2)

plot(lambda,delta, ylim=c(0,max(delta)), xlim=c(1,10), type="l",
 main="Optimal matching: Gamma=1.2")
abline(v=gamma, lty=2)
abline(h=gamma, lty=2)
legend(6,3, c("Amplification","Asymptote"),lty=c(1,2), lwd=c(0.5,0.5),cex=0.5)

gamma <- 1.3
```

```
delta <- seq(0,3, by=.01)
lambda <- ((gamma*delta) - 1)/(delta - gamma)
lambda <- round(lambda, 2)

plot(lambda,delta, ylim=c(0,max(delta)), xlim=c(1,10), type="l",
main="Genetic matching: Gamma=1.3")
abline(v=gamma, lty=2)
abline(h=gamma, lty=2)
legend(6,3, c("Amplification","Asymptote"),lty=c(1,2), lwd=c(0.5,0.5),cex=0.5)
#Amplification Set
cbind(delta,lambda)


gamma <- 1.6
delta <- seq(0,3, by=.01)
lambda <- ((gamma*delta) - 1)/(delta - gamma)
lambda <- round(lambda, 2)

plot(lambda,delta, ylim=c(0,max(delta)), xlim=c(1,10), type="l",
main="NN matching: Gamma=1.6")
abline(v=gamma, lty=2)
abline(h=gamma, lty=2)
legend(6,3, c("Amplification","Asymptote"),lty=c(1,2), lwd=c(1.5,1.5),cex=0.5)
#Amplification Set
cbind(delta,lambda)

gamma <- 1.7
delta <- seq(0,3, by=.01)
lambda <- ((gamma*delta) - 1)/(delta - gamma)
lambda <- round(lambda, 2)

plot(lambda,delta, ylim=c(0,max(delta)), xlim=c(1,10), type="l",
main="Full matching: Gamma=1.7")
abline(v=gamma, lty=2)
abline(h=gamma, lty=2)
legend(6,3, c("Amplification","Asymptote"),lty=c(1,2), lwd=c(1.5,1.5),cex=0.5)
#Amplification Set
cbind(delta,lambda)
```

Bibliography

Alex Bryson, R. D. and Purdon, S. (2002) The use of propensity score matching in the evaluation of active labour market policies. Working Paper No. 4, Department for Work and Pensions, 2002.

Alexis Diamond, J. S. S. (2012) Genetic matching for estimating causal effects: A general multivariate matching method for achieving balance in observational studies. Review of Economics and Statistics 95:932–945.

Austin, P. (2009) Some methods of propensity-score matching had superior performance to others: Results of an empirical investigation and monte carlo simulations. Biometrical Journal 51(1):171–184.

Bergersen, L., Gauvreau, K., Foerster, S., Marshall, A., McElhinney, D. B., III, R. B., Hirsch, R., Kreutzer, J., Balzer, D., Vincent, J., Hellenbrand, W., Holzer, R., Cheatham, J., Moore, J., Burch, G., Armsby, L., Lock, J. and Jenkins, K. (2011) Catheterization for congenital heart disease adjustment for risk method (charm). JACC: Cardiovascular Interventions 4(9):1037 – 1046.

Bergner, M., Warren, B., Damiano, A., Hakim, R., Murphy, D. J., Teno, J., Virnig, B., Wagner, D. P., Wu, A. W., Yasui, Y., Robinson, D. K., Connors, Alfred F., J., Kreling, B., Dulac, J., Baker, R., Holayel, S., Meeks, T., Mustafa, M., Vegarra, J., Alzola, C., Harrell, Frank E., J., Cook, E. F., Dawson, N. V., Hamel, M. B., Peterson, L., Phillips, R. S., Tsevat, J., Forrow, L., Lesky, L., Davis, R., Kressin, N., Solzan, J., Puopolo, A. L., Desbiens, N. A., Barrett, L. Q., Bucko, N., Brown, D., Burns, M., Foskett, C., Hozid, A., Keohane, C., Martinez, C., McWeeney, D., Melia, D., Fulkerson, William J., J., Otto, S., Sheehan, K., Smith, A., Tofias, L., Arthur, B., Collins, C., Cunnion, M., Dyer, D., Kulak, C., Michaels, M., Goldman, L., O'Keefe, M., Parker, M., Tuchin, L., Wax, D., Weld, D., Hiltunen, L., Marks, G., Mazzapica, N., Medich, C., Soukup, J., Knaus, W. A., Califf, R. M., Galanos, A. N., Kussin, P., Muhlbaier, L. H., Winchell, M., Mallatratt, L., Akin, E., Belcher, L., Buller, E., Clair, E., Lynn, J., Drew, L., Fogelman, L., Frye, D., Fraulo, B., Gessner, D., Hamilton, J., Kruse, K., Landis, D., Nobles, L., Oliverio, R., Oye, R. K., Wheeler, C., Banks, N., Berry, S., Clayton, M., Hartwell,

P., Hubbard, N., Kussin, I., Norman, B., Noveau, J., Read, H. and Investigators, S. P. (1995) A controlled trial to improve care for seriously iii hospitalized patients: The study to understand prognoses and preferences for outcomes and risks of treatments (support). JAMA: The Journal of the American Medical Association 274(20):1591–1598.

Bertsekas, D. P. (1991) Linear Network Optimization: Algorithms and Codes. MIT Press.

Bhattacharya, J., Shaikh, A. M. and Vytlacil, E. (2012) Treatment effect bounds: An application to swan-ganz catheterization. Journal of Econometrics 168(2):223 – 243.

Cochran, W. G. and Rubin, D. B. (1973) Controlling bias in observational studies: A review. Sankhya: The Indian Journal of Statistics, Series A 35(4):417–446.

Connors, A. F., Speroff, T., Dawson, N. V. and Thomas, C. (1996) The effectiveness of right heart catheterization in the initial care of critically ill patients. JAMA 276(11):889.

Diamond, A. and Sekhon, J. S. (2013) Genetic matching for estimating causal effects: A general multivariate matching method for achieving balance in observational studies. The review of economics and statistics 95(3):932–945.

Donald J. Murphy, W. A. K. and Lynn, J. (1990) Study population in support. Journal of Clinical Epidemiology 43:s11–s28.

Elliott, C., Brown, L., Farber, H., Poms, A., Liou, T., Raskob, G., Saydain, G., Turner, M. and McGoon, M. (2011) Right heart catheterization in patients with pulmonary arterial hypertension: Practice patterns observed in the reveal registry. The Journal of Heart and Lung Transplantation 30(4, Supplement):S80. Abstract Issue: International Society for Heart and Lung Transplantation Thirty-First Annual Meeting and Scientific Sessions.

Hansen, B. B. (2004) Full matching in an observational study of coaching for the sat. Journal of the American Statistical Association 99:609–619.

Heckman, J. J. (1979) Sample selection bias as a specification error. Econometrica 47:153–61.

Heckman, J. J., Ichimura, H. and Todd, P. E. (1997) Matching as an econometric evaluation estimator: Evidence from evaluating a job training programme. Review of Economic Studies 64:605–654.

Hemmerling, T. M., Cyr, S. and Terrasini, N. (2013) Epidural catheterization in cardiac surgery: The 2012 risk assessment. Annals of Cardiac Anaesthesia 16(3):169 – 177.

Hirano, K. and Imbens, G. W. (2001) Estimation of causal effects using propensity score weighting: An application to data on right heart catheterization. Health Services and Outcomes Research Methodology 2(3):259–278.

Ho, D. E., Imai, K., King, G. and Stuart, E. A. (2007) Matching as nonparametric preprocessing for reducing model dependence in parametric causal inference. Political Analysis 15:199–236.

Ho, D. E., Imai, K., King, G. and Stuart, E. A. (2011) Matchit: Nonparametric preprocessing for parametric causal inference. Journal of Statistical Software 42(8).

Hoeper, M. M. and Lee, S. H. (2006) Complications of right heart catheterization procedures in patients with pulmonary hypertension in experienced centers. Journal of the American College of Cardiology 48(12):2546–2552.

Holland, P. (1986) Statistics and causal inference. Statistics and causal inference 81:945–960.

Iacus, Stefano M., G. K. and Porro, G. (2011) Multivariate matching methods that are monotonic imbalance bounding. Journal of the Americian Statistical Association 106:345–361.

Imbens, G. W. (2000) The role of the propensity score in estimating dose-response functions. Biometrika 87(3):706–710.

Kang, J. D. and Schafer, J. L. (2007) Demystifying double robustness: A comparison of alternative strategies for estimating a population mean from incomplete data. Statistical Science 22:523–539.

Kaplan, D. (2005) Causal inference in educational policy research. Working paper, Wisconsin Center for Education Research, WI.

Keele, L. (2010) An overview of rbounds: An r package for rosenbaum bounds sensitivity analysis with matched data. Working Paper, Penn State University.

King, G. and Zeng, L. (2006) The dangers of extreme counterfactuals. Political Analysis 14:131–159.

Knaus, W. A., Harrell, F. E., Lynn, J., Goldman, L., Phillips, R. S., Connors, A. F., Dawson, N. V., Fulkerson, W. J., Califf, R. M., Desbiens, N., Layde, P., Oye, R. K., Bellamy, P. E., Hakim, R. B. and Wagner, D. P. (1995) The support prognostic model: Objective estimates of survival for seriously ill hospitalized adults. Annals of Internal Medicine 122(3):191–203.

Lechner, M. (2001) Identification and estimation of causal effects of multiple treatments under the conditional independence assumption. In: Lechner, M. and Pfeiffer, F. (eds.), Econometric Evaluation of Labour Market Policies, Physica-Verlag HD, vol. 13 of *ZEW Economic Studies*, pp. 43–58.

Morton, B. C., Higginson, L. A. and Beanlands, D. S. (1993) Death in a catheterization laboratory. Canadian Medical Association Journal 149(2):165–169.

Natarajan, M. K., Mehta, S. R., Holder, D. H., Afzal, Teo, K. and Yusuf, S. (2002) The risks of waiting for cardiac catheterization: a prospective study. Canadian Medical Association Journal 167(11):1233–1240.

Pearl, J. (2009) Causality: Models, Reasoning, and Inference. Cambridge University Press.

Polanczyk, C., Rohde, L. and Goldman, L. (2001) Right heart catheterization and cardiac complications in patients undergoing noncardiac surgery: An observational study. JAMA 286(3):309–314.

Robert, R. and Richard, G. (1968) Cooperative study on cardiac catheterization. coronary arteriography. Circulation 37(suppl 3):67–73.

Rosenbaum, P. R. (1980) Randomization analysis of experimental data: the fisher randomization test. Journal of American Statistics Association 75:591–593.

Rosenbaum, P. R. (1987) The role of a second control group in an observational study. Statistical Science 2:292–316.

Rosenbaum, P. R. (1989) Optimal matching for observational studie. Journal of theAmerican Statistical Association 84:1024–1032.

Rosenbaum, P. R. (2002) Observational Studies, 2nd ed. Springer, New York.

Rosenbaum, P. R. (2005) Sensitivity Analysis in Observational Studies. John Wiley & Sons.

Rosenbaum, P. R. and Rubin, D. B. (1983) The central role of the propensity score in observational studies for causal effects. Biometrika 70(1):41–45.

Rosenbaum, P. R. and Rubin, D. B. (1985) The bias due to incomplete matching. Biometrics 41:106–116.

Rubin, D. B. (1974) Estimating causal effects of treatments in randomized and nonrandomized studies. Journal of Educational Psychology 66(5):688–701.

Rubin, D. B. (1976) Multivariate matching methods that are equally percent bias reducing, ii: Maximums on bias reduction for fixed sampled sizes. Biometrics 32:121–132.

Rubin, D. B. (1980) Bias reduction using mahalanobis metric matching. Biometrics 36:293–298.

Rubin, D. B. (2001) Using propensity scores to help design observational studies: application to the tobacco litigation. Health Services and Outcomes Research Methodology 2:169–188.

Rubin, D. B. and Stuart, E. A. (2006) Affinely invariant matching methods with discriminant mixtures of proportional ellipsoidally symmetric distributions. The Annals of Statistics 34:1814–1826.

Rubin, D. B. and Thomas, N. (1992) Affinely invariant matching methods with ellipsoidal distributions. The Annals of Statistics 20:1079–93.

Sekhon, J. S. (2010) Opiates for the matches: Matching methods for causal inference. SSRN eLibrary, available at http://ssrn.com/paper=1600553. Accessed May 15, 2010.

Sekhon, J. S. (2011) Multivariate and propensity score matching software with automated balance optimization: The matching package for r. Journal of Statistical Software 42(7).

Setoguchi, S., Schneeweiss, S., Brookhart, M. A., Glynn, R. J. and Cook, E. F. (2008) Evaluating uses of data mining techniques in propensity score estimation: a simulation study. Pharmacoepidemiology and Drug Safety 17:546–555.

Shanes, J. G., Stein, M. A., Dierenfeldt, B. J. and Kondos, G. T. (1987) The value of routine right heart catheterization in patients undergoing coronary arteriography. Am Heart J 113(5):1261–1263.

Silber, J. H. and Rosenbaum, P. R. (2009) Amplification of sensitivity analysis in matched observational studies. Journal of the American Statistical Association 104(488):1398–1405.

Smith, H. L. (1997) Matching with multiple controls to estimate treatment effects in observational studies. Sociological Methodology 27:325–353.

Steyer, R. (2005) Analyzing individual and average causal effects via structural equation models. Methodology 1:39–64.

Stuart, E. A. (2010) Matching methods for causal inference: A review and a look forward. Statistical Science 25(1):1–21.

Stuart, E. A. and Green, K. (2008) Using full matching to estimate causal effects in nonexperimental studies: examining the relationship between adolescent marijuana use and adult outcomes. Developmental Psychology 44(2):395–406.

Vita

Qiong Yang

Qiong Yang was born in Jiangsu, China. After completing her schoolwork at siyang High School in 2000, Qiong entered Renmin University of China in Beijing, China and was awarded with a Bachelor of Arts in 2004. For the year 2004 to year 2006, she attended Peking University majoring in Economics and received a Master of Science degree. Since 2006, she left for US and started her studies in Pennsylvania State University. Master of Science degree in Economics was awarded to her in May 2008. Since 2008, she started her pursuit of degree in Agricultural Economics. Since 2011, she started her pursuit of concurrent Master of Science degree in Statistics.