Matching to Produce Causal Estimates in Non-experimental Settings

A Senior Comprehensive Paper

presented to
the Faculty of Carleton College
Department of Mathematics and Statistics
David Watson, Advisor

by
Kaitlyn Cook Tom Grodzicki Harrison Reeder Ji Min Yoo

March 2, 2015

Abstract

In perhaps every field, the most fundamental research questions often ask: "what is the relationship between X and Y? In particular, does X cause Y?" Causal inference is devoted to uncovering and improving research methods that enable investigators to measure the size and significance of causal effects of treatment on outcomes. In this paper, we reflect upon and extend the development of causal inference techniques in a variety of contexts: randomized experiments, observational studies, and retrospective studies. We use a dataset capturing the effect of a job training program on unemployment in order to evaluate the performances of causal inference techniques in each of these types of study design, and also propose a novel approach to retrospective causal inference called the Retrospective Exposure Case (REC) Score. We find that matching using the propensity score method in the prospective setting offers a marked improvement over non-matching approaches, and comes very close to recovering the experimental estimate; we further see promise in the REC Score matching analysis, which bests all other retrospective and prospective analyses we attempted.

Contents

1. Introduction to Causal Inference

In statistics, we are often interested in questions of causality: does a particular treatment cause an observed outcome? And if so, can we quantify the size and direction of that effect? A useful framework for conceptualizing a causal effect is potential outcomes. Throughout this paper, we consider the effect of a binary treatment ("treated" vs. "non-treated") on a binary outcome of interest ("case" vs. "non-case"). Let z be an indicator for the treatment $t \in \{0, 1\}$, where $z = 1$ indicates that an individual received the treatment and $z = 0$ indicates that they did not. Similarly, let r^{obs} be an indicator for the observed outcome $y \in \{0, 1\}$, where $r^{obs} = 1$ represents a "case" and $r^{obs} = 0$ represents a "non-case". Suppose we are interested in determining the effect of this treatment z, applied to a particular unit i, on this observable outcome r. We can then imagine quantifying the *causal effect* of this treatment by comparing the outcome given unit i received the treatment to the outcome given unit i did not receive the treatment. However, in reality unit i is either treated or non-treated, meaning that only one of these potential outcomes is ever observed. Thus, the problem of determining causal effect is—fundamentally—a missing data problem: how do we best estimate the causal effect when the other, unobserved potential outcome is unknown?

In order to perform this estimation, we need to broaden our understanding of causal effect.[1] Because we can only observe one outcome per unit, we must look at causal effect on a population, rather than an individual, level: instead of comparing one unit's outcome with the treatment to that same unit's outcome without the treatment, we compare the outcomes of a group of treated units with the outcomes of a group of non-treated units.

1.1 The Odds Ratio

The most critical question driving any quantitative research is simply "what relationship between the potential outcomes are we estimating?" The formal definitions of such relationships, which drive the estimation process, are called *estimands*. In causal inference for binary outcomes—i.e., $r^{obs} \in \{0, 1\}$—the most appropriate estimand is typically the odds ratio, denoted γ. We define γ as the odds ratio for a particular outcome in the treated group relative to the non-treated group:

$$\gamma = \frac{\frac{\bar{r}^{obs}_{z=1}}{1-\bar{r}^{obs}_{z=1}}}{\frac{\bar{r}^{obs}_{z=0}}{1-\bar{r}^{obs}_{z=0}}} \tag{1.1}$$

where $\bar{r}^{obs}_{z=1}$ and $\bar{r}^{obs}_{z=0}$ represent the mean proportion of cases in the treated group and non-treated group, respectively. A particular advantage of this estimand is that if we assume that treatment is randomly assigned, then there are equivalent estimators in each study design context—randomized experiment, prospective, and retrospective—for this estimand.

In the context of a randomized experiment (or a simple random sample), we write that the probability of having treatment status $z = t \in \{0, 1\}$ and outcome case status $r^{obs} = y \in \{0, 1\}$ is $\pi_{ty} = P(z = t, r^{obs} = y)$. We can thus construct the following contingency table:

	Non-Case (y=0)	Case (y=1)
Non-Treated (t=0)	π_{00}	π_{01}
Treated (t=1)	π_{10}	π_{11}

Table 1.1: Probabilities of Treatments and Outcomes in a Randomized Experiment

In this context, the estimator for γ is:

[1] While the present paper serves as an introduction to this topic, for a more complete survey of the field see [3].

$$\hat{\gamma}_{srs} = \frac{\frac{\pi_{11}}{\pi_{10}}}{\frac{\pi_{01}}{\pi_{00}}} = \frac{\pi_{00}\pi_{11}}{\pi_{01}\pi_{10}} \tag{1.2}$$

One common alternative to a randomized experiment is a prospective study, in which researchers sample a cohort of both treated and non-treated units, and then follow those cohorts for a set amount of time before comparing their final outcomes. The researchers do not intervene in the assignment of the treatment. Thus, in the case of a prospective study, we look not merely at probabilities relating treatment and outcome, but conditional probabilities of particular outcomes given treatment status. Let $\alpha_{ty} = P(r^{obs} = y|z = t) = \frac{P(z=t,r^{obs}=y)}{P(z=t)}$. Then in the prospective setting, we have:

	Non-Case (y=0)	Case (y=1)
Non-Treated (t=0)	α_{00}	α_{01}
Treated (t=1)	α_{10}	α_{11}

Table 1.2: Probabilities of Treatments and Outcomes in a Prospective Study

Conveniently we can still estimate the same odds ratio estimand in the prospective setting as the randomized experiment:

$$\hat{\gamma}_{pro} = \frac{\frac{\alpha_{11}}{\alpha_{10}}}{\frac{\alpha_{01}}{\alpha_{00}}} = \frac{\frac{\pi_{11}/(\pi_{11}+\pi_{10})}{\pi_{10}/(\pi_{11}+\pi_{10})}}{\frac{\pi_{01}/(\pi_{01}+\pi_{00})}{\pi_{00}/(\pi_{01}+\pi_{00})}} = \frac{\pi_{00}\pi_{11}}{\pi_{01}\pi_{10}} \tag{1.3}$$

Another problem that arises in prospective study designs is when one (or more) outcomes of interest are rare. Even a large sample may result in an extremely small number of units with this outcome, which impedes valid estimations of causal effects. When we are interested in binary outcomes $r^{obs} = 0$ and $r^{obs} = 1$, *retrospective* study designs solve this last problem by sampling "case" units from the population with outcome $r^{obs} = 1$ and "non-case" units from the population with outcome $r^{obs} = 0$. This allows the rare outcome $r^{obs} = 1$ to be "over-sampled".

However, given that units are sampled on the basis of their case/non-case status, when we look at retrospective study designs, we look at probabilities of being treated, conditioned on the known outcome. That is, if we let $\beta_{ty} = P(z = t|r^{obs} = y) = \frac{P(r^{obs}=y,z=t)}{P(r^{obs}=y)}$. Then in the retrospective setting we have:

	Non-Case (y=0)	Case (y=1)
Non-Treated (t=0)	β_{00}	β_{01}
Treated (t=1)	β_{10}	β_{11}

Table 1.3: Probabilities of Treatments and Outcomes in a Retrospective Study

Then the estimator of the odds ratio is given by:

$$\hat{\gamma}_{ret} = \frac{\frac{\beta_{11}}{\beta_{10}}}{\frac{\beta_{01}}{\beta_{00}}} = \frac{\frac{\pi_{11}/(\pi_{01}+\pi_{11})}{\pi_{10}/(\pi_{10}+\pi_{00})}}{\frac{\pi_{01}/(\pi_{01}+\pi_{11})}{\pi_{00}/(\pi_{10}+\pi_{00})}} = \frac{\pi_{00}\pi_{11}}{\pi_{01}\pi_{10}} \tag{1.4}$$

Thus, when there is a particular treatment and binary outcome of interest in a population, whether we perform a randomized experiment, a prospective study or a retrospective study, the odds ratio estimand allows for causal inference on the same quantity in each design.

1.2 Covariate Imbalance

To perform causal inference on γ, or any estimand, at a population level, it is critical to ensure that after the sampling procedure is finished, both groups—those receiving the treatment of interest and those not—are comparable. We want to ensure that any difference in observed outcomes between the groups is attributable to the treatment, and not to any other covariates. One way of measuring this similarity is to compare

the distributions of the covariates between both treated and non-treated groups. If the distributions of the covariates are reasonably similar, then we consider the two groups to have *covariate balance*; if the distributions differ, then there is *covariate imbalance* between the treatment and non-treatment groups. Covariate imbalance presents a substantial hurdle to causal inference techniques, and is often directly related to both the sampling method and treatment assignment mechanism. As such, the choice of study design—whether a randomized experiment, prospective study, or retrospective study is conducted—directly impacts our ability to draw causal conclusions.

In a randomized experiment, because the treatment assignment is random, it is independent from any observed or unobserved covariates. Thus, we can assume that there is, on average, covariate balance between the treated and non-treated groups. Yet there are several practical concerns that limit the feasibility of conducting a randomized experiment: they are often more expensive to conduct, and if the treatment of interest is potentially harmful, there are ethical barriers to randomizing and observing its effects. As such, observational studies—namely prospective studies and retrospective studies—are often more practical and convenient to perform. Yet these two study designs also present several hurdles to establishing covariate balance between the treatment and non-treatment groups. In prospective studies, because the researchers do not intervene in the assignment of the treatment, treated units may substantially differ from non-treated units; we thus cannot be sure that any observed difference in outcomes is due to a treatment effect, and not instead to this covariate imbalance. This problem is further exacerbated in the retrospective context. Not only is the sampling mechanism not fully random, but it also explicitly relies on characteristics of the subjects that are intimately related to their covariates—namely, their case status—to determine who is sampled. This process makes covariate imbalance—both between the treated and non-treated units as well as the case and non-case units—much more likely.

Given that causal inference theory relies on covariate balance between treated and non-treated units, the imbalance that results from either prospective or retrospective sampling poses challenges for estimating causal effects.

1.3 Basic Estimation Practices

1.3.1 Unadjusted Analysis

Because in a randomized experiment the assignment mechanism creates covariate balance between treated and non-treated groups, the observed outcomes in each group can simply be averaged to estimate potential outcomes. Such analysis does not involve modeling or any adjustments based on the covariates of individual subjects, and thus we call it *unadjusted* or *naive* analysis. So, if there were n subjects in each of the treated and non-treated groups, the unadjusted estimations of the potential outcomes are $\bar{r}^{obs}_{z=1} = \frac{1}{n}\sum_{i:z_i=1} r^{obs}_i$ and $\bar{r}^{obs}_{z=0} = \frac{1}{n}\sum_{i:z_i=0} r^{obs}_i$, and the estimator of γ is given by

$$\hat{\gamma} = \frac{\frac{\bar{r}^{obs}_{z=1}}{1-\bar{r}^{obs}_{z=1}}}{\frac{\bar{r}^{obs}_{z=0}}{1-\bar{r}^{obs}_{z=0}}} \tag{1.5}$$

Similarly, we can perform an unadjusted analysis with prospective or retrospective study designs, however this is almost never done because of the bias induced by covariate imbalance. The results of such estimation approaches are often overwhelmed by bias in the assignment mechanism. Thus, the utility of this approach is typically that it is an interesting point of comparison to measure the bias reduction of complex analytical approaches.

1.3.2 Regression Approaches to Addressing Covariate Imbalance

Performing a regression analysis as oppposed to a naive unadjusted analysis allows us to take into account the reality that other covariates may be contributing to the observed difference in outcomes. In particular, regression analysis attempts to isolate the effect of the treatment on the outcome from any other confounding covariate effects. As such, we consider regression analysis to be an *adjusted* analysis because the effect of treatment is adjusted in order to address any effects the covariates might have on the eventual observed outcome.

Let $p_i = P(r_i^{obs} = 1|x, z)$, where r_i^{obs} is the observed outcome for unit i. Then to calculate γ we use a logistic regression model of the form:

$$\log \frac{p_i}{1 - p_i} = \beta_0 + \beta^{treat} z_i + \beta x_i, i = 1, ..., n \tag{1.6}$$

where β is the vector of coefficients for the corresponding vector of covariates x_i of unit i. Then $\hat{\gamma} = e^{\hat{\beta}^{treat}}$, where $\hat{\beta}^{treat}$ is the estimated coefficient corresponding to the treatment variable—in other words, the adjusted treatment effect. We estimate coefficients using maximum likelihood estimation.

Critically, however, adjustment using regression approaches is typically insufficient to make causal inference. To many students of introductory regression analysis, the standard interpretation of regression coefficients after "controlling for" the effects of covariates may seem to inherently remove the bias and correct covariate imbalance. However, the major assumption at work when interpreting the coefficient of treatment variable is that it represents the isolated treatment effect *when all other covariates are held constant.* In many cases, however, the individuals who receive the treatment have notably different covariate distributions than non-treated individuals, such that there does not actually exist a corresponding non-treated person with similar covariates. Thus, to interpret the treatment coefficient of a regression analysis inherently extrapolates beyond the sample, which is not adequate to remove the bias introduced by covariate imbalance in non-randomized assignment mechanisms.

2. Matching

In prospective and retrospective studies, the covariate imbalance between treated and non-treated groups presents a substantial barrier to causal inference, which is not resolved by unadjusted estimation or typical logistic regression approaches. Another method of addressing this imbalance is matching: pairing individuals from the treatment group with corresponding individuals from the non-treatment group who have similar covariate values. This process creates treatment and non-treatment groups that "look alike", and for whom the confounding effect of the covariates on the outcome has been controlled; in short, we are able to "recreate" a randomized experimental setting. Thus, the observed outcome is independent of treatment assignment, meaning that any observed difference in outcomes can be attributed to the treatment itself.

The challenge to this approach comes from the difficulty of producing exact matches as the number of covariates increases. For example, for each treated person, it may be relatively straightforward to find a non-treated person of the same gender, but there won't necessarily be a non-treated person with the same gender, age, education level, income level, etc. Thus, exact matching is often both computationally intensive and impractical. In this section, we discuss current methods of summarizing the covariates in lower dimensions and thereby simplifying the matching process. We also propose a new method for matching in a retrospective setting.

2.1 Balancing Scores and the Propensity Score

We start by introducing the concept of a balancing score. Balancing scores are constructed so that the conditional distribution of the covariates x given the balancing score $b(x)$ is the same for both treated and non-treated units [5]. In other words, covariate balance has been restored.

Definition 2.1.1. A balancing score $b(x)$ is a function of the observed covariates x such that $x \perp\!\!\!\perp z|b(x)$.

We also introduce the concept of the propensity score $e(x)$, the conditional probability of being assigned the treatment given the observed covariates. We will further explain that $e(x)$ can serve as a one-dimensional balancing score.

Definition 2.1.2. $e(x) = P(z = 1|x)$

We now show that the propensity score is a balancing score. To do this, we first introduce the idea of a function—namely a balancing score—being "finer" or "coarser" than another function. In particular, a function $h(x)$ is finer than another function $g(x)$ if $g(x) = f(h(x))$ for a function f. We could equivalently say that $g(x)$ is coarser than $h(x)$.

Theorem 2.1.3. *Let $b(x)$ be a function of x. Then $b(x)$ is a balancing score, that is, $x \perp\!\!\!\perp z|b(x)$ if and only if $b(x)$ is finer than $e(x)$ in the sense that $e(x) = f(b(x))$ for some function f.*

Proof. First, we need to show that $x \perp\!\!\!\perp z|b(x)$ implies $e(x) = f(b(x))$.
A function is a relationship in which each input has a single output. So if $f(x)$ is a function and x_1, x_2 are inputs of the function, it follows that

$$x_1 = x_2 \implies f(x_1) = f(x_2).$$

Similarly, if $e(x)$ is a function of $b(x)$,

$$b(x_1) = b(x_2) \implies e(x_1) = e(x_2).$$

So, it is enough to show that, given $x \perp\!\!\!\perp z|b(x)$ and $b(x_1) = b(x_2)$, we can conclude $e(x_1) = e(x_2)$. Let's start with $e(x_1)$. By definition of the propensity score, $e(x) = P(z = 1|x)$. So

$$e(x_1) = P(z = 1|x_1)$$

Because $b(x)$ is a function of x, conditioning on $b(x)$ in addition to conditioning on x does not change the probability of z.

$$= P(z = 1|x_1) = P(z = 1|x_1, b(x_1))$$

Since $x \perp\!\!\!\perp z|b(x)$ and $b(x_1) = b(x_2)$,

$$\begin{aligned}
&= P(z = 1|b(x_1)) \\
&= P(z = 1|b(x_2)) \\
&= P(z = 1|b(x_2), x_2) \\
&= P(z = 1|x_2) \\
&= e(x_2)
\end{aligned}$$

Next, we need to show that $e(x) = f(b(x))$ implies $x \perp\!\!\!\perp z|b(x)$.
If $x \perp\!\!\!\perp z|b(x)$, then, using the conditional probability of independent events, it follows that $P(z|x, b(x)) = P(z|b(x))$. So, given $e(x) = f(b(x))$, it is sufficient to show that $P(z|b(x)) = P(z|x, b(x))$. We can write

$$\begin{aligned}
P(z|b(x)) &= 1 * P(z = 1|b(x)) + 0 * P(Z = 0|b(x)) \\
&= E[z|b(x)]
\end{aligned}$$

Using the law of iterated expectation,

$$\begin{aligned}
&= E[E[z|x, b(x)]|b(x)] \\
&= E[E[z|x]|b(x)] \\
&= E[e(x)|b(x)] \\
&= E[z|x, b(x)] \\
&= 1 * P(z = 1|x, b(x)) + 0 * P(z = 0|x, b(x)) \\
&= P(z = 1|x, b(x))
\end{aligned}$$

□

Theorem 2.1.4. *Treatment assignment and the observed covariates are conditionally independent given the propensity score $e(x)$, that is,*

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

Proof. Suppose that f is an identity function. Then the result follows immediately from Theorem 2.1.3 □

So we have that the propensity score $e(x)$—and any score that we can relate back to the propensity score—is a balancing score. As such, the propensity score is a way of "summarizing" the observed covariates simply to a number between 0 and 1, which is much easier to match on than matching on the full array of covariates x.

Furthermore, because matching on a balancing score helps to create covariate balance between the treated group and non-treated group, we are able to obtain unbiased causal estimates after the matching. The next two theorems explain that we can obtain unbiased causal estimates conditioning on a balancing score.

Definition 2.1.5. Treatment assignment is *strongly ignorable*, if the following is true:

$$r^{obs} \perp\!\!\!\perp z|x, \qquad 0 < P(z = 1|x) < 1$$

Theorem 2.1.6. *If treatment assignment is strongly ignorable given x, then it is strongly ignorable given any balancing score b(x); that is,*

$$r^{obs} \perp\!\!\!\perp z|x \text{ and } 0 < P(z = 1|x) < 1$$

for all x implies

$$r^{obs} \perp\!\!\!\perp z|b(x) \text{ and } 0 < P(z = 1|b(x)) < 1$$

for all b(x).

Proof. We can rephrase the above relationship as follows: $P(z = 1|r^{obs}, x) = P(z = 1|x)$ implies $P(z = 1|r^{obs}, b(x)) = P(z = 1|b(x))$. We will show this relationship using two initial assumptions. We have the given assumption

$$P(z = 1|r^{obs}, x) = P(z = 1|x)$$

Also, from Theorem 2.1.3, we can derive that

$$P(z = 1|b(x)) = e(x),$$

Now keeping in mind the two assumptions, we want to show $P(z = 1|r^{obs}, b(x)) = P(z = 1|b(x))$.

$$P(z = 1|r^{obs}, b(x)) = E[P(z = 1|r^{obs}, x)|r^{obs}, b(x)]$$

Using $r^{obs} \perp\!\!\!\perp z|x$,

$$\begin{aligned} &= E[P(z = 1|x)|r^{obs}, b(x)] \\ &= E[e(x)|r^{obs}, b(x)] \\ &= e(x) \\ &= P(z = 1|b(x)) \end{aligned}$$

□

Using Theorem 2.1.5, we can say that, if we can make treatment assignment independent from the outcome by conditioning on x, then we can create the same effect by conditioning on a reduced-dimension balancing score. Eventually, we want to say that we can measure an unbiased average treatment effect when conditioning on a balancing score, which the following theorem shows.

Theorem 2.1.7. *Suppose treatment assignment is strongly ignorable and b(x) is a balancing score. Then the expected difference in the observed responses to the two treatments at b(x) is equal to the average treatment effect at b(x), that is,*

$$E[r^{obs}_{z=1}|b(x), z = 1] - E[r^{obs}_{z=0}|b(x), z = 0] = E[r^{obs}_{z=1} - r^{obs}_{z=0}|b(x)].$$

Proof. Given strongly ignorable treatment assignment, that is, $r^{obs} \perp\!\!\!\perp z|x$, it follows from Theorem 2.1.5 that

$$\begin{aligned} E[r^{obs}_{z=1}|b(x), z = 1] - E[r^{obs}_{z=0}|b(x), z = 0] &= E[r^{obs}_{z=1}|b(x)] - E[r^{obs}_{z=0}|b(x)] \\ &= E[r^{obs}_{z=1} - r^{obs}_{z=0}|b(x)] \end{aligned}$$

□

The theorems are significant, because they set a theoretical basis on why matching on a balancing score helps to draw causal implications from non-experimental settings. Given the above theoretical basis, we can summarize a multi-dimensional vector of covariates into a balancing "score" with fewer dimensions. Conditioning on a balancing score generates the same effect as conditioning on all observed covariates, the effect being that the treated group and the non-treated group have covariate balance. In particular, the propensity score, $e(x)$, is a one dimensional balancing score, estimated by regressing all observed covariates on whether an individual was treated or not. The propensity score not only provides a way to match on one dimension instead of on multiple dimensions, but also helps to evaluate validity of a balancing score, because $b(x)$ is a balancing score if and only if the propensity score can be expressed as a function of $b(x)$.

2.2 Matching in a Prospective Setting

In a prospective study design, matching is used to minimize the covariate imbalance between the two study cohorts by pairing one treated individual with a demographically similar non-treated individual. As we have already shown, the propensity score $e(x)$ is a single-dimensional balancing score that can easily be used to match in prospective study designs. By Theorem 2.1.3 we propose a new balancing score on which to prospectively match: the *Prospective Exposure Case (PEC) Score*.

2.2.1 The Prospective Exposure Case Score

Definition 2.2.1. Given a treatment $t \in \{0,1\}$, an observed outcome $y \in \{0,1\}$, and covariates x, the Prospective Exposure Case Score is $b^{ty}_{pro}(x) = P(z = t, r^{obs} = y|x)$.

For any subject with covariates x, this metric takes on four dimensions, representing each of the four possible combinations of t and y. However, because these scores are probabilities, they have the constraint that $1 = b^{00}_{pro}(x) + b^{01}_{pro}(x) + b^{10}_{pro}(x) + b^{11}_{pro}(x)$, and the metric becomes a three-dimensional one in practice.

We conclude by demonstrating that the PEC Score is, in fact, a balancing score.

Theorem 2.2.2. *The PEC score $b^{ty}_{pro}(x)$ is a balancing score.*

Proof. Note that, by the law of total probability,

$$e(x) = P(z = 1|x) = P(z = 1, r^{obs} = 0|x) + P(z = 1, r^{obs} = 1|x) = b^{10}_{pro}(x) + b^{11}_{pro}(x)$$

So because there is a function relating the propensity score $e(x)$ and the PEC Score $b^{ty}_{pro}(x)$, the PEC Score is a balancing score. □

2.3 Matching in a Retrospective Setting

Matching in a retrospective setting is fundamentally different from matching in a prospective setting. Because individuals are sampled on the basis of their case/non-case status, both $z = t$ and $r^{obs} = y$ for a given individual i are conditional on that individual being sampled. Thus, finding a reduced-dimensional balancing score in the retrospective case is much harder.

Let s be a sampling indicator where

$$s = \begin{cases} 1 & \text{if retrospectively sampled} \\ 0 & \text{otherwise.} \end{cases}$$

If we were to regress all observed covariates against treatment assignment, as done when calculating the propensity score in the prospective case, the resulting probability would be conditional on both the covariates x *and* the fact that individual i was sampled, $s_i = 1$. In other words:

$$P(z = 1|x, s = 1)$$

This conditional probability is not equal to the propensity score in general; it also does not map to the propensity score. As such, this one-dimensional statistic is not a balancing score, and matching on it does not guarantee similar results to matching on all observed covariates x, and does not preserve treatment effect estimates. From this, we see that we cannot directly calculate the propensity score in the retrospective setting.

At the same time, very few methods for matching in a retrospective setting currently exist that both effectively reduce the dimensions of the matching problem and also take into account this conditioning on $s = 1$. In response to this methodological gap, we propose the *Retrospective Exposure Case (REC) Score* as a new balancing score suitable for retrospective causal inference.

2.3.1 The Retrospective Exposure Case Score

We define the REC score analogously to the PEC score in the prospective case, in that it is the joint probability of receiving treatment $z = t$ and displaying outcome $r^{obs} = y$, conditioned on the covariates. However, the REC score also conditions on being retrospectively sampled.

Definition 2.3.1. Given a treatment $t \in \{0, 1\}$, an observed outcome $y \in \{0, 1\}$, and covariates x, the Retrospective Exposure Case Score is $b^{ty}_{ret}(x) = P(z = t, r^{obs} = y|x, s = 1) = (b^{00}_{ret}(x), b^{10}_{ret}(x), b^{01}_{ret}(x), b^{11}_{ret}(x))$.

Theorem 2.3.2. *The REC score $b^{ty}_{ret}(x)$ is a balancing score.*

Proof. To show that b^{ty}_{ret} is a balancing score, it is sufficient to show that $e(x) = f(b^{ty}_{ret})$ for some function f.

Because retrospective sampling relies entirely on case status, we reasonably assume that $s \perp\!\!\!\perp x, z|r^{obs}$. Let $p_1 = P(s = 1|r^{obs} = 1) = P(s = 1|r^{obs} = 1, x, z = t)$, and $p_0 = P(s = 1|r^{obs} = 0) = P(s = 1|r^{obs} = 0, x, z = t)$.

Then

$$b^{ty}_{ret}(x) = P(z = t, r^{obs} = y|x, s = 1) \propto P(s = 1|z = t, r^{obs} = y, x)P(z = t, r^{obs} = y|x) = p_y b^{ty}_{pro}(x)$$

Using Bayes' Theorem, we find the particular relationship between the REC score and the PEC score is given by:

$$b^{ty}_{ret}(x) = \frac{p_y b^{ty}_{pro}(x)}{p_0 b^{00}_{pro}(x) + p_1 b^{01}_{pro}(x) + p_0 b^{10}_{pro}(x) + p_1 b^{11}_{pro}(x)}$$

Having defined the REC score in terms of the PEC score, we now invert it to give the PEC score in terms of the REC score, enabling us to conclude the REC score is a balancing score. So from the above equation, we derive

$$b^{ty}_{pro}(x) = \frac{p_{1-y} b^{ty}_{ret}(x)}{p_1 b^{00}_{ret}(x) + p_0 b^{01}_{ret}(x) + p_1 b^{10}_{ret}(x) + p_0 b^{11}_{ret}(x)}$$

So the PEC score $b^{ty}_{pro}(x)$ can be written as a function of the REC score $b^{ty}_{ret}(x)$. Let $b^{ty}_{pro}(x) = g(b^{ty}_{ret}(x))$ for the above function, g. Since $e(x) = f(b^{ty}_{pro}(x))$ for some function f, it follows that $e(x) = f(g(b^{ty}_{ret}(x)))$. So the REC score is a balancing score. □

Since the REC score is a balancing score, treatment assignment is independent of the observed covariates x given $b^{ty}_{ret}(x)$. Note that the REC score, like the aforementioned PEC score, is a three-dimensional metric, which has potential to greatly reduce the complexity and computational demands of the matching process. Thus, matching a treated individual to a non-treated individual on the basis of the REC score helps correct for any covariate imbalance and allows for the more efficient estimation of causal effects in retrospective studies.

3. Application: Data & Results

In order to determine the degree to which matching i) accounts for any existing covariate imbalance and ii) allows for accurate estimation of causal effects, we draw on the previous work of statisticians Dehejia and Wahba. Dehejia and Wahba used a composite data set including both experimental and survey data to examine how well prospective analytical approaches could recover experimental estimates of causal relationships [1]. Here, we extend their results, simulating both prospective *and* retrospective studies from this composite data set. Using these simulated studies, we examine to what degree prospective matching—using the propensity score $e(x)$—and retrospective matching—using the REC score $b^{ty}_{rec}(x)$—increase our ability to accurately quantify a causal effect in non-randomized experimental settings.

3.1 Data

In the mid-1970s, the National Supported Work (NSW) Demonstration job training program was implemented, providing work experience to individuals who faced economic or social barriers to finding work.[4] A total of 445 individuals enrolled in an experiment to determine the effect of this job training program on post-intervention wages. 185 of those individuals were randomly selected to receive the job training, and all participants were asked in 1978, four years after the experiment, if they had found a job and, if so, how much that job paid. The experimenters also collected information on the following pre-intervention covariates: age, number of years of school completed, race, marital status, degree status, wages in 1974, and wages in 1975. For our purposes, we converted all recorded wages into binary employment indicators: any subject with a reported wage of \$0 was considered unemployed; all other subjects were considered to be—in some capacity—employed.[1]

In order to later simulate both prospective and retrospective studies, we combined this experimental data with a contemporaneous non-experimental cohort from the Current Population Survey (CPS). This survey was administered jointly by the Census Bureau and the Bureau of Labor Statistics, and provides a substantial cohort entirely of non-treated individuals ($n = 15992$) with the same recorded covariates as in the NSW experiment.

Throughout our analysis, we considered the treatment of interest to be the NSW job training program: individuals who participated in the program are considered treated ($z = 1$) and those who did not participate are considered non-treated ($z = 0$). We considered as cases any individuals who were not employed in 1978 ($r^{obs} = 1$). All employed individuals were considered non-cases ($r^{obs} = 0$).

3.2 Randomized Experimental Results

We used both the naive estimation process described in section 1.3.1 and the regression analysis described in 1.3.3 to analyze the NSW experiment. We found that someone who took the job training was 41% less likely to be unemployed than an individual who did receive job training. That value changed to 42% when controlling for other covariates using logistic regression, as seen in Table 3.1 We used these results as a metric by which to evaluate the unmatched and matched estimation methods in both the prospective and retrospective settings. In particular, we were able to quantify "how well" each of these estimation practices adjusted for any existing covariate imbalance and "recovered" this known causal effect.

[1] We let *ed* be the number of reported years of education, *nodeg* be an indicator of whether an individual received ($nodeg = 0$) or did not receive ($nodeg = 1$) a high school degree, $re74$ and $re75$ be the recorded wages in 1974 and 1975, respectively, and $u74$ and $u75$ be employment indicators for those same years.

3.3 Prospective and Retrospective Studies: Unmatched Estimation

3.3.1 Simulation of Prospective and Retrospective Studies

In order to evaluate the performances of the estimation techniques for prospective and retrospective studies, we combined the treated observations from the NSW dataset with the CPS dataset to *simulate* both prospective and retrospective sampling techniques.

In simulating a prospective study, we simply combined the NSW and CPS datasets, pretending that they had been sampled from the same sampling frame and that the job training treatment was actually an observed variable. The two datasets were taken from the same time period and drew from the same subject demographic (middle-aged men), which makes it a plausible assertion that the subjects in each dataset could have come from the same sampling frame. Thus, we could reasonably perform prospective study analysis on this combined dataset to get estimates of the effect of the treatment on unemployment in an observational context.

To simulate a retrospective study, we treated the combined NSW and CPS dataset as a sampling frame of subjects from which we sampled cases (unemployed individuals) and non-cases (employed individuals). Specifically, we sampled all of the cases, and then sampled a number of non-cases such that the ratio of cases to non-cases in our sample was between 1:1 and 1:5. Unlike the simulated prospective study, where the entire combined dataset was considered to represent an observed sample, in this simulation there was randomness in the sampling of non-cases. Thus, when we performed retrospective inference, we took 1000 samples and averaged our estimates across the iterations.

3.3.2 Covariate Imbalance

To assess covariate imbalance in each study design, we compared the absolute standardized differences in mean recorded covariate values between the treated and non-treated groups. For binary variables, we calculated this standardized absolute difference using a z-statistic; for continuous variables, we used a t-statistic. For the retrospective study design estimates, we calculated these statistics across 30 samples, again seeking to account for the randomness in the sampling of the non-cases. As such, the covariate imbalance estimates were a single point in the experimental and prospective cases, and a distribution of points in the retrospective cases. The results were then presented on a Love plot. Any standardized differences smaller than 1.96 indicated inconsequential covariate imbalance; differences larger than that amount constituted covariate imbalance and presented a challenge to causal effects estimation.

3.3.3 Unmatched Estimation Results

In the experimental NSW data, the randomized treatment assignment mechanism created relative balance across all observed covariates (Figure 3.1). However, without matching, both the prospective and retrospective datasets displayed significant covariate imbalance, especially in comparison to the experimental dataset.

Using the same estimation techniques as for the randomized experiment, we found that, in both the prospective and retrospective settings, the unmatched analysis was insufficient to remove the bias created by the covariate imbalance in these study designs. Table 3.1 shows a comparison of the prospective results to the experimental benchmarks, and demonstrates this bias; the naive plug-in approach estimated that job training actually *increased* the chances of being unemployed. However, this result directly reflects the covariate imbalance we know to exist in the observational datasets: those who took job-training programs generally did so because they were otherwise less likely to find gainful employment in the workforce. As such, the subjects who received the treatment tended to be those who were less likely to become employed, i.e., they previously earned less or were previously unemployed, relative to the general population in the CPS survey who were not treated.

The prospective study estimation bias was reduced with the logistic regression approach, which found that job training reduced the odds of being unemployed by 13%. However, even this finding incorrectly estimated the experimentally-determined value. Thus, regression alone did not resolve the problem of covariate imbalance.

The retrospective approach yielded similar results. Table 3.2 also reflects the extreme bias of naive unadjusted ("plug-in") estimation, and once again logistic regression was not sufficient to remove that bias.

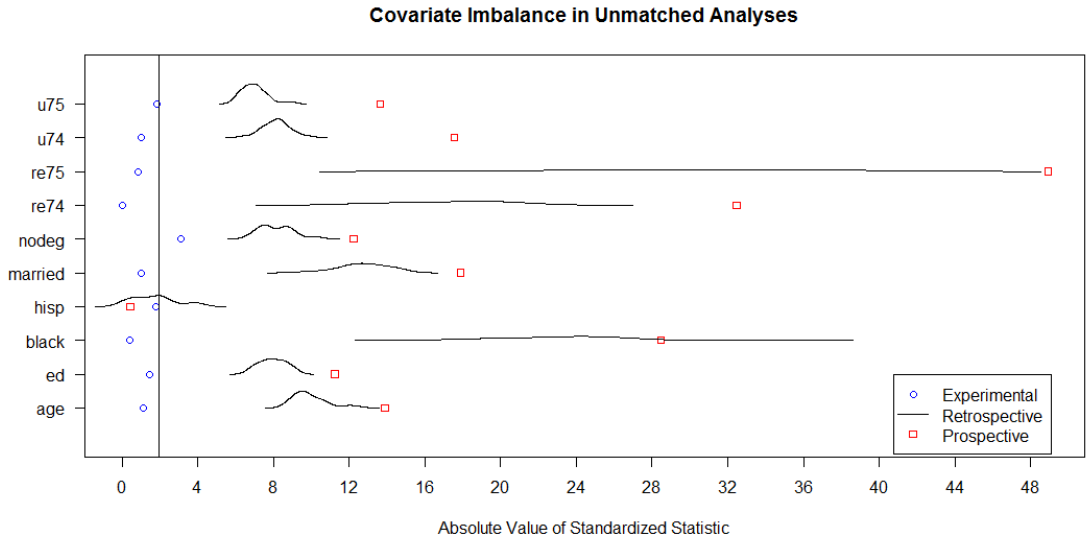

Figure 3.1: Love plot depicting covariate imbalance in unmatched analyses.

Study	Naive Plug-in	Logistic Regression
Experimental	0.59	0.58
Prospective	2.04	0.87

Table 3.1: Prospective Study Treated v. Non-Treated, unmatched estimation of odds ratio γ for unemployment indicator

Also note that as the retrospective sample increased in size (i.e., the number of non-cases sampled per case increased) the estimates approached the prospective estimates, which were calculated from the full dataset.

Study	Naive Plug-in	Logistic Regression
Experimental	0.59	0.58
Retrospective 1:1	2.04	1.15
Retrospective 1:2	2.04	1.06
Retrospective 1:3	2.04	0.99
Retrospective 1:4	2.04	0.94
Retrospective 1:5	2.04	0.90

Table 3.2: Treated v. Non-Treated, unmatched estimation of odds ratio γ for unemployment indicator, averaged over 1000 simulations

3.4 Prospective and Retrospective Studies: Matched Estimation

Given that the unmatched analyses were insufficient in adjusting for covariate imbalance in either the prospective or the retrospective setting, we next examined to what extent matching addressed this issue. From a theoretical perspective, matching treated individuals to similar non-treated individuals should address the underlying covariate imbalance issue for both the prospective and retrospective cases, and thus should improve on unmatched estimation of causal effects. We thus applied common matched analysis techniques to both the simulated prospective and retrospective NSW/CPS datasets to determine whether matching is also an effective causal inference tool in practice.

3.4.1 Implementation of Matching Procedures

We took the estimated propensity score $e(\widehat{x})$ as the balancing score of interest for the prospective setting, and the estimated REC Score as the balancing score of interest for the retrospective setting. We then used these metrics to match treated individuals to similar non-treated individuals.

3.4.1.1 Prospective Setting

In our simulated prospective study, the binary balancing score was computed using a logistic regression model.[2] We then used the R package "Matching" to match a single treated unit to the single non-treated unit on the basis of this propensity score.

3.4.1.2 Retrospective Setting

To estimate the multidimensional REC Score, we used a regression technique called *multinomial regression*.[3] This model generates four predicted probabilities for each subject: one corresponding to each combination of treatment and case statuses. We then used these probabilities to estimate the REC Scores for each subject (see Section 2.3.1).

Matching based on a multidimensional score such as the REC Score requires a different approach than matching on a singular metric, such as the propensity score. In particular, we employed a technique called *coarsened exact matching* using the R package "cem". Coarsened exact matching does not strictly match 1-to-1, but instead assigns subjects to strata, where within each stratum the covariates are considered balanced. This assignment process works by binning each covariate, and then considering all subjects that fall into the same covariate bins as belonging to the same stratum.

For the purposes of matching, any strata containing only a single individual were discarded, as were all strata containing only "treated" individuals or only "non-treated" individuals. We then considered all treated and non-treated individuals in the same stratum to be matched to one another.

3.4.2 Estimation Techniques using Matched Data

3.4.2.1 Off-Diagonal Matched Analysis

For the matched samples in which one treated individual was matched to exactly one non-treated individual, the odds ratio was estimated using a discordant pairs approach [2]. In the prospective setting, we considered the discordant pairs to be the set of paired treated and non-treated individuals with different observed outcomes. For example, in Table 3.3, there are $u + v$ discordant pairs. Note that u is the number of pairs composed of a non-treated non-case individual matched to a treated case individual. Similarly, v is the number of pairs composed of a non-treated case individual matched to a treated non-case individual.

		Non-Treated ($z = 0$) Case ($r^{obs} = 1$)	Non-Treated ($z = 0$) Non-case ($r^{obs} = 0$)
Treated ($z = 1$)	Case ($r^{obs} = 1$)	t	u
Treated ($z = 1$)	Non-Case ($r^{obs} = 0$)	v	w

Table 3.3: Display of Matched Pairs in Prospective Data

In the retrospective setting, we considered the discordant pairs to be the set of paired case and non-case individuals with different treatment assignments. In Table 3.4, u is now the number of matched pairs with a

[2] The specifications for the model are given by

$$P(z = 1) = f(age, age^2, ed, ed^2, married, nodeg, black, hisp, re74, re75, u74, u75, ed * re74, age^3) \quad (3.1)$$

.

[3] In particular, we used identical covariate specifications to our model for propensity score:

$$P(z = t, r^{obs} = y) = f(age, age^2, ed, ed^2, married, nodeg, black, hisp, re74, re75, u74, u75, age^3, ed * re74) \quad (3.2)$$

treated case individual and a non-treated non-case individual, while v is now the number of matched pairs with a non-treated case individual and a treated non-case individual.

		Non-Case($r^{obs} = 0$) Treated ($z = 1$)	Non-Case ($r^{obs} = 0$) Non-Treated ($z = 0$)
Case ($r^{obs} = 1$)	Treated ($z = 1$)	t	u
Case ($r^{obs} = 1$)	Non-Treated ($z = 0$)	v	w

Table 3.4: Display of Matched Pairs in Retrospective Data

In both the prospective and retrospective settings, we estimated the odds ratio by the ratio u/v.

3.4.2.2 Conditional Logistic Regression Analysis

In order to implement this off-diagonal analysis, we utilized a method called *conditional logistic regression*. Conditional logistic regression is quite similar to logistic regression but with one major difference: the model fitting process incorporates information about stratification of the subjects, and thus leverages the matching process to improve the effect estimate. This estimation technique is suited both to 1-to-1 matching—where each pair is considered its own stratum—as well as more general stratification—such as we employed when matching on the REC score. Moreover, this approach turns out to replicate the off-diagonal analysis when it is fit using only strata information.

3.4.3 Matched Analysis Results

In both the prospective and retrospective setting, matching treated individuals with non-treated individuals on the basis of a balancing score substantially reduced covariate imbalance. This reduction was particularly noticeable in the prospective dataset; after matching on the propensity score, only one covariate (years of education) remained significantly imbalanced between the treated and non-treated groups (Figure 3.2).

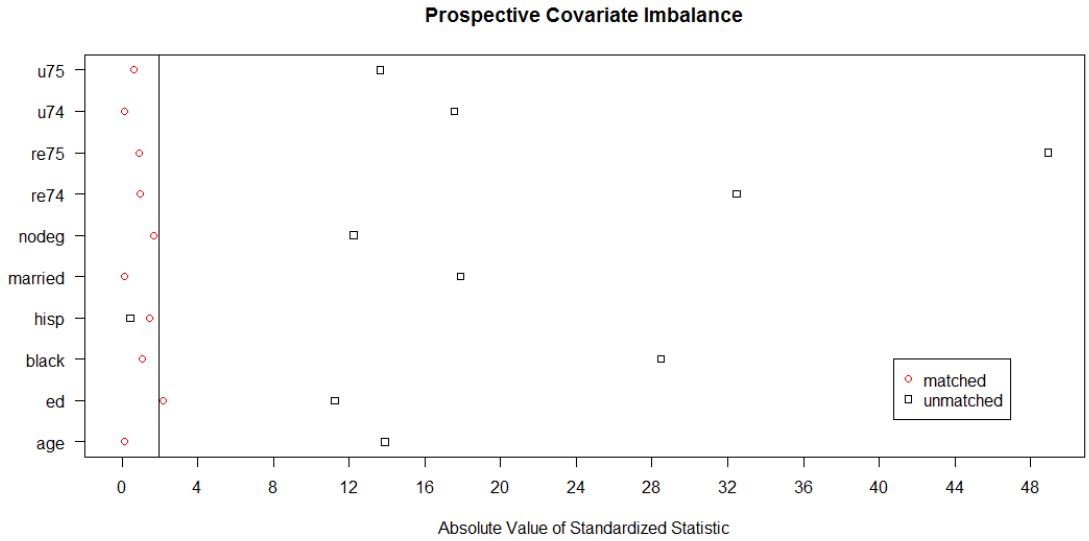

Figure 3.2: Love plot comparing covariate imbalance in the prospective matched and unmatched datasets

In the retrospective setting, matching on the REC Score also reduced covariate imbalance, though this reduction was much less pronounced than in the prospective case (Figure 3.3). All but two covariates (namely the indicator for whether an individual was Hispanic and age) remained significantly imbalanced between the treated and non-treated groups; one of those two "balanced" covariates (the Hispanic indicator) had already

been approximately balanced in the unmatched retrospective sample. In several instances, namely for the 1974 and 1975 unemployment indicators, the matched retrospective samples displayed a greater magnitude of covariate imbalance than the unmatched samples. That being said, the absolute standardized difference between the mean covariate values of the treated and non-treated groups did generally decrease after matching on the REC Score, indicating that matching did—in part—address issues of covariate imbalance.

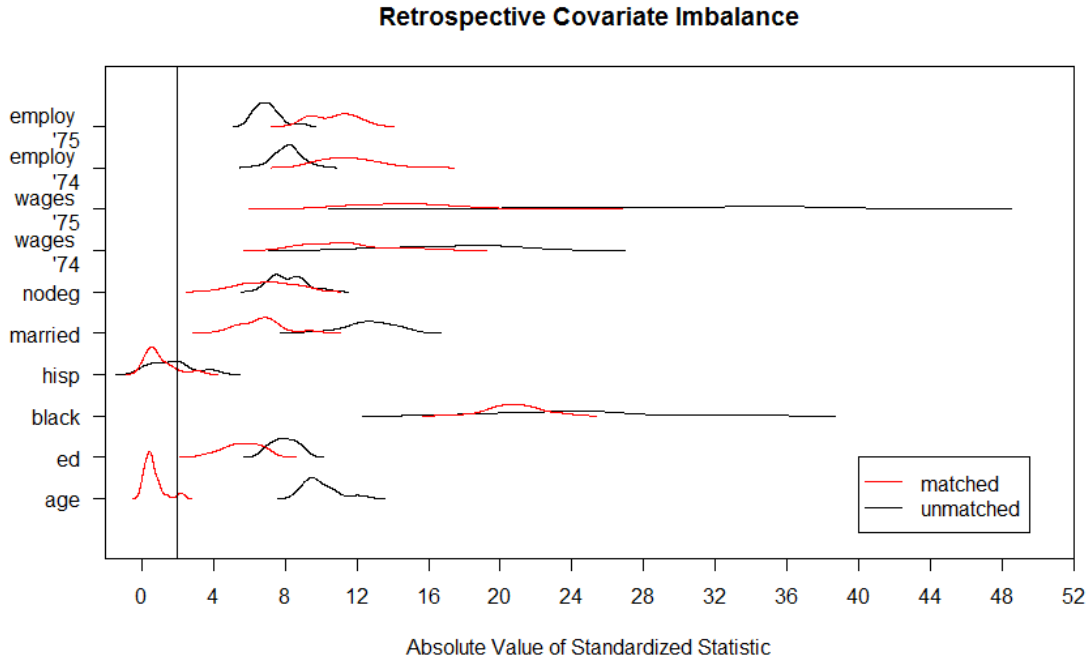

Figure 3.3: Love Plot comparing covariate imbalance in the retrospective matched and unmatched datasets. All retrospective samples represented above consisted initially of five non-cases per case. Distribution over 30 simulations.

Both the prospective and retrospective odds ratio estimates reflected this reduction in covariate imbalance (Table 3.5, Table 3.6). In the prospective setting, we found that job training reduced the odds of being unemployed by 34% (Table 3.5). Though this result still underestimated the actual experimental result—that job training reduced the odds of unemployment by 42%—it vastly improved upon the unmatched prospective estimates.

Study	Conditional Logistic Regression
Experimental	0.58
Prospective	0.66

Table 3.5: Estimation of unemployment odds ratio γ, Prospective Study Treated v. Non-Treated, Matched analysis

Although matching on the REC Score did not fully address issues of covariate imbalance in the retrospective samples, the retrospective odds ratio estimates very nearly recovered the experimental value (Table 3.6). The smallest retrospective sample—with one non-case per case—yielded an odds ratio estimate of 0.94, which was similar to the odds ratio found in the "best-performing" unmatched retrospective sample (see Table 3.2). As in the unmatched case, the matched retrospective odds ratio estimate approached the experimental value as the size of the retrospective sample increased. In the matched retrospective sample with five non-cases per case, we estimated that job training reduced the odds of unemployment by 40%, an estimate that was within one percentage point of the experimental estimate.

Study	Conditional Logistic Regression
Experimental	0.58
Retrospective 1:1	0.94
Retrospective 1:2	0.80
Retrospective 1:3	0.71
Retrospective 1:4	0.68
Retrospective 1:5	0.60

Table 3.6: Estimation of unemployment odds ratio γ, Retrospective Study Treated v. Non-treated, Matched analysis

4. Conclusion

4.1 Primary Findings

Having demonstrated the improvement that matching on a balancing score in the prospective setting produces in Table 3.5, we established the REC score, a balancing score for the retrospective study setting. We designed this score specifically to overcome the main challenge that in the retrospective setting, covariates of subjects are unknown until they are actually sampled. Thus, the REC score incorporates the probabilities of being retrospectively sampled as a case and as a non-case, enabling it to serve as a balancing score in the retrospective setting.

In our simulations, we found that retrospective matching using the REC score successfully produces causal estimates similar to those from a randomized experiment, particularly as the retrospective sample size increases (Table 3.6). Ultimately, we found that by sampling with a ratio of 1:5 case to non-cases, we saw an average estimated odds ratio of 0.6, very near to the experimentally determined value of 0.58. This estimate even bests the prospective matched estimate of 0.66 (Table 3.5).

However, we also find that, despite coming very close to recovering the experimental estimate, our retrospective matching process did not eliminate the problem of covariate imbalance. As demonstrated in Figure 3.3, while matching did decrease covariate imbalance, it did not eliminate it. This comes as a particular surprise not only because we still observed a substantially improved estimate, but also because our matching in the prospective case dramatically eliminated covariate imbalance between the treated and non-treated groups (Table 3.2).

The exact reason for the decoupling of unbiased causal estimation and covariate balance—at least in the retrospective case—is not entirely clear, and deserves further consideration and empirical study. One hypothesis locates the issue specifically in the type of analysis we performed, and the way we stratified based on the REC score. In particular, our estimation procedures all leveraged the concept of off-diagonal comparison—that is, comparison of matched pairs of treated and non-treated individuals with differing observed outcomes. Because of this analysis, the individuals who were not in those categories did not affect the estimation, even though they were part of the samples and thus could contribute to covariate imbalance. Finally, because our retrospective matching approach was not 1-to-1, but rather stratified, it follows that this may have let untreated individuals who did not affect the casual estimate remained in the matched samples by virtue of their presence in the same strata as treated individuals, and thus the covariate balance may have been affected. This is merely a speculation, made particularly limited by our inability to test the performance of the REC score on any other data. Thus, we merely present the results as we observed them, and remain cautiously optimistic about the ability of REC score matching to recover experimental causal estimates.

4.2 Limitations and Future Considerations

From reading this paper, one might conclude that matching is uniformly more convenient and cheaper than randomized experiments and just as effective—so why bother with experiments at all? However, we have yet to touch on the most critical shortcoming of matching techniques in general, which persists even as we continue to improve and extend matching procedures and techniques: matching can only be done on observed and measured covariates. This means that researchers have to identify the covariates that they assume are serving to confound the relationship between treatment and outcome, and measure those covariates with enough accuracy to enable matching. This becomes a particular challenge in the retrospective context because it often involves mining historical data collected about each individual's conditions and status at times before the outcome was observed, where data is likely to be more sparse. Compare this situation to the randomized experiment, where all characteristics of the subjects, whether observed or unobserved, observable or unobservable, are independent of the random treatment assignment mechanism. This produces balance not just on the measured covariates, but on all possible personal characteristics. In sum, matching works to

balance measured covariates, and can be quite effective if the covariates chosen explain most of the estimation bias, but also do not completely balance the subjects on all possible confounding factors.

Another possible limitation of matching is that it may create little or no effect on covariate imbalance in a small data set. This happens when the size of the non-treatment group is small, making it harder to find a non-treated individual who looks similar to the treated individual. In other words, as the pool to find a matching non-treated unit is smaller, it is harder to find a "good match" to the treated unit. This limitation becomes more critical as the size of the non-treated sample is smaller and as initial covariate imbalance is bigger. This also provides insight into why, in our simulation of retrospective matching, a ratio of 1 case per 5 non-cases produced a better result than a ratio of 1 case to 1 non-case. It may be that, as the sample progressively included a larger number of non-cases, we had a larger pool of both treated and non-treated individuals from which to find good matches.

4.3 Impact on the Statistical Community

While the community has been searching for a while, the concept of producing causal results in a non-randomized setting is especially tantalizing today as data is gathered faster than ever before. The theory that we could find causal results without running an experiment would allow anyone to collect a sample of people with some specified treatment and outcome and as long as you match the cases and controls. Causal results would be churned out faster than ever before. Wouldn't we all like to know exactly which of the hundreds of possible carcinogenic products are actually giving us cancer?

4.4 Acknowledgments

The authors would like to acknowledge Professor Dave Watson for his guidance throughout the project. We also would like to thank professors in Carleton College's Mathematics and Statistics Department for their valuable teachings and discussions, and all those who have mentored us on our statistical journeys.

Bibliography

[1] Rajeev H Dehejia and Sadek Wahba. "Causal effects in nonexperimental studies: Reevaluating the evaluation of training programs." In: *Journal of the American statistical Association* 94.448 (1999), pp. 1053–1062.

[2] B. Bert Gertman. *Case-Control Studies: Odds Ratios*. 2006. URL: `http://www.sjsu.edu/faculty/gerstman/StatPrimer/case-control.pdf`.

[3] Guido W Imbens and Donald B Rubin. *Causal Inference in Statistics, Social, and Biomedical Sciences: An Introduction*. Cambridge University Press, 2015.

[4] Robert J LaLonde. "Evaluating the econometric evaluations of training programs with experimental data." In: *The American Economic Review* (1986), pp. 604–620.

[5] Paul R Rosenbaum and Donald B Rubin. "The central role of the propensity score in observational studies for causal effects." In: *Biometrika* 70.1 (1983), pp. 41–55.