Assessing the Ability of Matching to Address Attrition Bias in a Randomized Experiment using Data from the Rand Health Insurance Experiment*

John C. Ham

University of Maryland,

IFAU, IRP (Madison) and IZA

Xianghong Li

York University

Revised May 2011

JEL Classification: C52, C31, C14; I11

Keywords: Non-random attrition; Propensity score matching; Rand Health Insurance Experiment

* We are grateful for very helpful comments received at several seminars, as well as those we received from Miguel Almanzar and Jeff Smith. As is often the case in our work, Geert Ridder provided us with substantial help at several stages of the research. This research has been supported by grants from the National Science Foundation (SES-0136928 and SES-0627968). Any opinions, findings and conclusions or recommendations in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation. We are responsible for any errors. Contact information: *John Ham (corresponding author)*, ham@econ.umd.edu, Department of Economics, 3105 Tydings Hall, University of Maryland, College Park, MD 20742, phone: (202) 380-8806, fax: (301) 405-3542; *Xianghong Li,* xli@econ.yorku.ca, Department of Economics, York University, 4700 Keele Street, Vari Hall 1068, Toronto, ON M3J 1P3, Canada, phone: (416) 736-2100 ext. 77036, fax: (416)736-5987.

ABSTRACT

It is well known that non-random attrition can lead to bias in estimating treatment effects from a social experiment that is based on random assignment. If the randomized intervention suffers from non-random attrition, the intent to treat (ITT) estimator is biased and the IV estimator of the treatment effect is also inconsistent (Frangakis and Rubin 1999). DiNardo, McCrary and Sanbonmatsu (2006) propose an approach to correct attrition bias, but their approach requires an additional randomization.

In principle, propensity score matching can eliminate or mitigate the bias due to non-random attrition. This study investigates how well matching achieves this goal after we introduce a plausible form of non-random attrition in the well-known Rand Health Insurance Experiment (RHIE). Specifically, we introduce attrition in a less generous insurance plan sample at the end of year 1 of the experiment on the basis of high health care expenditures in year 2. We then assess whether matching can eliminate or mitigate the bias in estimating the treatment effect of switching plans by comparing our matching estimates to the experimental results. Since the data on pre-experimental health care expenditures is self-reported and unreliable, we use the individuals' year 1 expenditure rank within their respective plans as one of our conditioning variables. While this is a post-experiment (treatment) variable, its use will be valid if an individual's health expenditure ranking within a plan is not affected by health insurance plan assignment, i.e. the individuals are rank-order stable across the plans. We find that when we use year 1 expenditure rank as a conditioning variable and the outcome variable is defined as chronic condition expenditures, matching eliminates about half of the attrition bias. However, this is not the case when the outcome variable is defined as the total health care expenditures. Finally, without conditioning on year 1 rank, matching cannot mitigate attrition bias for either outcome variable.

1. Introduction

Social experiments are considered the gold standard in terms of estimating causal effects. However, these experiments can encounter problems in at least two dimensions: non-compliance and non-random attrition, both of which may require the use of nonexperimental methods. For example, in terms of non-compliance, some of the controls in the Job Training Partnership Act experiment obtained the training by other means, while some of the treatments did not undertake the training. In this case one can estimate the intent-to-treat (ITT) parameter or use the random assignment indicator as an instrument (or exclusion restriction) in an IV procedure for estimating the effect of training participation.[1]

Non-random attrition has long attracted the attention of statisticians and economists (e.g. Hausman and Wise 1979, Barnard, Du, Hill and Rubin 1998 and Heckman, LaLonde and Smith 1999), but general solutions to the problem have received less attention. Frangakis and Rubin (1999) demonstrate that if the randomized intervention suffers from non-ignorable attrition where subjects leave the study in a non-random manner, the ITT estimator is biased and the IV estimator of the treatment effect is also inconsistent. In principle, propensity score matching offers the potential to reduce or eliminate the bias in treatment effect estimates when the social experiment has a rich set of baseline variables, but matching appears to have been used sparingly in this context.[2]

Our goal is to assess the efficacy of propensity score matching in addressing bias in estimated treatment effects due to non-random attrition in a social experiment setting.[3] We

[1] See, e.g., the discussion in Eberwein, Ham and LaLonde (1997), who estimate both the ITT effect, and the effect of actually participating training, on the labor market histories of disadvantaged women.

[2] Ding and Lehrer (2010) apply a dynamic matching model proposed by Robins (1986) to a multi-period experiment contaminated by selective attrition to recover dynamic treatment effects. Because the social experiment they analyze (Project STAR) was contaminated by attrition, their evidence cannot be used to asses how well the dynamic matching model performs in correcting for biases in a "broken experiment."

[3] One can also use a control function approach. Hausman and Wise (1979) present an MLE version of such an approach. For a comparison of matching and control function estimates in addressing possible bias in a

introduce a plausible form of non-random attrition in the well-known Rand Health Insurance Experiment (RHIE) data. Specifically, we artificially induce non-random attrition at the end of year 1 post-baseline in one health insurance plan sample, and then investigate how well matching deals with the bias when the year 2 expenditures are the outcome variables of interest. Since the data come from a social experiment, we can compare our matching estimates to the true effects provided by the experimental data. We focus on two RHIE insurance plans: the 95 percent co-payment plan (the least generous fee-for-service plan in RHIE, hereafter the 95% plan) and the no co-payment plan (the most generous fee-for-service plan in RHIE, hereafter the free plan). Participants in both plans are randomly assigned from a representative (random) sample of families in six US cities. Our artificially introduced attrition takes the form of randomly excluding 50% of the individuals with *year 2* expenditures above the median in the 95% plan group (i.e. a quarter of the 95% plan group is deleted in year 2). We do not introduce any attrition in the free plan group. This is an economically plausible form of attrition: individuals in the less generous plan with potential high health care expenditures are more likely to leave in the absence of side payments.[4]

We define the treatment as being in the 95% plan. Thus, our treatment group consists of individuals in the 95% plan (after our induced attrition) and our comparison group consists of individuals in the free plan. We consider two treatment effects: (i) the treatment effect of moving from the 95% plan to the free plan for the individuals who "chose" to stay in the 95% plan (after attrition), i.e. the average treatment effect on the treated (ATET); and (ii) the treatment effect of moving from the 95% plan to the free plan

---

different context, see Blundell, Dearden and Sianesi (2005). Note that one can use Stata to estimate both types of models; in terms of using Stata to estimate matching models, see Sianesi (2010).

[4] During the first two years, there was actually very little attrition in either the treatment or comparison group because the RHIE employed side payments (called Participation Incentive payments) to ensure that families were equally likely to participate in their plans regardless of plan generosity. Thus, we have little need to worry about actual attrition in the RHIE.

for the underlying RHIE population, i.e. the average treatment effect (ATE). Since matching may perform differently in addressing the bias due to non-random attrition for different components of health expenditures, we consider two outcome variables in year 2: (i) health care expenditures on chronic conditions only and (ii) total health care expenditures.

Finally, we investigate a modification of standard matching procedures that seems particularly useful in minimizing bias in estimated treatment effects when using health expenditure data, but should also have other applications. Appropriate conditioning variables in matching should not be affected by the treatment; thus, generally one only uses pre-treatment variables. A natural conditioning variable in the RHIE is the pre-experimental medical expenditures; however, this variable is self-reported and has little correlation with post-treatment expenditures.[5] As a result, we are unlikely to be able to achieve the Ignorable Treatment Assignment (ITA, Rosenbaum and Rubin 1983) assumption that underlies matching, and indeed we find that matching is ineffective in eliminating bias when we use the pre-baseline expenditures as a conditioning variable. To address this problem, we use the individual's year 1 expenditure (within their respective plan) *rank* as one of our conditioning variables in the propensity score. While this is a post-experiment (treatment) variable, its use will be valid if the individual's ranking *within* a plan is not affected by the plan's generosity, i.e. the individuals are rank-order stable across the plans. Given this assumption, individuals who have relatively high expenditures in year 1 in the free plan are comparable to those who have relatively high expenditures in year 1 in the 95% plan.[6] Because we have experimental data, we can investigate whether matching is more successful with this conditioning variable.

[5] A recent review of 42 studies evaluating the accuracy of household-reported health utilization data found that the most common problem is underreporting (Bhandari and Wagner 2006). Of course, if the under-reporting were constant across individuals, it would not create a problem for matching. Our results suggest that it is not constant across individuals.

[6] Rank order stability plays a role in other areas of applied research. For example, see Borghans, Duckworth, Heckman and ter Weel (2008) for its use when discussing the stability of personality traits over time.

We find that using this variable improves matching's ability to eliminate bias from non-random attrition when chronic condition expenditures is the outcome variable but not when the outcome variable is chosen to be the total health expenditures.

The RHIE was conducted from 1974 to 1982 at six sites across the United States. Since health costs differ across cities, we also consider a finer balancing procedure (Rosenbaum and Rubin 1985) where we match an observation only with individuals who live in the same city. Our use of finer balancing is motivated by the results of Heckman, Ichimura and Todd (1997). They show that matching individuals operating in the same local labor market improves the performance of matching estimators in reducing biases due to non-random selection when evaluating the effect of JTPA training. We consider both kernel regression matching (KRM) and local linear regression matching (LLRM) because LLRM has better asymptotic properties (Fan and Gijbels 1996), but poorer small sample properties (Frölich 2004), than KRM.

The rest of the paper proceeds as follows. Section 2 contains our literature review. It first provides the background for the RHIE and describes the data we use from the experiment. Next it discusses previous work using experimental data to assess the efficacy of different approaches to minimizing bias due to sample selection since non-random attrition leads to one form of sample selection. Finally, it reviews other approaches in the literature specifically dealing with attrition. Section 3 describes our method of introducing non-random attrition into the RHIE data and describes, in some detail, our matching procedures. We present our empirical results in Section 4. When we condition on year 1 rank, we find that matching performs reasonably well in terms of reducing bias in the estimated treatment effects for chronic condition expenditures but not for total health care expenditures. Without year 1 rank as a conditioning variable, we find that matching is ineffective in

reducing biases in treatment effects for either type of health care expenditure. Somewhat surprisingly, using finer balancing does not affect the ability of matching to reduce bias. Further, matching does a slightly better job of reducing bias when estimating the ATET than when estimating the ATE. Section 5 concludes the paper.

2. The Rand Health Insurance Experiment and Literature Review

2.1 The Rand Health Insurance Experiment

The RHIE was conducted from 1974 to 1982 at six sites across the United States: Dayton, Ohio; Seattle, Washington; Fitchburg and Franklin counties, Massachusetts; and Charleston and Georgetown counties, South Carolina. In each site, the experiment lasted for 3 or 5 years. The experiment was designed primarily to assess how different rules for sharing health care costs between individuals and health insurance companies would affect utilization of services, the quality of health care provided, patient satisfaction, and subject health status.

The experimental samples (with the family as the unit of sampling) were randomly chosen from each site subject to eligibility criteria. The participating families then were randomly assigned to one of up to eighteen experimental fee-for-service (FFS) health insurance plans (distinguished by coinsurance rates and maximum out-of-pocket expenditures) in each location, and one HMO plan in Seattle.[7] We consider only the FFS members of the experiment and exclude the HMO enrollees for three reasons. First, the FFS participants were randomly sampled from each city's population, while the HMO participants were enrollees of the Group Health Cooperative of Puget Sound, an HMO in Seattle, before the experiment. Second, due to the different modes of health care delivery,

[7] See Health Insurance Experiment: Master Sample Series Volume 1 by Rand Corporation for details.

the total health care expenditures from HMO and FFS plans are not directly comparable. Third, for individuals participating in the FFS plans, the experiment provides their total health expenditures and a breakdown of expenditures into five categories: acute, chronic, chronic flare-up, well-being, and pregnancy, but this breakdown is not available for the HMO sample in Seattle. As noted above, in our analysis we focus on chronic condition health expenditures and total health care expenditures.

To simplify the analysis, we restrict the sample to adults only and exclude children.[8] Appendix Table A presents the adult sample size for each of the six sites and up to eighteen FFS plans. We exclude Dayton, Ohio because it was the pilot site and different questionnaires were used. Finally, we exclude the sites and insurance plans with very small sample sizes.[9] Our final sample consists of three sites - Seattle, Washington, and Fitchburg and Franklin counties, Massachusetts - and two insurance plans in each site: the free plan (Plan 11) and the 95% plan (Plan 13).

Although families were randomly assigned to each plan, there is the potential problem that those in the less generous plans were worse off than those assigned to more generous plans. For example, in the free plan a participant pays nothing out-of-pocket for *all* covered services, while in the 95% plan a participant pays nothing out-of-pocket for covered inpatient services, but 95% of covered outpatient services until the deductible is met. To ensure that all families drawn by the experiment were equally likely to participate in their plans, the experiment developed a method of side payments (called Participation Incentive payments) to make families approximately equally well off across plans.

[8] RHIE defines adults as enrollees who were 14 years of age or older at the time of exit and who completed at least one of the enrollment and exit medical history questionnaires.

[9] One option would have been to aggregate these small insurance plan groups into a larger "other" category. We did not do this because we felt that the treatment effect estimated for this artificially constructed group would not be informative.

Table 1 summarizes our construction of the final sample. Panel A shows that there were 3648 adults across the three sites we considered; among them, 735 were in the free plan and 527 in the 95% plan, for a total sample size of 1262. Panel A also presents the breakdown of the free plan and 95% plan enrollees by sites. Panel B shows that, after deleting observations with missing values for the variables used in this study and outliers in terms of health expenditure variables, our final sample consists of 371 and 260 individual in the free plan and 95% plan respectively.[10]

Table 2 presents descriptive statistics. The top panel contains individual socioeconomic variables. The middle panel contains the self-reported health expenditures in the year before the experiment. The bottom panel provides total and chronic health care expenditures in the first two years of the experiment. The first and second columns present the means of the 95% and free plan samples respectively. The third column reports the mean difference for each variable between the two samples, while the last column provides the respective *t* statistic for the null hypothesis that the difference equals zero. The two samples are very similar in terms of pre-experiment variables and none of the *t* tests rejects the null hypothesis that the means are equal across the two samples, indicating that our sample selection procedures have not compromised the random assignments in the experiment. During the first two years of the experiment, the free plan sample has significantly higher total health expenditures than the 95% sample. However, the expenditures related to chronic conditions are comparable between the two groups. It is worth noting that our selected sample exhibits the usual distribution of health expenses, where around one-third of the enrollees use no medical service and a small percent of users account for half or more than half of the total expenses.

[10] We deleted extreme outliers in terms of expenditures, where we defined an outlier as an observation that was more than three standard deviations from the corresponding mean.

2.2 Previous Work Using Experimental Data to Assess the Ability of Matching to Mimic the Experimental Results

Previous studies in this strand of literature often use data from a social experiment as gold standard to investigate the efficacy of non-experimental methods. LaLonde (1986), in his seminal paper, found that nonexperimental methods (widely used by economists at that time) performed poorly in terms of replicating the experimental results. The problem of course is that those entering training are a very selective group and it is difficult to mimic this selection in random samples from the general population (LaLonde 1986). More recently, economists have turned to investigating the ability of propensity score matching to address the absence of random assignment when evaluating training programs. Dehejia (2005) and Dehejia and Wahba (1999, 2002) argue that matching can adequately address selection issues in the National Supported Work (NSW) Demonstration data, but Smith and Todd (2005) take a more pessimistic view. Furthermore, Heckman, Ichimura and Todd (1997) and Heckman, Ichimura, Smith and Todd (1998) present evidence of mixed success using matching to mimic the experimental results with the Job Training Partnership Act (JTPA) data. Our work provides valuable new information in two ways. First, we provide important evidence on how well matching works for dealing with selection issues outside of job training. Second, we deal with a different problem than that considered by the above studies. Instead of the absence of a control group, we are concerned with attrition problems within the context of social experiments. We review the relatively fewer studies dealing with non-random attrition in the following section.

2.3 Previous Approaches Addressing Non-random Attritions in Social Experiments

Hausman and Wise (1979) use an MLE version of the sample selection/control function models to account for attrition in the Negative Income Tax experiment. One problem with this approach is that it is difficult to find a variable that credibly affects attrition but not income (the outcome of interest in their study). In fact, they do not have such an exclusion restriction in their model, and thus their model is instead identified by their normality assumption. Frangakis and Rubin (1999) assume a two-arm randomized experiment comparing a new versus a standard treatment, where those assigned standard treatment cannot receive the new treatment, and those assigned the new treatment receive either the new treatment or the standard treatment. This rule will obviously result in non-compliance. They show that when, in addition, there is non-random attrition ("non-ignorable missing outcome") further complicating the experiment setting, the simple intention-to-treat estimator ignoring non-random attrition is biased. They propose a new estimation procedure for the intention-to-treat effect. Note that they deal with a double-contaminated experiment, non-compliance and attrition, which is a more complicated scenario than the problem we address. However, their solution requires two strong identification assumptions (i.e., stronger than standard matching assumptions): "latent ignorability" and the "compound exclusion restriction for never-takers".[11] DiNardo, McCrary and Sanbonmatsu (2006) suggest addressing non-random attrition under the framework of the conventional sample selection models, but their approach requires an additional random assignment other than the treatment versus control. The additional random assignment involves assigning experimental subjects to groups with varying intensity

[11] See Frangakis and Rubin (1999, pp. 369) for details about both assumptions.

in terms of follow-up effort.

3. Methodology

3.1 Introducing Non-random Attrition in the RHIE

This section describes our procedure for creating artificial attrition on the basis of chronic condition expenditures; our approach, when using total health care expenditures as the outcome variable, is completely analogous. We introduce attrition into the 95% plan sample at the end of year 1 on the basis of high chronic condition expenditures in year 2, but leave the free plan sample intact. Specifically, our induced attrition took the following form: we ranked individuals in the 95% group by their *year 2* chronic condition expenditures and randomly excluded half of the individuals above the median (i.e., a quarter of our 95% sample is deleted in year 2). We felt that this captured a plausible form of selective attrition, since in the real world we would expect individuals with potential high expenditures to drop out of the less generous plan in the absence of side payments. Because RHIE did not allow individuals to switch plans after the random assignment, we rule out the possibility of non-compliance accordingly. Thus, we are dealing with an attrition issue affecting the less generous plan without the confounding effect of non-compliance.[12]

3.2 Two Treatment Effects and Propensity Score Matching

Our goal is to assess the ability of matching to mitigate or eliminate bias due to non-random attrition when estimating treatment effects in a social experiment. We first introduce a triplet (R,T,D) of indicator functions to characterize the experiment involving attrition as

[12] In a medical trial, switching from new treatment (treatment group) to standard treatment (control group) is often allowed, and such a design forces the researcher to deal with both attrition and non-compliance (e.g. Frangakis and Rubin 1999).

described in Section 3.1. Let $R=1$ if a person is *randomized into* the treatment group (the 95% plan) and $R=0$ if a person is *randomized into* the control group (the free plan). Let T be an indicator for staying in treatment *if a person is randomized into the treatment group* $(R=1)$. Thus, among individuals with $R=1$, $T=1$ for those actually staying in treatment and $T=0$ for those leaving treatment (attrition). Note that given the definition above, T is a counterfactual for individuals assigned to the control group $(R=0)$, and it indicates whether an individual in the control group would actually have stayed in treatment had she been assigned to the treatment group. Finally, we denote *actual participation* in the 95% plan as $D=1$ and *actual participation* in the free plan as $D=0$. We have $D=1$ when $R=1$ and $T=1$, while D is undefined for individuals with $R=1$ and $T=0$ (assigned to the 95% plan but not staying); thus $D=1$ is a selective sample due to the non-random attrition. We have $D=0$ equivalent to $R=0$ because we do not introduce attrition into the free plan. Thus $D=0$, or equivalently $R=0$, is a representative sample of the underlying population. The outcome variable is year 2 chronic condition expenditures. We define two potential outcomes by treatment status, Y_0 for $D=0$ and Y_1 for $D=1$. The two treatment effects we are interested in are

$$ATET = E\left(Y^1 - Y^0 \middle| D=1\right) = E\left(Y^1 \middle| D=1\right) - E\left(Y^0 \middle| D=1\right) \tag{3.1}$$

$$ATE = E\left(Y^1 - Y^0\right) = E\left(Y^1\right) - E\left(Y^0\right). \tag{3.2}$$

The objective of most job training experiment designs (e.g. NSW and JTPA) is to estimate the ATET. The underlying population of job training experimental subjects often differs substantially from the general population; thus, the ATE (the job training effect for a random person chosen from the general population) usually is not a parameter of interest. However, the RHIE was designed to assess how different rules for sharing health care costs

between individuals and health insurance companies would affect the health care outcomes of the general population, so the ATE was, of course, the main parameter of interest. In addition, we believe that the ATET is also of interest because it reflects the potential treatment effect of switching plans for those who are more likely to choose a less generous plan when participation is up to individual choice. Hence, we investigate the effectiveness of matching in minimizing bias when estimating both ATE and ATET. The difficulty of estimating each effect arises from the fact that for each individual we observe only one of the two potential outcomes. We rely on propensity score matching to solving this missing data problem.

To estimate ATET and ATE using matching, certain assumptions must hold. The two well-known matching assumptions (Rosenbaum and Rubin 1983) are

$$\left(Y^0, Y^1\right) \perp D \middle| X, \tag{3.3}$$

$$0 < \Pr\left(D = 1 \middle| X\right) < 1, \tag{3.4}$$

where X is a vector of variables that are unaffected by the treatment. Condition (3.3), the ignorable treatment assumption (ITA, Rosenbaum and Rubin 1983), states that conditional on a set of observables X the potential outcomes are independent of actual treatment status. In empirical work, X often contains pretreatment variables and time-invariant individual characteristics. Condition (3.4) is a common support condition. It requires that at each level of X, the probabilities of observing treated and non-treated individuals are positive.[13]

In our setup, $D = 1$ if and only if $R = 1$ and $T = 1$, and $D = 0$ if and only if $R = 0$.

[13] This condition can be enforced by adding a common support constraint. The experimental setting and the way we introduce attrition (the existence of a stochastic component in the attrition rule) guarantee that this condition is satisfied. Thus, unlike most matching studies, in our study we do not have to impose a common support constraint.

Since R denotes the random assignment, it is independent of the potential outcomes by design. *Thus T is the only variable causing sample selection.* Consequently, the assumptions (3.3) and (3.4) can be rewritten as

$$\left(Y^0, Y^1\right) \perp T \left| X \right., \tag{3.5}$$

$$0 < \Pr\left(T = 1 \left| X \right.\right) < 1. \tag{3.6}$$

Let $P(X) = \Pr\left(T = 1 \left| X \right.\right)$ be the propensity of staying in treatment given X**.** Matching on all variables in X becomes impractical as the number of variables increases. To overcome this curse of dimensionality, Rosenbaum and Rubin (1983) propose propensity score matching, which reduces the multidimensional matching problem to a one-dimensional problem. They show that given conditions (3.5) and (3.6), it follows that

$$\left(Y^0, Y^1\right) \perp T \left| P(X) \right., \tag{3.7}$$

$$0 < \Pr\left(T = 1 \left| P(X) \right.\right) < 1. \tag{3.8}$$

Matching on the propensity score of staying in the treatment group (non-attrition), $P(X) = \Pr\left(T = 1 \left| X \right.\right)$, solves the dimensionality problem. The treatment effect given $p(x) = P(X = x)$ is defined as

$$\tau_{p(x)} = E\left(Y^1 \left| p(x) \right.\right) - E\left(Y^0 \left| p(x) \right.\right). \tag{3.9}$$

Following the ITA condition of (3.7), we have

$$E\left(Y^1 \left| p(x) \right.\right) = E\left(Y^1 \left| p(x), T = 1 \right.\right) = E\left(Y^1 \left| p(x), T = 0 \right.\right) \text{ and}$$

$$E\left(Y^0 \left| p(x) \right.\right) = E\left(Y^0 \left| p(x), T = 1 \right.\right) = E\left(Y^0 \left| p(x), T = 0 \right.\right). \tag{3.10}$$

Further, since $\left(Y^0, Y^1\right) \perp T \left| P(X) \right.$ implies $\left(Y^0, Y^1\right) \perp D \left| P(X) \right.$, we have

$$E\left(Y^1 \middle| p(x)\right) = E\left(Y^1 \middle| p(x), D=1\right) = E\left(Y^1 \middle| p(x), D=0\right) \text{ and}$$

$$E\left(Y^0 \middle| p(x)\right) = E\left(Y^0 \middle| p(x), D=1\right) = E\left(Y^0 \middle| p(x), D=0\right). \quad (3.11)$$

Equation (3.11) shows how propensity score matching solves the problem of observing only one potential outcome for each individual. For individuals in the $D=1$ sample, we construct the unobserved potential outcomes by relying on $E\left(Y^0 \middle| p(x)\right) = E\left(Y^0 \middle| p(x), D=1\right) = E\left(Y^0 \middle| p(x), D=0\right)$. Likewise, for individuals in the $D=0$ sample, we construct the unobserved potential outcome by using $E\left(Y^1 \middle| p(x)\right) = E\left(Y^1 \middle| p(x), D=0\right) = E\left(Y^1 \middle| p(x), D=1\right)$. We first estimate the treatment effect at specific values of $p(x)$; then taking the expectation of $\tau_{p(x)}$ with respect to the distribution of $p(x)$ in the treatment group yields the ATET, while taking the expectation of $\tau_{p(x)}$ with respect to the distribution of $p(x)$ in the general population yields the ATE.

Since the propensity score of staying in the treatment group $P(X) = \Pr(T=1|X)$ is unknown in a real world setting, we estimate it for each individual in our $D=1$ and $D=0$ samples. For those in the $D=1$ sample, this estimation is straightforward. We observe both $T=1$ and $T=0$ for individuals in the $R=1$ group (from our simulated attrition process), and we use a logit model to estimate $\Pr(T=1|X=x)$ using the whole $R=1$ sample. However, for individuals in the $D=0$ (or equivalently the $R=0$) group, T is a counterfactual. Since both $R=1$ and $R=0$ are representative samples of the same underlying population, we assume that individuals in the $R=0$ sample will exhibit the same attrition behavior had they been assigned to $R=1$ conditional on X. We impute $\Pr(T=1|X=x)$ for individuals in $D=0$ sample by using their X realizations and borrowing

the logit coefficients estimated from the $R=1$ sample. We discuss our detailed estimation procedures in the two sections below.

3.3 Our Method for Estimating the ATET

In this section we focus on the ATET, the bias arising from ignoring the non-random attrition, and the ability of matching to mitigate this bias when the outcome variable is year 2 chronic condition expenditures. Our approach for year 2 total health care expenditures is completely analogous.[14]

Our matching approach for estimating the ATET is as follows. For each individual $i, i=1,...,n_1$, in the $D=1$ sample with the estimated propensity score $p(x_i)$ (the estimation process described in Section 3.2), we first estimate the treatment effect defined by equation (3.9). The observed year 2 chronic condition expenditure is used to estimate $E\left(Y_i^1 \middle| \hat{p}(x_i)\right)$, and we use propensity score matching to estimate the counterfactual $E\left(Y_i^0 \middle| \hat{p}(x_i)\right)$. Next we take the average of $\widehat{\tau_{p(x_i)}}$ over the $D=1$ sample to yield the $\widehat{ATET}$

$$\widehat{ATET} = \frac{1}{n_1}\sum_{i=1}^{n_1} \widehat{\tau_{p(x_i)}} = \frac{1}{n_1}\sum_{i=1}^{n_1} \widehat{E\left(Y_i^1 \middle| \hat{p}(x_i)\right) - E\left(Y_i^0 \middle| \hat{p}(x_i)\right)}. \qquad (3.12)$$

Taking the average over the $D=1$ sample reflects the notion discussed in the previous section: taking the expectation of $\tau_{p(x)}$ with respect to the distribution of $p(x)$ in the treatment group.

Because there is a stochastic component in our attrition process plus the usual data

[14] As noted above we induce attrition on the basis of year 2 total expenditures when it is our outcome of interest.

structure of health care expenditures[15], the sample distribution of health expenditures after attrition may exhibit large variation from simulation to simulation. To avoid extreme cases in a single simulation, we simulate the attrition, and match 100 times and take averages over the 100 simulations. Although researchers dealing with "broken experiments" with non-random attrition have only one data set available, our estimates will be informative of the average biases involved. Specifically, we take the following estimation steps for each of the 100 simulations:

1. Estimate the "true" ATET as the benchmark for matching estimates.

   The RHIE does not provide the true ATET directly. To yield the true ATET, we first introduce symmetric non-random attrition to the free plan sample (as we have done for the 95% plan sample), randomly excluding half of the free plan sample above the median of the year 2 expenditures. We then take the difference in means between the two groups (both after attrition).

2. Estimate the ATET by ignoring non-random attrition.

   To show the potential bias resulting from ignoring non-random attrition, we take the mean difference of the outcome between the $D=1$ sample (95% plan after attrition) and the entire $D=0$ sample. This difference in means provides the naïve ATET estimate obtained by ignoring non-random attrition.

3. Estimate the ATET using the propensity score matching.

   We first construct the counterfactual $E\left(Y_i^0 \middle| \hat{p}\left(x_i\right)\right), i=1...,n_1$ for each individual in the $D=1$ sample, using Kernel regression matching (KRM) and local linear

[15] As discussed in Section 2.1, approximately one-third of the enrollees use no medical services and a small percent of users account for half or more than half of the total expenses.

regression matching (LLRM).[16] Then we plug in the estimated counterfactuals into equation (3.12) to yield ATET estimates.

Finally, we obtain our objects of interest - the "true," the naïve, and the matching ATET estimates - by averaging the results from the above three steps over 100 simulations.

3.4 Our Method for Estimating the ATE

In this section we discuss estimating the ATE when the outcome variable is year 2 chronic condition expenditures; again our approach for total health care expenditures is exactly analogous. For each individual $j,\ j=1,...,n_0$ in the $D=0$ sample with the imputed propensity score $p(x_j)$ (the imputation process described in Section 3.2), we first estimate the treatment effect defined by equation (3.9). The observed year 2 chronic condition expenditure is used to estimate $E\left(Y_j^0 \middle| \hat{p}(x_j)\right)$, and we use propensity score matching to estimate the counterfactual $E\left(Y_j^1 \middle| \hat{p}(x_j)\right)$. Next we take the average of $\widehat{\tau_{p(x_j)}}$ over the $D=0$ sample to yield $\widehat{ATE}$ such that

$$\widehat{ATE} = \frac{1}{n_0}\sum_{j=1}^{n_0}\widehat{\tau_{p(x_j)}} = \frac{1}{n_0}\sum_{j=1}^{n_0}\widehat{E\left(Y_j^1 \middle| \hat{p}(x_j)\right)} - \widehat{E\left(Y_j^0 \middle| \hat{p}(x_j)\right)}. \qquad (3.13)$$

Taking the average over the $D=0$ sample (a representative sample of the general population), reflects the notion discussed in the previous section: taking the expectation of $\tau_{p(x)}$ with respect to the distribution of $p(x)$ in the general population.

To investigate the ability of matching to mitigate the bias due to attrition in estimating the ATE, we take the following steps:

[16] See Appendix A for a brief review of KRM and LLRM; for more detailed descriptions of these procedures, see e.g. Frölich (2004). Hirano, Imbens and Ridder (2003) provide an alternative matching procedure that would be interesting to investigate in future work.

1. Obtain the true ATE estimate by taking the mean difference in outcomes for the full 95% and the free plan.[17]

Then for each of 100 simulations we:

2. Estimate the ATE by ignoring non-random attrition.

   Again, we take the mean difference of the outcome between the $D=1$ sample (95% plan after attrition) and the entire $D=0$ sample. This difference in means provides the naïve ATE estimate by ignoring non-random attrition.[18]

3. Estimate the ATE using the propensity score matching.

   We first construct the counterfactual $E\left(Y_j^1 \middle| \hat{p}\left(x_j\right)\right),\ j=1...,n_0$ for each individual in the $D=0$ sample, again using KRM and LLRM. Then we plug in the estimated counterfactuals into equation (3.13) to yield the ATE estimates.

Finally, we obtain our objects of interest - the ATE estimates from the naïve and matching estimators - by taking the averages of the results from steps 2 and 3 over 100 simulations.

3.5 The Conditioning Variables used in our Propensity Score Matching

Given our attrition problem, the most crucial identification condition for matching is the ITA assumption given in equation (3.3) or equivalently (3.5) for our attrition problem, where X is a vector of variables that are unaffected by the treatment. Thus, one of the most important decisions when using matching is choosing the conditioning variables X, so that one can argue convincingly that the ITA assumption is satisfied. Of course, we cannot test whether the ITA will be satisfied given the X variables available to us. However, because

[17] Note that this is done once over the original sample, not over the 100 simulated samples as in estimating ATET, because there is no stochastic component in estimating the ATE.
[18] Since the naïve estimator ignores the non-random attrition, it is the same for both the ATE and the ATET.

we have the complete experimental data, we can assess how well our choice of X allows matching to mimic the experimental evidence.

We first include in X the demographic variables shown in the top panel of Table 2: gender, age, highest grade completed, and marital status. However, these variables are unlikely by themselves to allow us to achieve the ITA, so we also include two potentially more powerful conditioning variables: family income[19] and a dummy variable for having health insurance at the workplace pre-baseline. Finally, since attrition is positively correlated with health expenditures in year 2, it would be extremely useful to include variables that capture an individual's propensity toward health care expenditure under either plan. A natural first choice for this variable is pre-experimental health care expenditure; however, this variable is self-reported and is thus may be unreliable. Indeed, our empirical results indicate that including it in the conditioning variables is not sufficient to allow matching to mitigate the bias due to attrition in our application.

To address this deficiency, we construct an additional conditioning variable based on post-experimental expenditures, which will be valid under the assumption that the insurance plan assignment does not affect the rank of a participant's health care spending in a given plan, i.e. individuals are rank-order stable across the two plans. Given this assumption, we order the individuals in each plan in terms of their health spending in year 1, and then use this rank (normalized by sample size) as an explanatory variable in the propensity score.[20] The intuition behind this ranking is that we want to match high-expenditure individuals in the 95% plan with high-expenditure individuals in the free plan. To construct a counterfactual for an individual in the 95% plan whose health expenditure is at the 80th percentile in year 1, we can use individuals whose health expenditures are close to

[19] We use the logarithm of family income normalized by family size.
[20] When considering chronic (total) care expenditures, we use their year 1 rank for chronic (total) expenditures.

the 80th percentile in the free plan in year 1. Our empirical evidence suggests that using the year 1 rank in a given plan as a conditioning variable can significantly improve the performance of matching. Thus, our modification should be helpful to other researchers.

Finally, the families in our sample live in different cities, and health care costs are likely to vary across these cities. Therefore, it may be appropriate to match only individuals living in the same city. We also consider this approach, which is known as finer balancing (Rosenbaum and Rubin 1985).[21]

4. Empirical Results

4.1 *A Priori* Expectations for the Results

Before examining the empirical results, it is worth discussing how we might expect matching to perform in different circumstances. To estimate the ATET, we need to construct a counterfactual only for a relatively small group (the 95% plan sample after attrition), given the observed outcomes for a relatively large group (the whole free plan sample). However, to estimate the ATE we need to construct a counterfactual for the larger group, the whole free plan sample, given the outcomes we observed for the smaller group, i.e. the remaining 95% group after attrition. Frölich (2004) found that matching is likely to perform better when the group for which the counterfactual needs to be constructed is considerably smaller than the group whose outcomes form the basis for counterfactual construction. Therefore, we would expect matching to do a better job of eliminating bias for the ATET estimates than for the ATE estimates.

Secondly, it is unclear whether matching will perform better when the conditioning variables include the rank of year 1 health expenditures. On the one hand, the rank of year 1

[21] Ham, Li and Reagan (2011) found using finer balancing (based on education category) to be important in their study of the returns to migration for those who moved.

expenditures will be a much better proxy variable for omitted factors determining the year 2 expenditures than self-reported pre-experimental health expenditures. On the other hand, year 1 rank will not be a valid conditioning variable if it is affected by treatment assignment, contrary to our assumption that individual are rank-order stable across plans. In our sample, pre-baseline expenditures have a correlation of 0.16 with the year 2 chronic condition expenditures and 0.1 with the year 2 total expenditures while the correlation between chronic expenditure rank in year 1 and chronic expenditures in year 2 is 0.37, and the corresponding correlation for total expenditures is 0.18.

Finally, we would also expect matching to perform better in terms of eliminating bias for the more predictable (by the conditioning variables) aspects of health expenditure. Thus, we believe that matching will be more effective in reducing bias in estimates with respect to chronic condition expenditures than with respect to total health care expenditures. Total health care expenditures include expenditures related to both acute conditions and chronic conditions, and thus are much less correlated over time than chronic care expenditures. For example, in our final sample, the correlations between year 1 and year 2 expenditures are 0.39 and 0.19 for chronic and total expenditure respectively.

4.2 Results for Estimating the ATET

Table 3 presents our results for estimating the ATET as described in Section 3.3. In Panel A, the propensity score model includes the socioeconomic variables discussed in Section 3.5, plus the (self-reported) pre-experimental expenditures and the rank of year 1 expenditures. The results when the outcome variable is defined as chronic condition expenditures are presented in Row (1), and those when the outcome variable is defined as

total health care expenditures are presented in Row (2). The "true" ATET in Column (1),[22] i.e. the average effect of moving those remaining in the 95% plan group after the attrition to the free plan, shows that this "external" switching would, on average, reduce chronic condition expenditures by $1.96.[23] The analogous ATET estimate for total health expenditures is $185. Column (2) of Panel A presents estimates when we ignore the attrition bias and simply take the mean difference between the 95% plan sample (after attrition) and the whole free plan sample: the ATET estimates would be approximately $40 and $436 for chronic condition expenditures and total expenditures respectively. Column (3) presents the corresponding biases (Column (2) minus Column (1)) resulting from this naïve estimator. The biases are in the expected direction, since the free plan without attrition contains relatively more high-expenditure individuals than the 95% plan after attrition.

Columns (4) and (7) of Panel A present the LLRM and KRM matching estimates respectively.[24] Row (1) indicates that the matching ATET estimates for chronic condition expenditures using LLRM and KRM to be $21 and $20 respectively.[25] Columns (5) and (8) demonstrate that the bias involved in using LLRM and KRM is $19 and $18 respectively. Finally, Column (6) presents the ratio of Column (5) to Column (3), and Column (9) presents the ratio of Column (8) to Column (3). Columns (6) and (9) show that both matching methods reduce the magnitude of the bias by around half. Row (2) presents the

[22] The "true" ATET is estimated by introducing symmetric attrition in the 95% and free plans as described in Section 3.3.

[23] All dollar amounts are in nominal values.

[24] We use a variable bandwidth, as suggested by Fan Gijbels (1992), for all of our matching estimates. See Appendix A for a discussion of fixed (global) versus variable bandwidths for LLRM and KRM.

[25] We use the bootstrap to calculate the standard errors for the "true" ATET, the estimated treatment effects when we ignore attrition, and the LLRM and KRM estimates. While there is no theoretical justification for using the bootstrap with the LLRM and KRM estimators, and Abadie and Imbens (2008) show that the bootstrap is inappropriate when using nearest neighbor matching, Ham, Li and Reagan (2011) present Monte Carlo evidence that the bootstrap works extremely well for LLRM. Since LLRM and KRM are similar methods, we would expect their results to hold for KRM. In each case we use the Andrews-Buchinsky (2000, 2001) algorithm to choose the number of bootstrap replications.

corresponding estimates for total health care expenditures. Here we see that ignoring attrition would overestimate the ATET for total health care expenditure by about $250. Further, the rest of the columns show that both LLRM and KRM only reduce this bias by a trivial amount.

Panel B of Table 3 uses the same conditioning variables as in Panel A when we implement finer balancing matching, i.e. when matching we use only individuals in the same city.[26] The true ATET and the bias from the simple estimator ignoring attrition in Columns (1) and (2) of Row (1) of Panel B are, of course, the same as in Panel A, since only the matching procedure changes across the two panels. The matching estimates using finer balancing, and the corresponding entries shown in Columns (4) to (9) are quite similar to those in Panel A.[27]

Finally, Panel C of Table 3 shows the results when we drop the rank of year 1 health expenditures as a conditioning variable in the propensity score. The matching estimators perform strikingly poorly since they essentially eliminate none of the biases due to attrition. Based on our results, we conclude that year 1 rank is a "good" conditioning variable and that using such a variable can be effective in reducing bias introduced by non-random attrition.

4.3 Results for Estimating the ATE

Table 4 presents the corresponding results when we estimate the ATE. Recall that the ATE captures the treatment effect on expenditure when "externally" moving an average person in the population from the 95% plan to the free plan, and as noted above we would

[26] We do not have enough data to estimate the propensity score model separately for the three locations.
[27] Interestingly, the standard errors for the LLRM estimates increase substantially compared to panel A. This suggests that KRM will outperform LLRM when the number of observations in local regressions is reduced due to finer balancing. Since we will see this phenomenon again when estimating ATE, it seems more appropriate to use KRM with finer balancing than LLRM unless the sample is large.

expect it to be more challenging to estimate than the ATET. First, it requires constructing a counterfactual for the larger group, the whole free plan sample, given the outcomes we observed for the smaller group (the remaining 95% group). Second, the remaining 95% group has relatively smaller sample variation due to (our induced) non-random attrition. As noted in Section 2.1, the samples in this study follow the usual distribution of health expenses, where more than one-third of the enrollees use no medical services. Thus, much of the sample variation comes from individuals with expenditures above median. When we induce attrition by randomly excluding half of the individuals with above median expenditures, a significant amount of sample variation is lost in the outcome variable for the 95% plan sample.

Table 4 is structured symmetrically to Table 3. Again, in Panel A, the propensity score model includes the socioeconomic variables stated in Section 3.5, plus the pre-experiment (self-reported) expenditures and the rank of year 1 expenditures within plan; Rows (1) and (2) of the panel present results when chronic and total health care expenditure respectively are the outcome variables. Column (1) presents the true ATE estimated from the experimental data. The effect of externally switching an average person from the 95% plan to the free plan would, in year 2, increase chronic condition expenditures by $1.45 and total health expenditures by $216. The ATE estimate of $216 on total health expenditures is very close to the published evidence from RHIE (e.g. Newhouse et al. 1987). Evidence on how the RHIE affected chronic condition related utilizations is often based on episodes of care and not on the level of annual expenditures (see e.g. Lohr et al. 1986) and thus not directly comparable to our results. Columns (2) and (3) demonstrate that substantial biases arise if we ignore the non-random attrition and take the difference in means between the two

groups.[28] Columns (4) and (7) of Row (1) of the panel present the estimates for LLRM and KRM respectively, and show that they estimate the ATE to be $24 and $27 respectively for chronic condition expenditures; thus for the ATE LLRM and KRM eliminate only 43% and 35% of the bias respectively. As suggested by Frölich's (2004) Monte Carlo results, matching does a better job of estimating the ATET than estimating the ATE (for chronic condition expenditures). Row (2) indicates that LLRM and KRM estimate the ATE to be about $435 and $431 respectively for total health care expenditures. Since these point estimates are very close to the estimate from the naïve estimator (ignoring the attrition) in Column (2) of $436, matching again does not eliminate any of the attrition bias for total expenditures.

Panel B in Table 4 uses the same conditioning variables as in Panel A when we implement finer balancing matching; again, the matching results and bias reductions are similar to those in Panel A.[29] In Panel C of Table 4, we drop the rank of year 1 health expenditures as a conditioning variable in the propensity score. Matching again performs poorly in estimating treatment effects and eliminating bias for both chronic and total health care expenditures.

5. Conclusion

Both non-compliance and non-random attrition can contaminate social experiments. However, non-random attrition in social experiments has received less attention than non-compliance in the literature. Furthermore, it cannot be addressed by looking at an intent-to-treat parameter or by an IV approach, as is the case with non-compliance. In principle, propensity score matching offers an attractive solution to the attrition problem when the

[28] Note that this difference is the same as that in Table 3 by definition.
[29] However, note that we get a very large standard error for the LLRM estimate on total expenditure.

social experiment has a rich set of baseline variables and the selection-on-observables assumption becomes more plausible.

In this paper we investigate how well matching works in terms of eliminating bias in estimated treatment effects after we induce plausible non-random attrition in the well-known Rand Health Insurance Experiment (RHIE) data. Additionally, we investigate a modification of standard matching procedures that seems particularly useful when using health expenditure data, but should also have applications outside of health economics. Specifically, we assume that a family's expenditures are rank-order stable across the two plans, and then use their year 1 rank in their respective plan as a conditioning variable in the propensity score. Since matching may perform differently for different components of health expenditures, we consider two outcome variables: health care expenditures on chronic conditions only and total health care expenditures. Because the data come from the first two years of a social experiment with little contamination, we can use the experimental data to benchmark how well our matching estimates mitigate the biases resulting from our artificially induced non-random attrition.

We focus on two RHIE insurance plans: the 95% plan (the least generous fee-for-service plan in RHIE) and the free plan (the most generous fee-for-service plan in RHIE). Our artificially introduced attrition takes the form of randomly excluding, at the end of year 1, 50% of the individuals with year 2 expenditures above the median in the 95% plan. We do not introduce any attrition in the free plan group. We estimate both the treatment effect for those remaining in the 95% plan - the ATET - and for the general population underlying the RHIE - the ATE. We find first that when we use year 1 rank as a conditioning variable, matching performs moderately well in terms of mitigating the biases in chronic condition expenditures but not in total health care expenditures. Second, if we do not use year 1 rank

as a conditioning variable, matching does a poor job of eliminating biases in treatment effects for both types of health care expenditures. Third, we find matching does a slightly better job of reducing biases in the ATET than in the ATE estimates, as expected given that the remaining 95% plan sample is smaller and has less variation than the full free plan sample.

References:

Abadie, A. and Imbens, G. (2008). "On the Failure of the Bootstrap for Matching Estimators." *Econometrica*, 76, 1537-1557.

Andrews, D. W. K. and Buchinsky, M. (2000). "A Three-Step Method for Choosing the Number of Bootstrap Repetitions." *Econometrica*, 67, 23-51.

——— (2001). "Evaluation of a Three-Step Method for Choosing the Number of Bootstrap Repetitions." *Journal of Econometrics*, 103, 345-386.

Barnard, J., Du, J., Hill, J.L. and Rubin, D.B. (1998). "A Broader Template for Analyzing Broken Randomized Experiments." *Sociological Methods and Research*, 27, 285–317.

Bhandari, A., and Wagner, T. (2006). "Self-Reported Utilization of Health Care Services: Improving Measurement and Accuracy." *Medical Care Research and Review*, 63, 217-235.

Blundell, R., Dearden, L. and Sianesi, B. (2005). "Evaluating the effect of education on earnings: models, methods and results from the National Child Development Survey." *Journal of the Royal Statistical Society Series A*, Royal Statistical Society, 168, 473-512.

Borghans, L., Duckworth, A. L., Heckman, J. J. and ter Weel, B. (2008). "The Economics and Psychology of Personality Traits." *Journal of Human Resources*, 43, 972-1059.

Dehija, R. (2005). "Practical Propensity Score Matching." *Journal of Econometrics*, 125, 255-264.

Dehija, R. and Wahba, S. (1999). "Causal Effects in Non-Experimental Studies: Re-Evaluating the Evaluation of Training Programs." *Journal of the American Statistical Association*, 94, 1053-1062.

——— (2002). "Propensity Score Matching Methods for Non-Experimental Causal Studies." *Review of Economics and Statistics*, 84, 151-161.

DiNardo, J., McCrary, J., and Sanbonmatsu, L. (2006). "Constructive Proposals for Dealing with Attrition: An Empirical Example." Working paper.

Ding, W., and Lehrer, S. (2010). "Estimating Treatment Effects from Contaminated Multi-Period Education Experiments: The Dynamic Impacts of Class Size Reductions." *Review of Economics and Statistics,* 92, 31-42.

Eberwein, C., Ham, J. C. and LaLonde, R. (1997). "The Impact of Being Offered and Receiving Classroom Training on the Employment Histories of Disadvantaged Women: Evidence From Experimental Data." *The Review of Economic Studies*, 64, 655-682.

Fan, J. and Gijbels, I. (1996). *Local Polynomial Modeling and Its Applications*. Monographs on Statistics and Applied Probability 66 (London: Chapman & Hall).

——— (1992). "Variable Bandwidth and Local Regression Smoothers." *Annals of Statistics*, 20, 2008-2036.

Frangakis, C.E., and Rubin, D.B. (1999). "Addressing Complications of Intention-to-Treat Analysis in the Presence of All-or-None Treatment-Noncompliance and Subsequent Missing Outcomes." *Biometrika*, 86, 365–379.

Frölich, M. (2004). "Finite Sample Properties of Propensity-Score Matching and Weighting Estimators," *Review of Economics and Statistics*, 86, 77-90.

Ham, J., Li, X., and Reagan, P. (2011). "Propensity Score Matching, a Distance-Based Measure of Migration, and the Wage Growth of Young Men." *Journal of Econometrics*, 161, 208-227.

Hausman, J., and Wise, D. (1979). "Attrition Bias in Experimental and Panel Data: The Gary Income Maintenance Experiment," *Econometrica*, 47, 455-473.

Heckman, J., Ichimura, H. and Todd, P. (1997). Matching as an Econometric Evaluation

Estimator: Evidence from Evaluating a Job Training Program." *Review of Economic Studies*, 64, 605-654.

Heckman, J., Ichimura, H., Smith, J. and Todd, P. (1998). "Characterizing Selection Bias Using Experimental Data." *Econometrica*, 66, 1017-1098.

Heckman, J., LaLonde, R., and Smith, J. (1999). "The Economics and Econometrics of Active Labor Market Programs" in the *Handbook of Labor Economics,* Volume 3, O. Ashenfelter and D. Card (eds), North Holland, Amsterdam.

Hirano, K., Imbens, G. and Ridder, G. (2003), "Efficient Estimation of Average Treatment Effects Using the Estimated Propensity Score." *Econometrica,* 71, 1161-1189.

Lalonde, R. J. (1986). "Evaluating the Econometric Evaluations of Training Programs Using Experimental Data." *American Economic Review*, 76, 602-620.

Lohr, K. N., Brook, R. H., Kamberg, C. J., Goldberg, G. A., Leibowitz, A., Keesey, J., Reboussin, D., Newhouse, J.P. (1986). "Use of Medical Care in the Rand Health Insurance Experiment: Diagnosis- and Service-Specific Analyses in a Randomized Controlled Trial," *Medical Care*, 24, S1-S87.

Newhouse, J.P., Manning, W.G., Duan, N., Morris, C.N., Keeler, E.B., Leibowitz, A., Marquis, M.S., Rogers, W.H., Davies, A.R., Lohr, K. N., Ware, J.E., Brook, R.H. (1987). "The Findings of the Rand Health Insurance Experiment - A Response to Welch et al.," Medical Care, 25, 157-179.

Rosenbaum, P. R. and Rubin, D. B. (1983). "The Central Role of the Propensity Score in Observational Studies for Casual Effects." *Biometrika*, 70, 41-55.

Rosenbaum, P. R. and Rubin, D. B. (1985). "Constructing a Control Group Using Multivariate Matched Sampling Methods That Incorporate the Propensity Score." *The American Statistician*, 39, 33-38.

Robins, J. M. (1986) "A New Approach to Causal Inference in Mortality Studies with Sustained Exposure Periods - Application to Control of the Healthy Worker Survivor Effect." *Mathematical Modelling*, 7, 1393-1512, with 1987 Errata, *Computers and Mathematics with Applications*, 14, 917-21; 1987 Addendum, *Computers and Mathematics with Applications*, 14, 923-45; and 1987 Errata to Addendum, *Computers and Mathematics with Applications*, 18, 477.

Sianesi, B. (2010). "An introduction to matching methods for causal inference and their implementation in Stata." *United Kingdom Stata Users' Group Meetings 2010 13*, *Stata Users Group.* (Downloadable at http://ideas.repec.org/e/psi28.html .)

Smith, J. and Todd, P. (2005). "Does Matching Overcome LaLonde's Critique of Nonexperimental Methods?" *Journal of Econometrics* 125, 305-353.

Table 1. Sample Selection Criteria and Sample Construction

Panel A: Sample Selection

Seattle, Washington, Fitchburg, Massachusetts, and Franklin County, Massachusetts (site = 2,3,4) insured adult enrollees (14 years of age or older at the time of exit and completed at least one of the enrollment and exit medical history questionnaires).	3648
plan = 11 and plan = 13	1262
Sample composition by plan	
Plan 11 (Free Plan)	735
Plan 13 (95% Plan)	527
Total	1262
Sample composition by site	
site = 2 (Seattle)	552
site = 3 (Fitchburg)	332
site = 4 (Franklin Country)	378
Total	1262

Panel B: Sample Construction

	Plan 11 (Free Plan)	Plan 13 (95% Plan)
	735	527
After constraining socioeconomic variables (conditioning variables in the propensity score model), pre-experiment self-reported health expenditure, and year 1 and year 2 health expenditures to be available	385	270
After deleting outliers for pre-experimental expenditure and year 2 total health expenditure and year 2 health expenditure related to chronic conditions.	371	260

Notes:
1. Plan=11 (free plan): Participant pays nothing out-of-pocket for covered services.
2. Plan=13 (95% plan): Participant pays nothing out-of-pocket for covered inpatient services but pays 95% of covered outpatient services until deductible is met. Then plan pays 100%. The deductible is $150 per person or $450 per family.
3. Outliers are defined as observations more than three standard deviations from their respective means.

Table 2. Variable Definitions and Descriptive Statistics

Variable Name	Variable Definition	95% Plan	Free Plan	Difference	*t* statistic for Difference
Socioeconomic Variables					
Sex	=1 if male	0.44	0.47	-0.03	-0.69
Age	Age in years	34.58	34.84	-0.26	-0.27
HGC	Highest grade completed	12.68	12.60	0.09	0.42
Income	logarithm family income normalized by family size	9.44	9.39	0.05	1.19
Work-Ins	=1 if having insurance at the workplace	0.82	0.81	0.01	0.21
MSP	=1 if married, spouse present	0.75	0.72	0.03	0.81
Pre-experiment Health Expenditure					
TotExp0	Total health expenditure in the year before the experiment (self-reported)	84.67	91.01	-6.34	-0.58
Health Expenditures During the Experiment					
TotExp1	Total health expenditure in year 1	905.10	1140.11	-235.01	-1.84
TotExp2	Total health expenditure in year 2	751.02	966.62	-215.61	-2.16
Chronic1	Chronic condition health expenditure in year 1	114.86	138.91	-24.05	-0.91
Chronic2	Chronic condition health expenditure in year 2	113.66	115.11	-1.45	-0.08

Table 3. Experimental and Matching Estimates of the ATET

Panel A: Conditioning Variable = Pre-experiement Expenditure + Year 1 Expenditure Rank + Other Controls

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
	True ATET	**Sample Diff**	**Bias**	**LLRM**	**Bias**	**Bias Ratio**	**KRM**	**Bias**	**Bias Ratio**
Chronic Condtion Expenditure	**1.96**	**40.17**	**38.21**	**21.31**	**19.35**	**0.51**	**20.46**	**18.50**	**0.48**
Standard Error (200 bootstrap replications)	(13.00)	(16.12)		(14.35)			(14.51)		
Total Expenditure	**185.51**	**435.68**	**250.17**	**414.60**	**229.09**	**0.92**	**414.07**	**228.56**	**0.91**
Standard Error (200 bootstrap replications)	(71.82)	(85.03)		(81.28)			(81.78)		

Panel B: Conditioning Variable = Pre-experiement Expenditure + Year 1 Expenditure Rank + Other Controls
Finer Balancing Matching

	True ATET	**Sample Diff**	**Bias**	**LLRM**	**Bias**	**Bias Ratio**	**KRM**	**Bias**	**Bias Ratio**
Chronic Condtion Expenditure	**1.96**	**40.17**	**38.21**	**20.48**	**18.52**	**0.48**	**19.35**	**17.39**	**0.46**
Standard Error (200 bootstrap replications)	(13.00)	(16.12)		(20.65)			(14.95)		
Total Expenditure	**185.51**	**435.68**	**250.17**	**437.00**	**251.49**	**1.01**	**403.78**	**218.27**	**0.87**
Standard Error (200 bootstrap replications)	(71.82)	(85.03)		(127.21)			(81.77)		

Panel C: Conditioning Variable = Pre-experiement Expenditure + Other Controls

	True ATET	**Sample Diff**	**Bias**	**LLRM**	**Bias**	**Bias Ratio**	**KRM**	**Bias**	**Bias Ratio**
Chronic Condtion Expenditure	**1.96**	**40.17**	**38.21**	**35.89**	**33.93**	**0.89**	**35.84**	**33.88**	**0.89**
Standard Error (200 bootstrap replications)	(13.00)	(16.12)		(15.62)			(15.28)		
Total Expenditure	**185.51**	**435.68**	**250.17**	**416.78**	**231.27**	**0.92**	**418.61**	**233.1**	**0.93**
Standard Error (200 bootstrap replications)	(71.82)	(85.03)		(80.50)			(81.13)		

Notes:
1. The "true" ATET is calculated by introducing the same attrition (see Section 3.3 for details) into the free plan sample as in the 95% sample and then taking the difference in means between the two remaining samples. Since there is a stochastic component in this attrition rule, we use 100 simulations and then take the averages across simulations to estimate this effect.
2. The sample differences after attrition are calculated by taking the difference in means between the remaining 95% group and the whole free plan group. Again, they are the averages taken over 100 simulations.
3. We implement matching for each of the 100 simulations, and the matching estimates in this table are averages across simulations.

Table 4. Experimental and Matching Estimates of the ATE

Panel A: Conditioning Variable = Pre-experiement Expenditure + Year 1 Expenditure Rank + Other Controls

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
	True ATE	**Sample Diff**	**Bias**	**LLRM**	**Bias**	**Bias Ratio**	**KRM**	**Bias**	**Bias Ratio**
Chronic Condtion Expenditure	**1.45**	**40.17**	**38.72**	**24.38**	**22.93**	**0.57**	**26.65**	**25.20**	**0.65**
Standard Error (200 bootstrap replications)	(19.27)	(16.12)		(17.17)			(17.12)		
Total Expenditure	**215.6**	**435.68**	**220.08**	**435.1**	**219.50**	**1.00**	**431.24**	**215.64**	**0.98**
Standard Error (200 bootstrap replications)	(99.67)	(85.03)		(85.93)			(86.54)		

Panel B: Conditioning Variable = Pre-experiement Expenditure + Year 1 Expenditure Rank + Other Controls **Finer Balancing Matching**

	True ATE	**Sample Diff**	**Bias**	**LLRM**	**Bias**	**Bias Ratio**	**KRM**	**Bias**	**Bias Ratio**
Chronic Condtion Expenditure	**1.45**	**40.17**	**38.72**	**16.20**	**14.75**	**0.37**	**18.59**	**17.14**	**0.44**
Standard Error (200 bootstrap replications)	(19.27)	(16.12)		(30.82)			(18.53)		
Total Expenditure	**215.6**	**435.68**	**220.08**	**437.91**	**222.31**	**1.01**	**427.98**	**212.38**	**0.97**
Standard Error (200 bootstrap replications)	(99.67)	(85.03)		(379.88)			(88.54)		

Panel C: Conditioning Variable = Pre-experiement Expenditure + Other Controls

	True ATE	**Sample Diff**	**Bias**	**LLRM**	**Bias**	**Bias Ratio**	**KRM**	**Bias**	**Bias Ratio**
Chronic Condtion Expenditure	**1.45**	**40.17**	**38.72**	**40.31**	**38.86**	**0.97**	**39.14**	**37.69**	**0.97**
Standard Error (200 bootstrap replications)	(19.27)	(16.12)		(16.49)			(16.28)		
Total Expenditure	**215.6**	**435.68**	**220.08**	**435.02**	**219.42**	**1.00**	**427.7**	**212.10**	**0.96**
Standard Error (200 bootstrap replications)	(99.67)	(85.03)		(85.73)			(86.04)		

Notes:
1. The true ATE is calculated by taking the difference in means between the 95% group (before attrition) and the free plan group.
2. The sample difference after attrition is the same as those in Table 3.
3. We implement matching for each of the 100 simulations, and the matching estimates in this table are averages across simulations.

Appendix A: A Brief Review of Kernel and Local Linear Regression Matching Estimators

One faces a trade-off when choosing between Kernel Regression Matching (KRM) and Local Linear Matching (LLRM). On the one hand, LLRM estimators have better asymptotic properties (Fan and Gijbels 1996). On the other hand, the Monte Carlo evidence in Frölich (2004) suggests that KRM is more robust in finite samples. We compare our estimates using the two matching estimators. The next decision a researcher faces is the choice of the bandwidth for these procedures: the smaller the bandwidth, the smaller the bias but the larger the variance. Further, there is the issue of whether one should use a fixed bandwidth or use variable bandwidths, which vary with distribution of the estimated propensity score. The variable bandwidth approach has the advantage of allowing a smaller bandwidth where the data are concentrated and a larger bandwidth where the data are sparse. Fan and Gijbels (1992, p. 2013) argue that it is advantageous to combine local regression with variable bandwidth, and we follow their suggestion below by using the adaptive variable bandwidth proposed by Fan and Gijbels (1996). In their procedure the size of the window is chosen to include the same fraction of the sample closest to the point at which we conduct local regression. Our variable bandwidth includes 15% of the data for each point at which we run a local regression to estimate a counterfactual.[30]

KRM and LLRM are two special cases of local regression matching. Below we discuss KRM and LLRM in the context of Section 3.3, and the use of KRM and LLRM in Section 3.4 is symmetric. For each observation $i\,(i=1,...,N_1)$ in the treatment group with estimated propensity score $\hat{p}(x_i)=\hat{p}_i$, local regression matching considers all observations j

[30] Many researchers choose a fixed bandwidth. In Ham, Li and Reagan (2011) we compared local linear ridge regression matching estimates (Frölich 2004) with a fixed bandwidth chosen by cross-validation to our LLRM estimates and found that the results were very similar.

in the comparison group $j\,(j=1,...,N_0)$ with imputed propensity scores $\hat{p}(x_j)=\hat{p}_j$ that satisfy $\left|\hat{p}_i-\hat{p}_j\right|<h_i$, where h_i is the local bandwidth for individual i. Then local regression matching estimates $\mu(\hat{p}_i)=E\left[Y^0\mid D=1, p(X)=p_i\right]$. Within the local bandwidth h_i, the closer $\hat{p}_j$ is to $\hat{p}_i$, the greater the weight, given by the Kernel function $K\left(\frac{\hat{p}_j-\hat{p}_i}{h_i}\right)$, the observation j gets in estimating $\hat{\mu}(p_i)$. Specifically, local regression matching solves the following minimization problem for each treated individual i

$$\min_{\beta_0,\beta_1,...,\beta_L}\sum_{j=1}^{N_0}\left\{Y_j^0-\sum_{l=0}^{L}\beta_l\left(\hat{p}(x_j)-p_i\right)^l\right\}^2 K\left(\frac{\hat{p}(x_j)-p_i}{h}\right). \qquad \text{(A.1)}$$

and sets $\hat{\mu}(\hat{p}_i)=\hat{\beta}_0$. This procedure becomes KRM when we set $l=0$ and it becomes LLRM when we set $l=1$.

Appendix Table A: Adult Sample Size by Site and Insurance Plan

Insurance Plan	Dayton, Ohio site=1	Seattle, Washington site=2	Fitchburg, Massachusetts site=3	Franklin County, Massachusetts site=4	Charleston, South Carolina site=5	Georgetown County, South Carolina site=6	Total
1	85	0	0	0	0	0	85
2	80	0	0	0	0	0	80
3	62	0	0	0	0	0	62
4	70	0	0	0	0	0	70
5	66	43	7	13	19	20	168
6	63	27	14	22	12	29	167
7	59	29	7	12	24	18	149
8	51	0	0	0	0	0	51
9	42	0	0	0	0	0	42
10	46	0	39	44	21	40	190
11	243	334	185	216	197	263	1438
13	0	218	147	162	142	205	874
14	0	75	28	41	30	34	208
15	0	56	28	40	30	30	184
16	0	71	27	41	46	50	235
17	0	37	20	18	24	31	130
18	0	32	29	26	26	34	147
19	0	23	25	22	13	18	101
Total	867	945	556	657	584	772	4381

Notes:
1. The two HMO groups in Seattle are not included in this table.
2. The sample sizes represent the experiment assignment without considering data availability, non-compliance, and attrition.
3. Dayton was the pilot site and used different questionnaires.