

A **brief** introduction to Propensity Score Matching

March 2020

QS Update

When will this be useful?

- Assessing the [effects of kindergarten retention](#) on children's social-emotional development (Hong & Yu, 2008)
- [Effectiveness of Alcoholics Anonymous](#) (Ye & Kaskutas, 2009)
- [Effect of small school size](#) on mathematics achievement (Wyse, Keesler, & Schneider, 2008)

and of course..

- Studying the [impact of divorce](#) on families ☺

The problem with observational studies

If parallel universes exist, and we want to study the impact of divorce on financial outcomes of an individual, we can simply compare his finances in both worlds – one where he divorced, and where he did not divorce.

y_{0i} : financial status of individual i if he divorced

y_{1i} : financial status of individual i if he remains married

Effect for individual $i = y_{1i} - y_{0i}$

Average treatment effect = $E[Y_1 - Y_0] = \frac{1}{N} \sum y_{1i} - y_{0i}$

The problem with observational studies

Unfortunately, we can only observe one outcome per individual. There is no counterfactual to speak of.

In a large study, it is plausible that we can compare the means of the group which received treatment and the group which did not, and conclude that the difference is the effect of divorce.

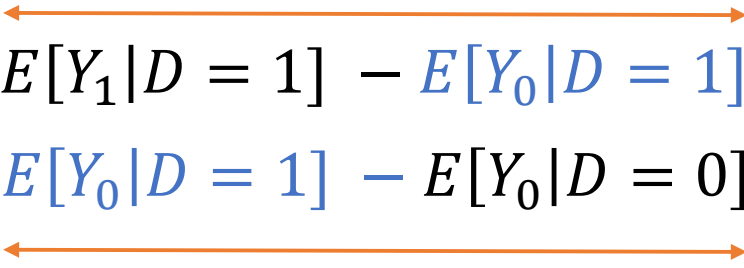
Really?

Potential outcomes framework

	Outcome if treatment is not received (y_0)	Outcome if treatment is received (y_1)
Individuals who do not receive treatment ($D = 0$)	$E[y_{i0} D = 0]$	$E[y_{i1} D = 0]$ (Not observed)
Individuals who receive treatment ($D = 1$)	$E[y_{i0} D = 1]$ (Not observed)	$E[y_{i1} D = 1]$

If we compare group means we are doing this:

$$\begin{aligned}
 E[Y_1|D = 1] - E[Y_0|D = 0] &= E[Y_1|D = 1] - E[Y_0|D = 1] \\
 &\quad + E[Y_0|D = 1] - E[Y_0|D = 0]
 \end{aligned}$$



Average treatment effect on the treated

Selection bias

Do divorcees and non-divorcees differ

because divorcees divorce

$$E[Y_1|D = 1] - E[Y_0|D = 0] = E[Y_1|D = 1] - E[Y_0|D = 1] \\ + E[Y_0|D = 1] - E[Y_0|D = 0]$$

*or because divorcees are divorcees,
and they fare differently in financial
outcomes compared to non-divorcees
to begin with?*

Randomized controlled trials are considered the gold standard for estimating effects of treatments

Because treatment assignment is random, and this eliminates any form of selection bias.

Intuition: If we are able to randomly choose people to get divorces (i.e. people have **no choice** in this case), when sample size is large enough, it is likely that the characteristics of divorcees and non-divorcees will be similar.


$$\begin{aligned} E[Y_1|D = 1] - E[Y_0|D = 0] &= E[Y_1|D = 1] - E[Y_0|D = 1] \\ &\quad + E[Y_0|D = 1] - E[Y_0|D = 0] \end{aligned}$$

Randomized controlled trials are considered the gold standard for estimating effects of treatments

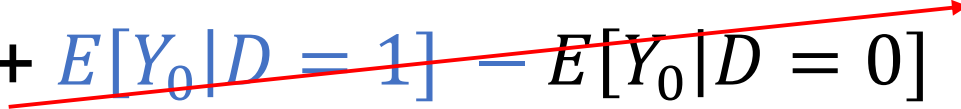
Because treatment assignment is random, and this eliminates any form of selection bias.

Intuition: If we are able to randomly choose people to get divorces (i.e. people have **no choice** in this case), when sample size is large enough, it is likely that the characteristics of divorcees and non-divorcees will be similar.

Average treatment effect on the treated



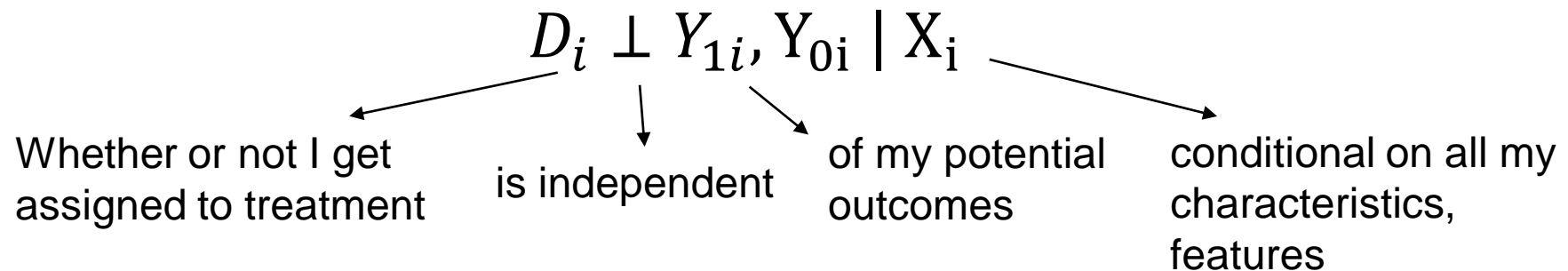
$$\begin{aligned} E[Y_1|D = 1] - E[Y_0|D = 0] &= E[Y_1|D = 1] - E[Y_0|D = 1] \\ &\quad + E[Y_0|D = 1] - E[Y_0|D = 0] \\ &= E[Y_1 - Y_0] \end{aligned}$$



We are, however, rarely in the world of randomized experiments with perfect compliance

The next best setting seems to be one in which **treatment is conditioned on observables**. What this means is that if I look at the pool of divorcees and non-divorcees, and I magically find a pair of guys named *Jane* and *Jane'* who are exactly the same in all aspects except for their choice to divorce, I can compare the outcomes of both and if they differ, I conclude that it is because of divorce.

Conditional Independence Assumption:



Conditional Independence Assumption

Let's move away from divorce to get a clearer example of what I mean. Assuming WF has to decide to send either Rory or Jovi to a Data Science workshop.

Y_{1J} : Jovi's competency if he gets sent to boot-camp.

Y_{0J} : Jovi's competency if he does not get sent to the boot-camp

Y_{1R} : Rory's competency if he gets sent to the boot-camp

Y_{0R} : Rory's competency if he does not get sent to the boot-camp

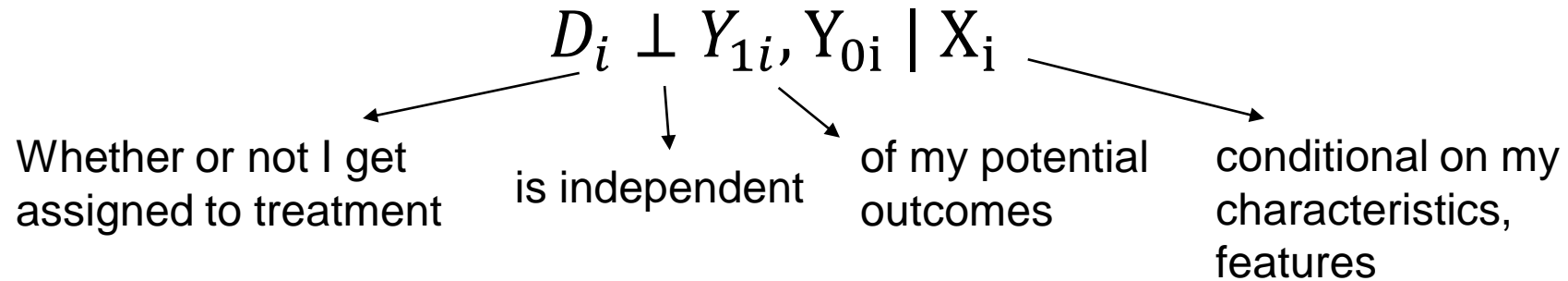
It is clear that $Y_{0R} \gg Y_{0J}$ and $Y_{1R} \gg Y_{1J}$. In this case, assignment treatment D_i is not independent of Y_{1i}, Y_{0i} since it is likely for Jovi to receive the treatment.

$$D_i \not\perp Y_{1i}, Y_{0i}$$

Conditional Independence Assumption

The goal of matching then, is to find someone that is as similar to Jovi as possible. Assuming we have the same characteristics ($X_{Jovi} = X_{Jovi'}$), we are likely to have the same potential outcomes ($Y_{1,Jovi} = Y_{1,Jovi'}$ and $Y_{0,Jovi} = Y_{0,Jovi'}$), and any treatment we receive is independent of our characteristics.

Conditional Independence Assumption:



Matching

It is clearly not possible to find exact matches. Matching can be done on appropriate covariate **distance metric**, e.g. Euclidean, Gower. If you do match based on a covariate distance metric, that you should scale the covariates before calculating the metric based on your a priori knowledge about the **relative importance of the covariates** in their effects on outcome.

As number of covariates increase, matching becomes **more and more difficult**.

Solution by Rosenbaum and Rubin (1983)

- Match on a **single index (propensity/balancing score)** which condenses all relevant information contained in covariates.
- **Propensity score theorem**: For given value of index, distribution of X should be the same for participants and non-participants.

$$D_i \perp Y_{1i}, Y_{0i} \mid \mathbf{X}_i \text{ implies } D_i \perp Y_{1i}, Y_{0i} \mid \mathbf{p}(\mathbf{X}_i)$$

- For those interested in the proof (using law of iterated expectations), I refer you to *Mostly Harmless Econometrics*. Or you can look at Ben Lambert videos on YouTube.

Solution by Rosenbaum and Rubin (1983)

- Propensity score is the probability of assignment to one treatment conditional on a subject's measured baseline covariates.

$$p(X_i) = P(D_i = 1 | X_i)$$

- Why this *theoretically* works? Two individuals who are equally likely to receive treatment are somewhat similar. There is no selection bias involve that increases one/or the other chances of being put into treatment.

PSM Steps

1. Estimation of propensity score
2. Check the assumptions: common support
3. Match participants with non-participants (& pruning)
4. Check the assumptions: covariates' balance
5. Compute the average treatment effect
6. Compute the standard error of the treatment effect

Step 1: Estimation of propensity score

- Use binary model to estimate γ_0 and γ_1 .

$$p(X_i) = P(D_i = 1 | X_i) = G(\gamma_0 + \gamma_1 X_i)$$

- $G(.)$ can be either probit or logit.
- Compute predicted values.

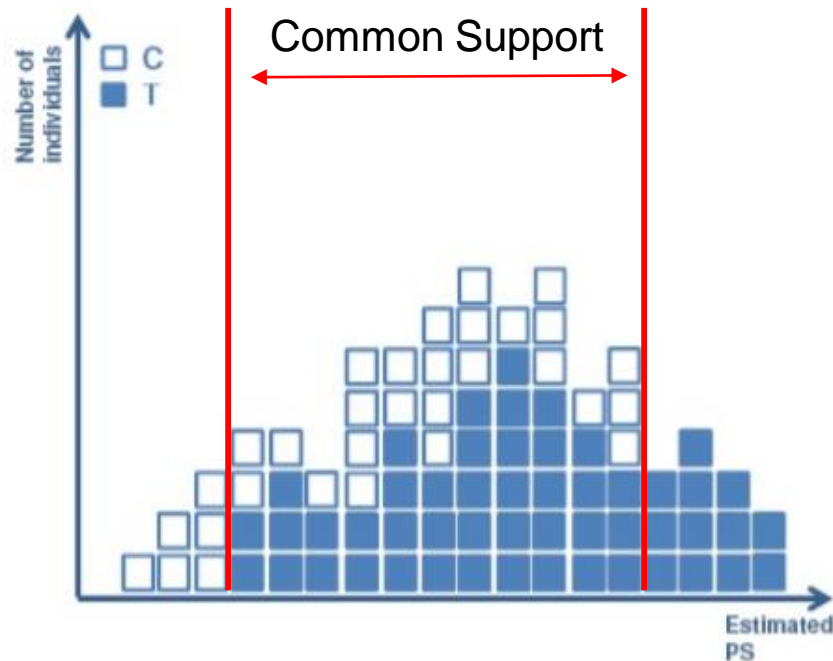
$$\hat{P}(D_i = 1 | X_i) = G(\hat{\gamma}_0 + \hat{\gamma}_1 X_i) = \hat{P}\hat{S}_i$$

What are the variables to include in my estimation of PSM?

- There is a lack of consensus in the literature.
- Possible sets of variables for inclusion in the propensity score model include the following: all measured baseline covariates, all baseline covariates that are associated with treatment assignment, all covariates that affect the outcome (i.e., the potential confounders), and all covariates that affect both treatment assignment and the outcome (i.e., the true confounders).
- **Not** post-baseline covariates that may be influenced or modified by the model.

Step 2: Check assumptions: common support

Hypothetical Example

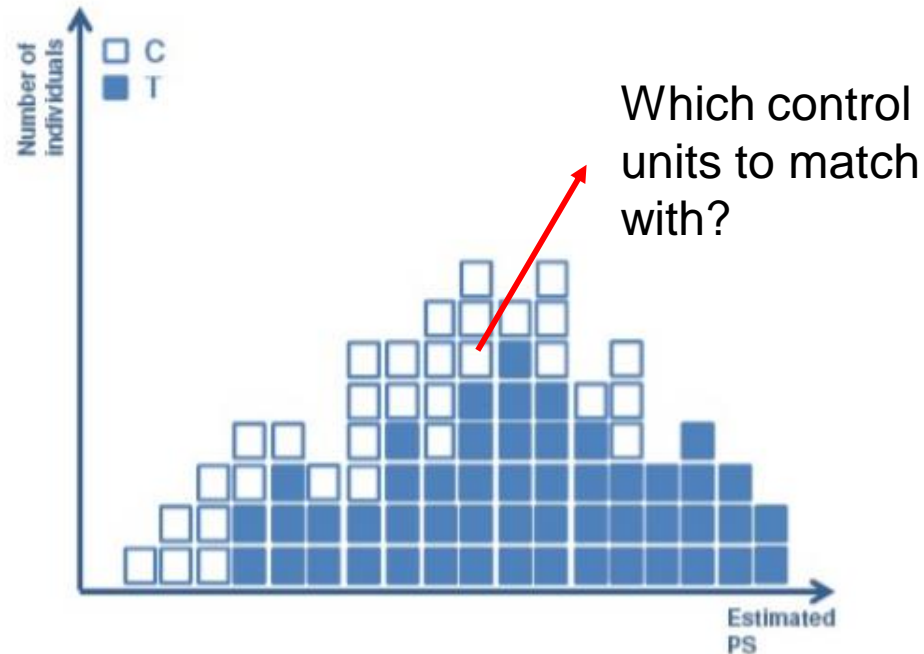


- Compare only similar individuals with similar propensity score.
- **Drop** treated units that have no units with similar PS in control group.

Note: If too many observations are dropped, there may be bias in the analysis because the remaining treated may not be representative of the whole treated group.

Step 3: Matching

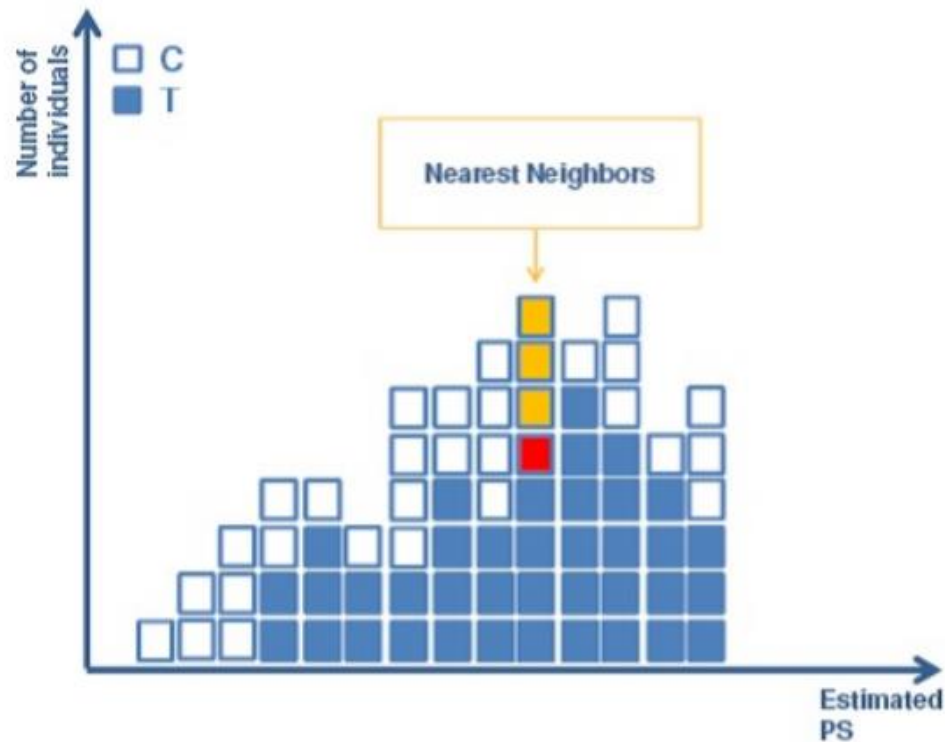
Hypothetical Example



- Matching is done **individually**.
- Have to decide:
 - Number of control units
 - Weight attributed to each control
 - Replacement, or without replacement

Step 3: Matching

Hypothetical Example

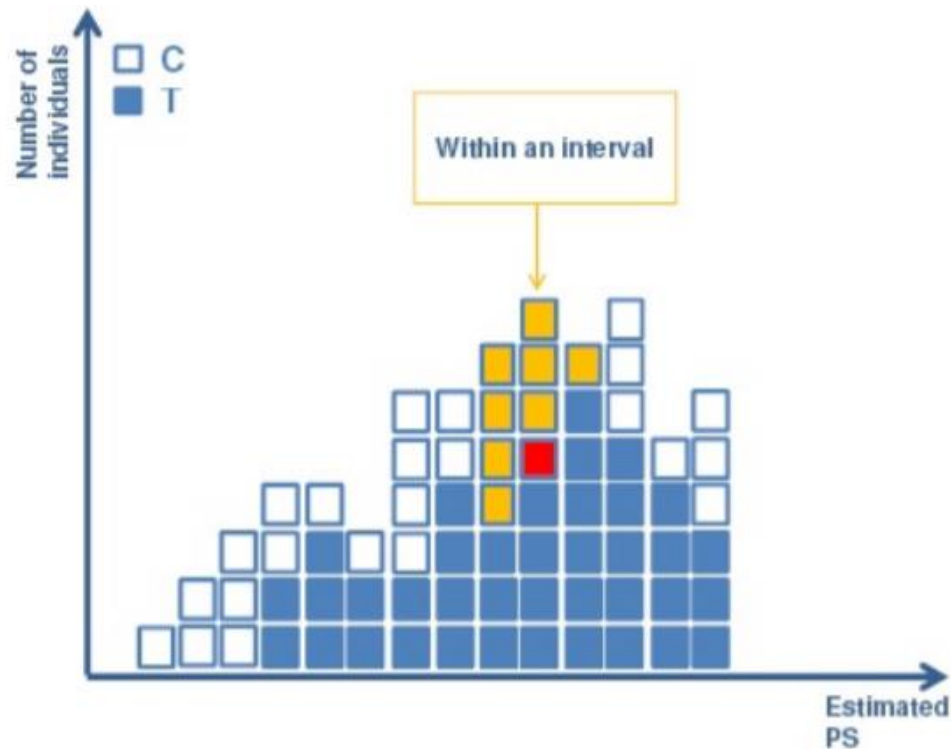


Nearest Neighbors

- Choose nearest neighbors control units.
- Have to decide:
 - Number of control units
 - Replacement, or without replacement

Step 3: Matching

Hypothetical Example



Radius Matching

- Choose controls within maximum interval/range.
- Have to decide:
 - Maximum distance (caliper width)
 - Replacement, or without replacement

Step 4: Check covariates' balance

- Checking if the distribution of all covariates is the same for participants and matched non-participants.
- If covariate balance is not satisfactory it may indicate a fundamental lack of comparability between the two groups ⇒ **alternative** evaluation approaches should be considered

Hypothetical example: t-test to compare age

	D = 1	D = 0	Difference
Before Matching	34.41	43.46	9.04***
After Matching	34.69	34.7	0

No difference in age
between treated and
matched control


Austin, 2011

.. conducted an extensive series of Monte Carlo simulations to determine the optimal caliper width for estimating differences in means (for continuous outcomes) and risk differences (for binary outcomes).

When estimating differences in means or risk differences, we recommend that researchers match on the logit of the propensity score using calipers of width equal to 0.2 of the standard deviation of the logit of the propensity score.

Step 5: Compute average treatment effect

- Treatment effect for treated i .

$$TT_i = Y_i - \sum_{j \in C_i} w_{ij} Y_j$$


Outcome of control, can be **equal weights**
(Radius matching), or **inverse to distance from**
treated (Kernel matching)

- SATT: Sample average treatment effect on treated. Simply just an average of the above.
- FSATT: Feasible SATT (prune badly matched treatment observations)

However ...

- There has been some studies which argued against the use of PSM.
- In that PSM sometimes achieves the opposite of what it seeks to achieve – increasing imbalance, model dependency, bias.
- Does not mean propensity score is useless – they can be used for other purposes such as regression adjustment, inverse weighting, stratification.
- I refer you to this brilliant lecture *Why Propensity Scores should not be used for matching* by Gary King (King & Nielson, 2016).

PSM is not as powerful as other matching methods

Types of Experiments

Covariates Balance?	Complete Randomization	Fully Blocked
Observed	On average	Exact
Unobserved	On average	On average

- Fully blocked methods dominates complete randomization for imbalance, model dependence, power, efficiency, bias

Goal of Matching Method (in observational data)

- PSM: Complete randomization
- Other methods: Fully blocked
- Other matching methods **dominate** PSM.

Coarsened Exact Matching

Approximates Fully Blocked Experiment

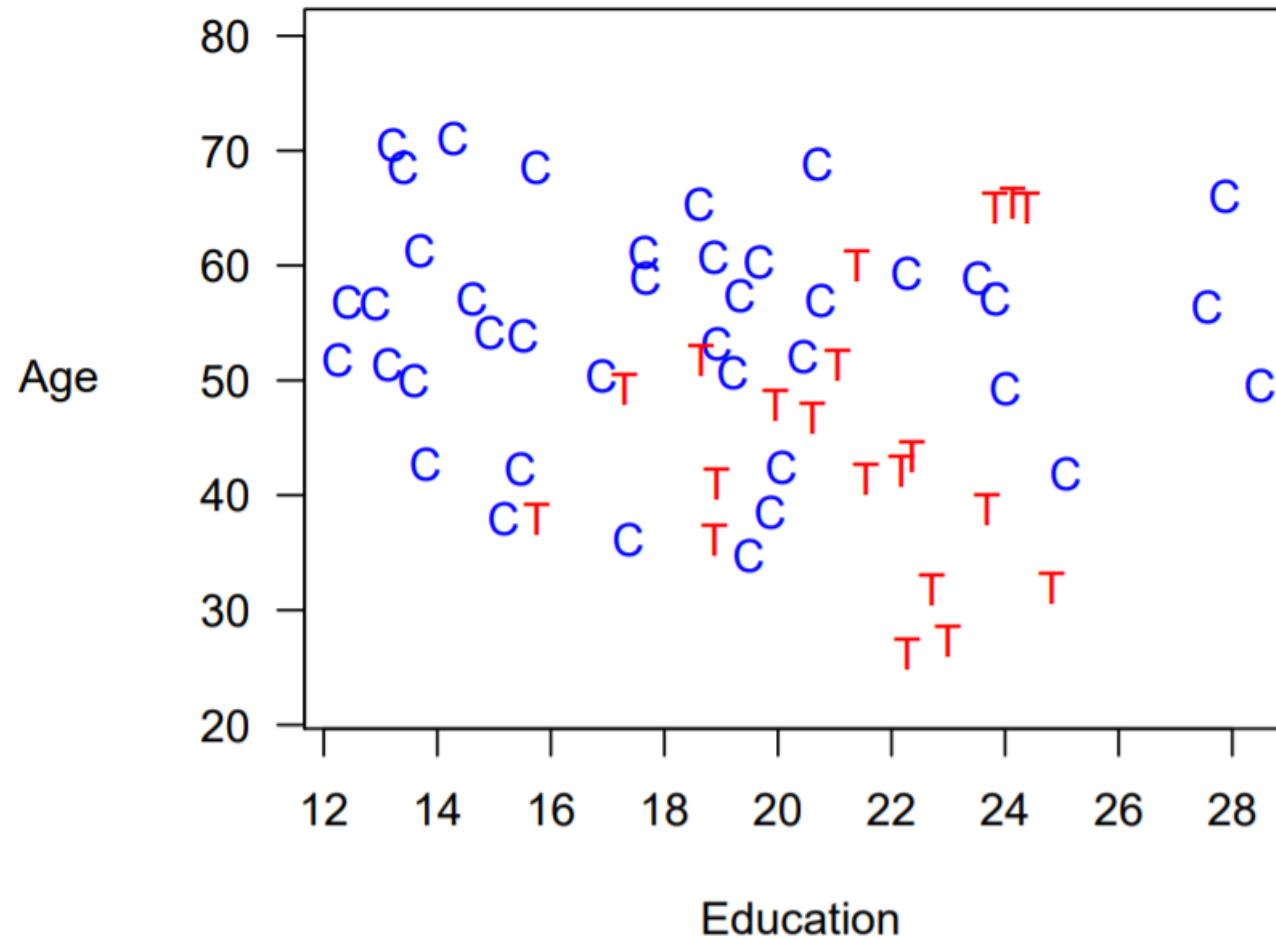
1) Pre-process (Matching)

- Temporarily coarsen X
 - E.g. Education (primary, secondary, high school, college)
- Apply exact matching to coarsened X
 - Sort observations into strata

2) Estimation using difference in means or other models

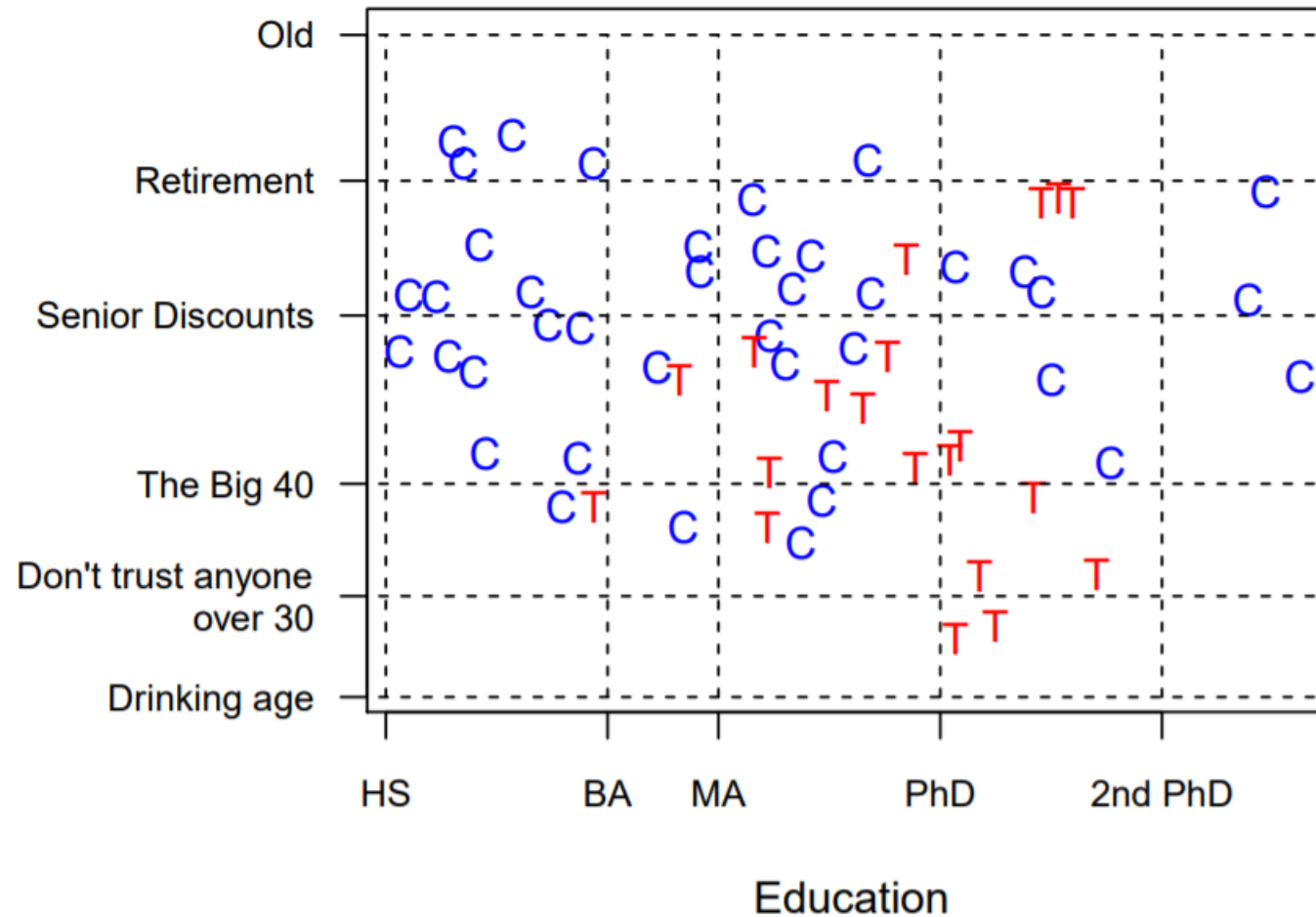
Coarsened Exact Matching

Approximates Fully Blocked Experiment



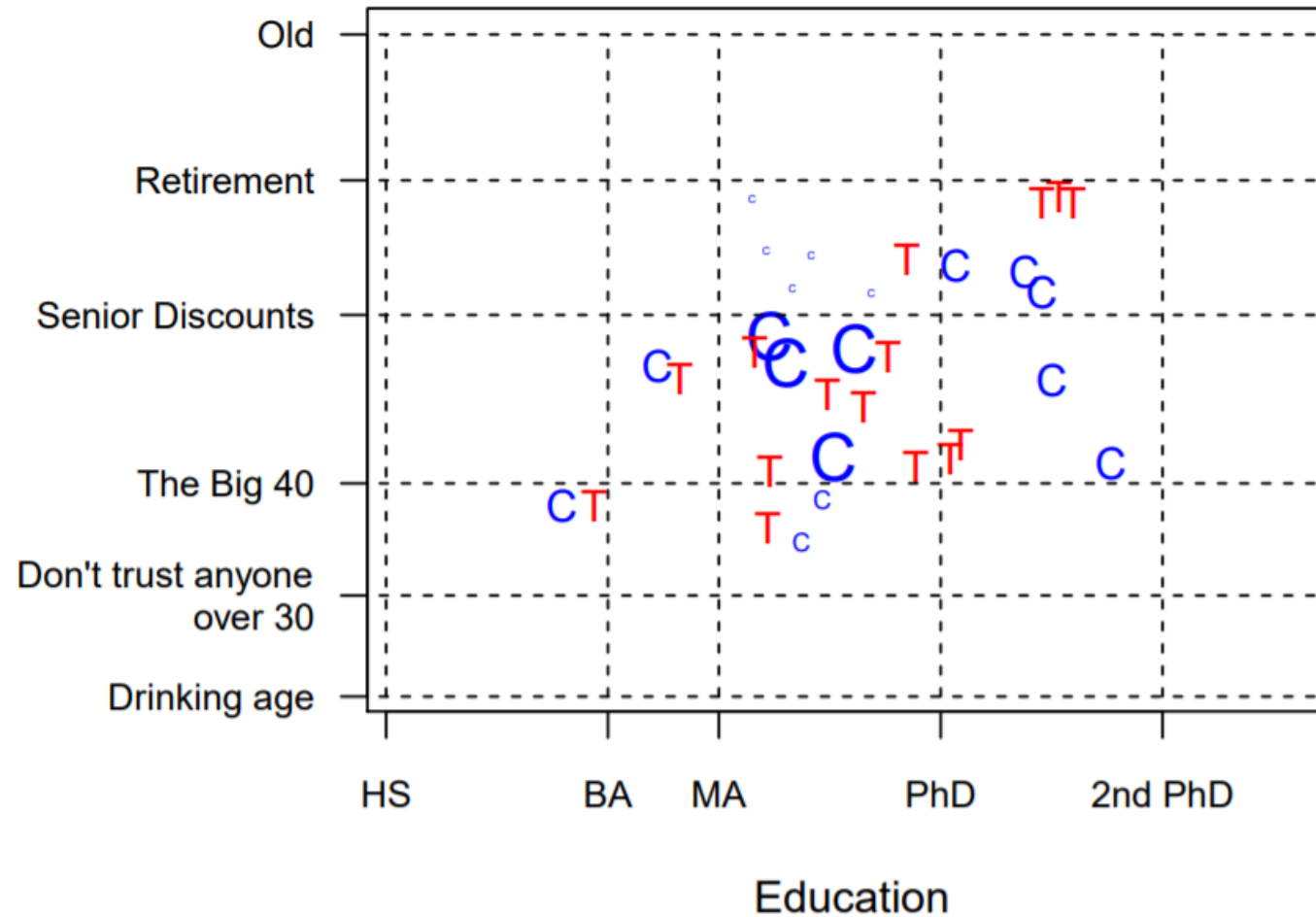
Coarsened Exact Matching

Approximates Fully Blocked Experiment



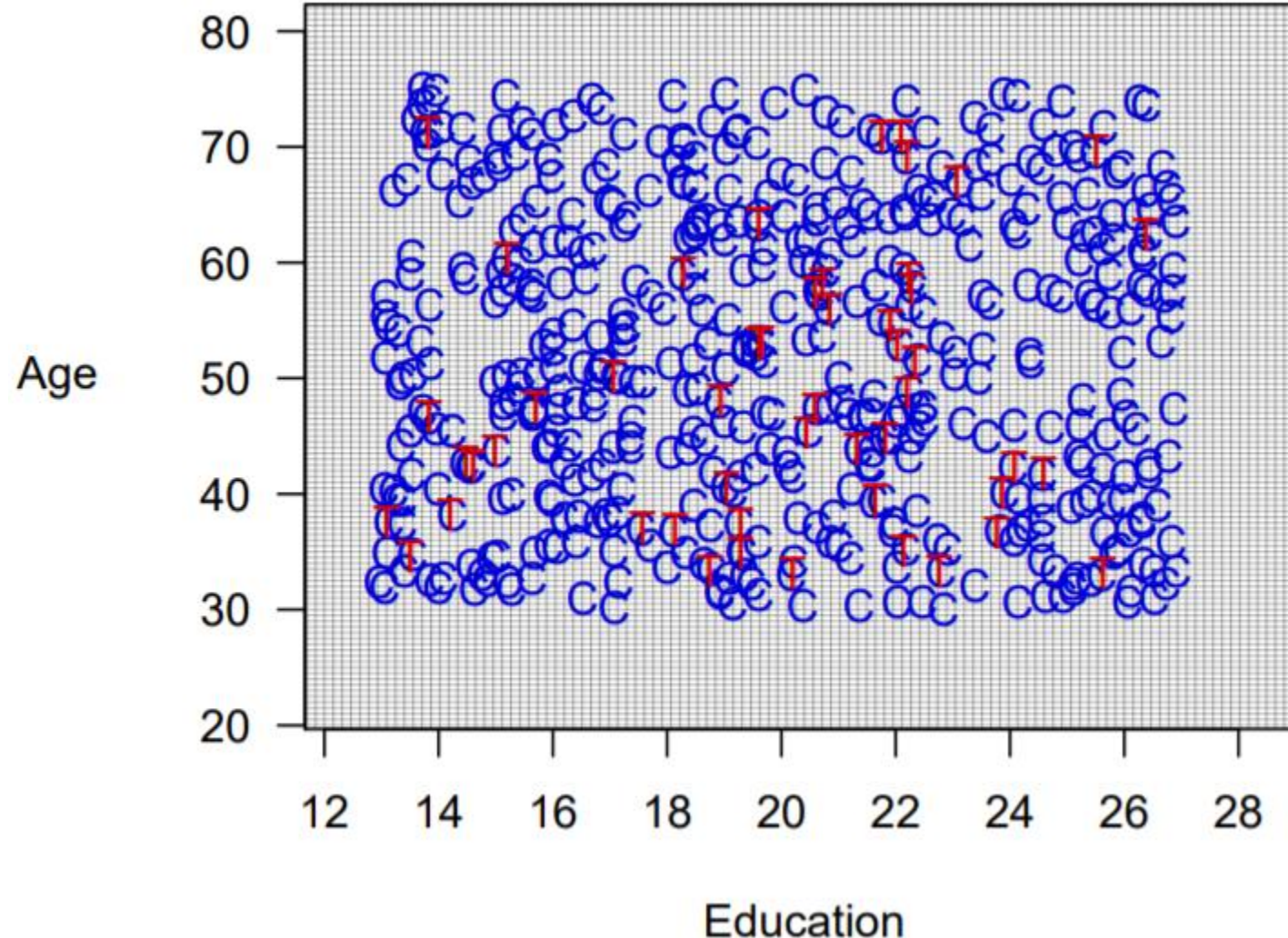
Approximates Fully Blocked Experiment

Approximates Fully Blocked Experiment



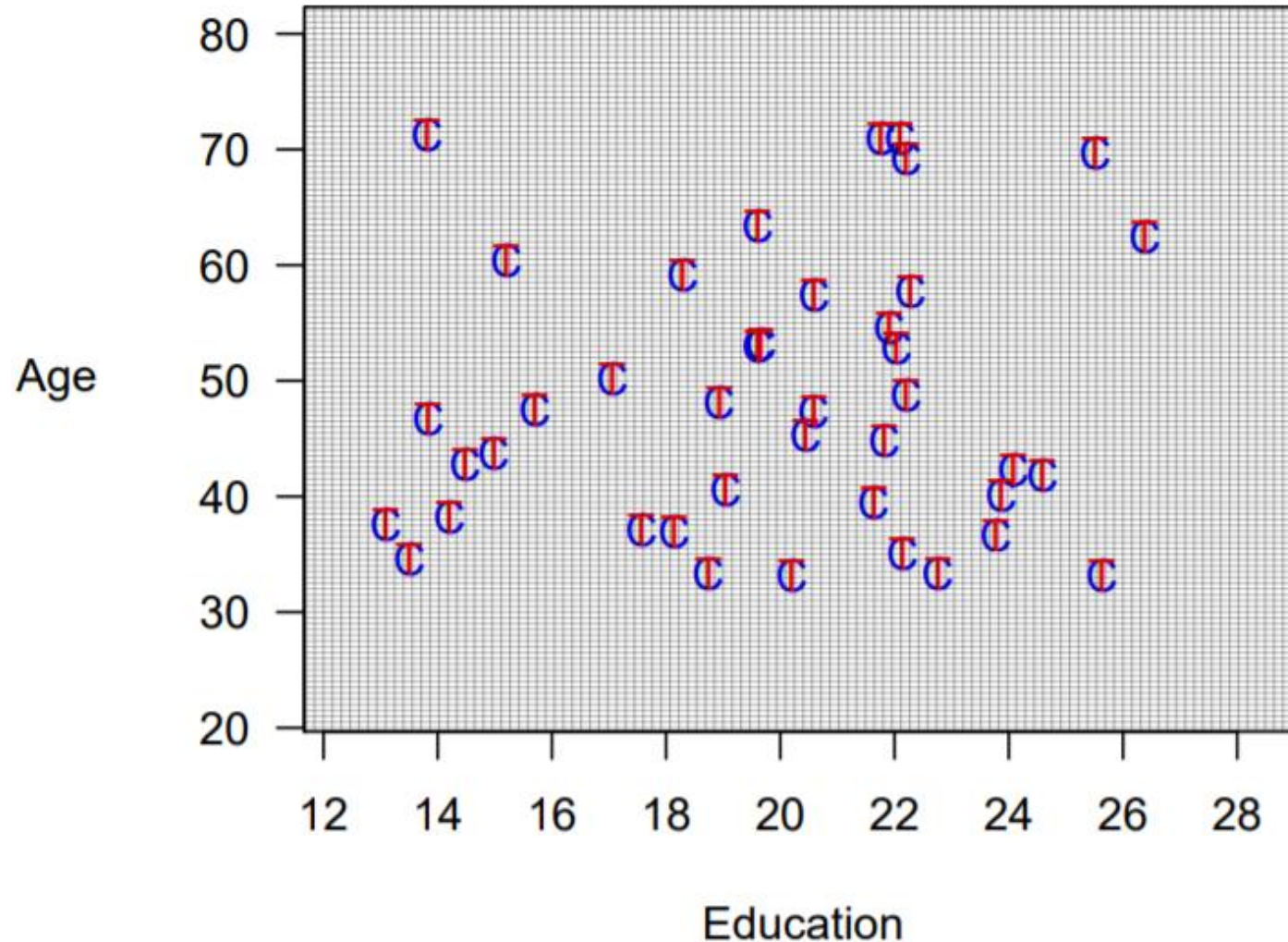
Coarsened Exact Matching – Best Case

Approximates Fully Blocked Experiment



Coarsened Exact Matching – Best Case

Approximates Fully Blocked Experiment



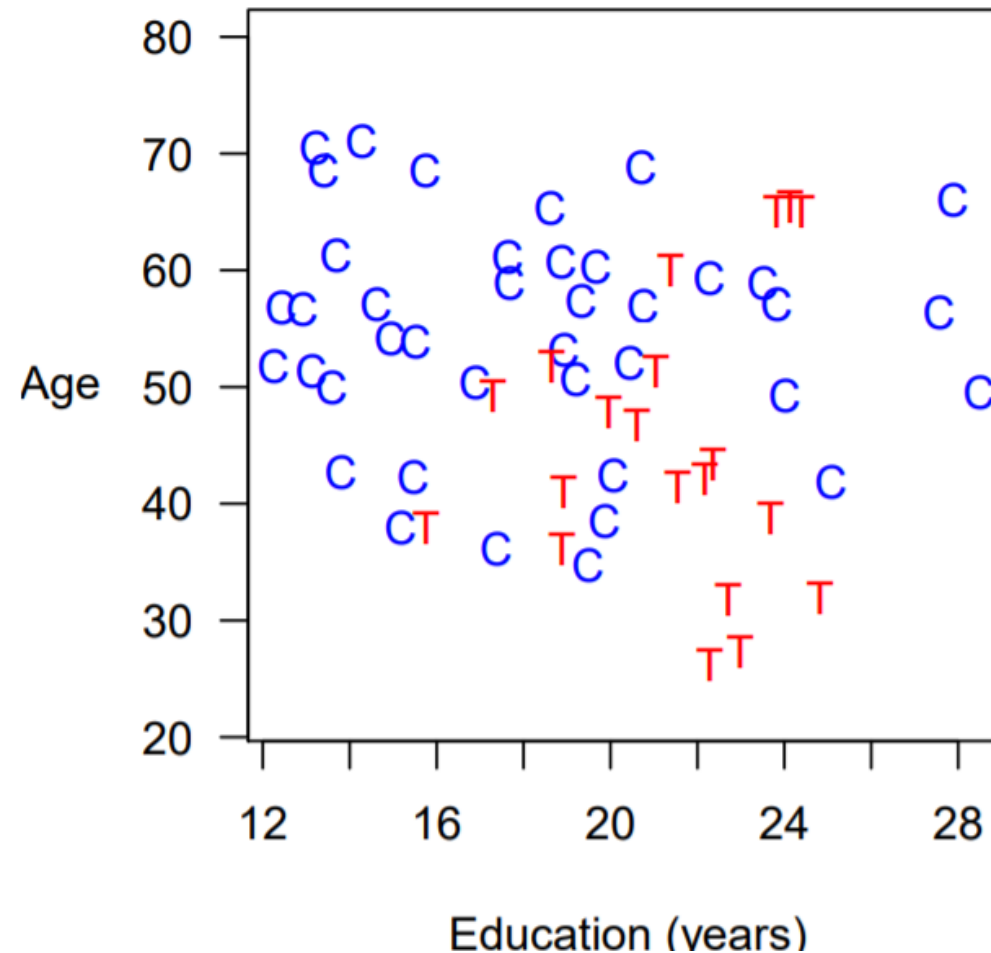
Propensity Score Matching

Approximates Completely Randomized Experiment

- 1) Pre-process (Matching)
 - Estimate propensity score
 - Match each treated unit to nearest control unit
- 2) Estimation using difference in means or other models

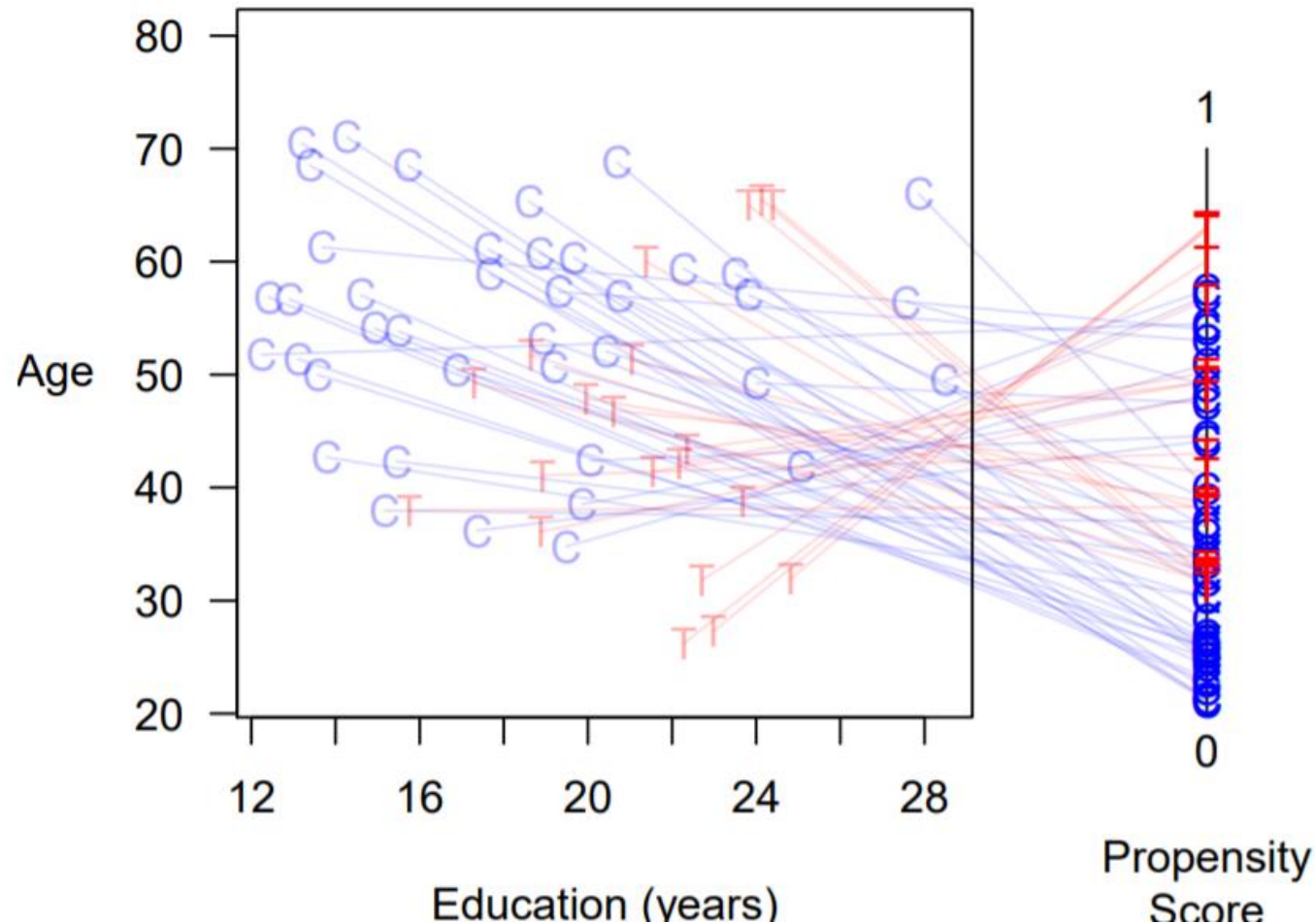
Propensity Score Matching

Approximates Completely Randomized Experiment



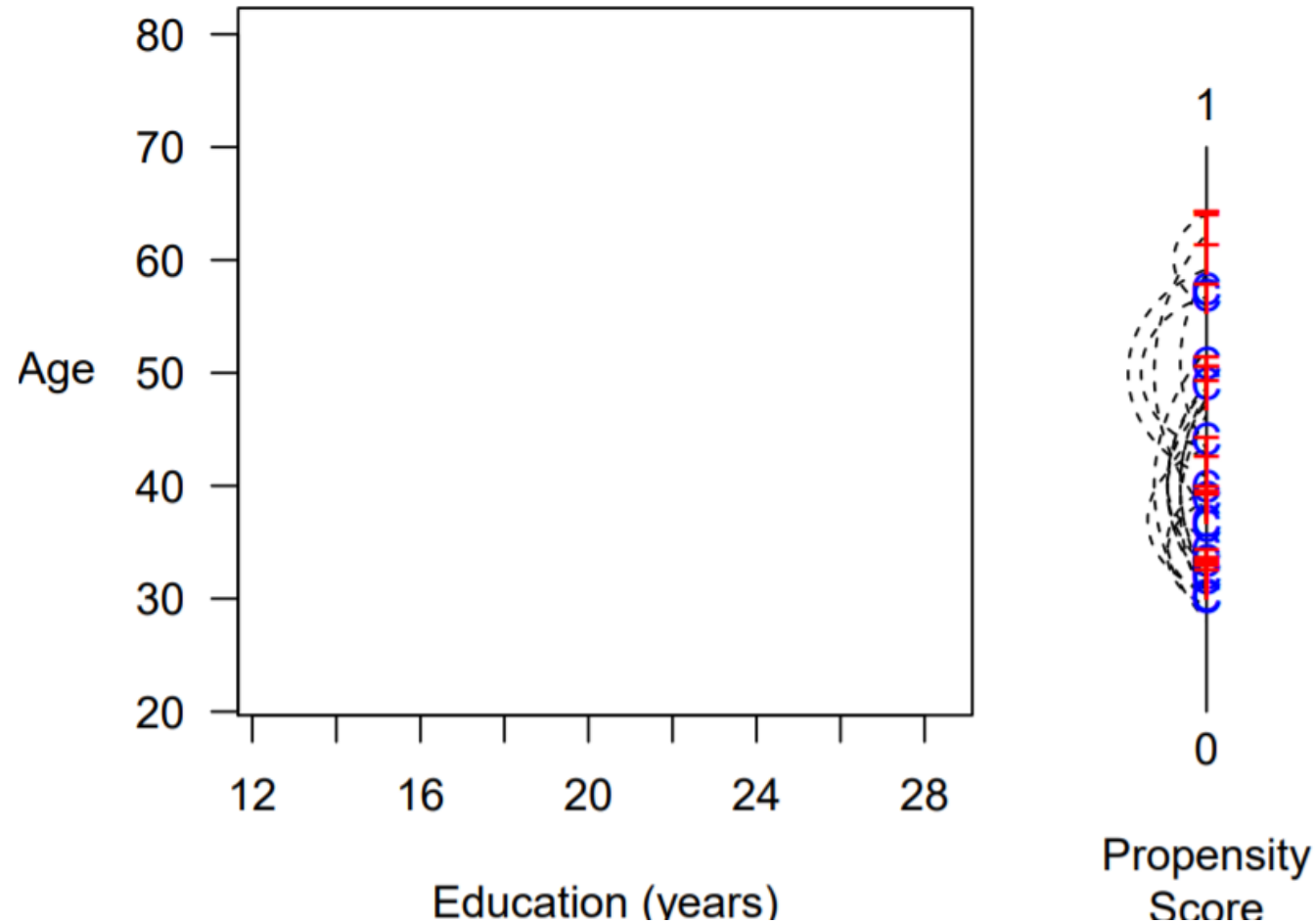
Propensity Score Matching

Approximates Completely Randomized Experiment



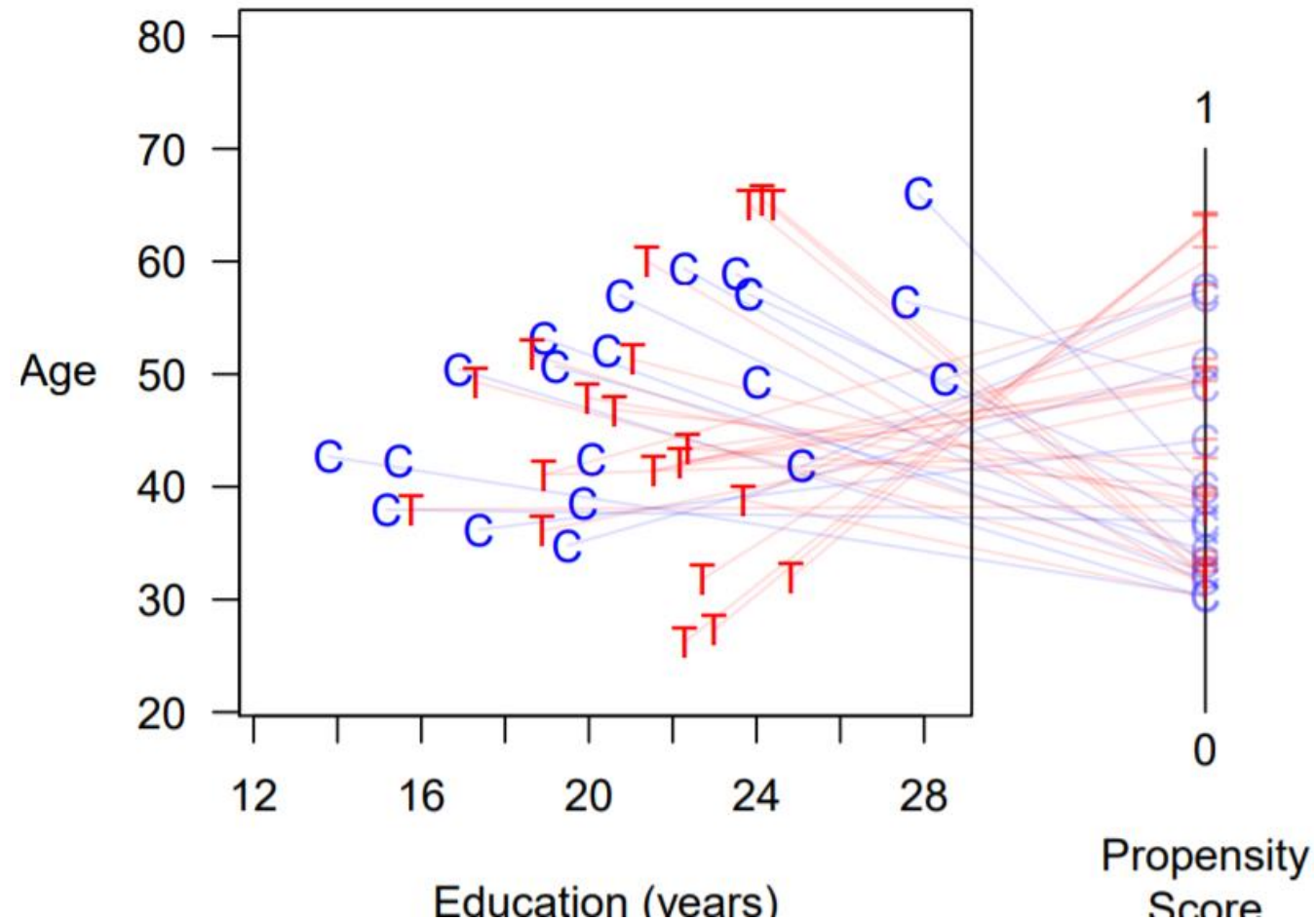
Propensity Score Matching

Approximates Completely Randomized Experiment



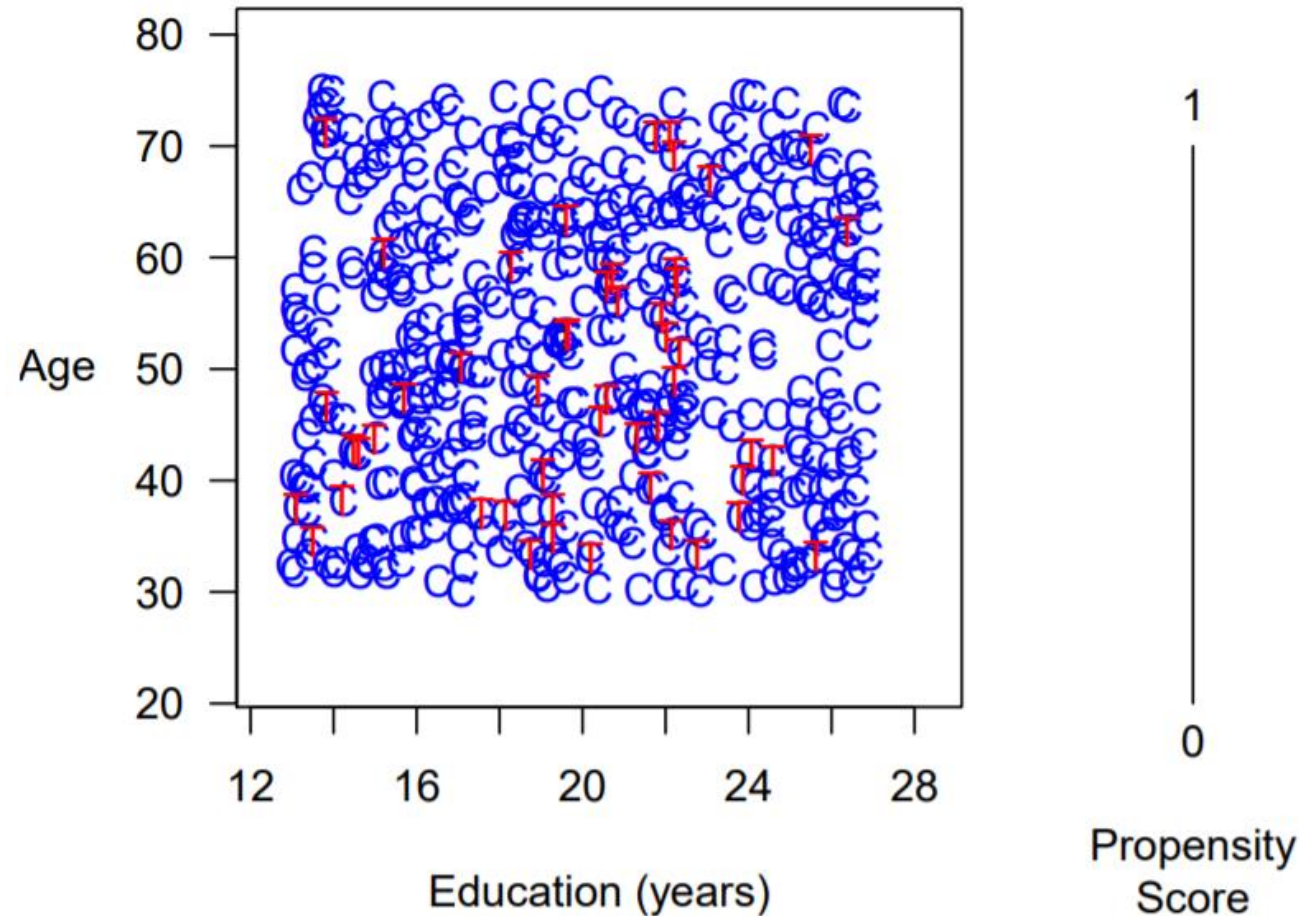
Propensity Score Matching

Approximates Completely Randomized Experiment



Propensity Score Matching: Best Case

Approximates Completely Randomized Experiment



Propensity Score Matching: Best Case

Approximates Completely Randomized Experiment

