Vaccine Effectiveness Analysis on a Balanced Synthetic Cohort:

A NON-EXPERIMENTAL APPROACH TO CASUAL INFERENCE

Miss. Avishag Nevo
Data Science student, Technion
assisted by Prof. Yair Goldberg

OUTLINE

- **INTRODUCTION**

- **OVERVIEW ON PROPENSITY SCORE MATCHING**
  - Introduction
  - Calculating propencity scores
  - Matching
  - Balance diagnostic
  - Estimating treatment effect – sevral approches
  - Tips for practice

- **DATA**
  - Generating synthetic data
  - Calculating propensity scores

- **STUDY DESIGN**
  - Patient follow up
  - Matching
  - Balance checkinhg

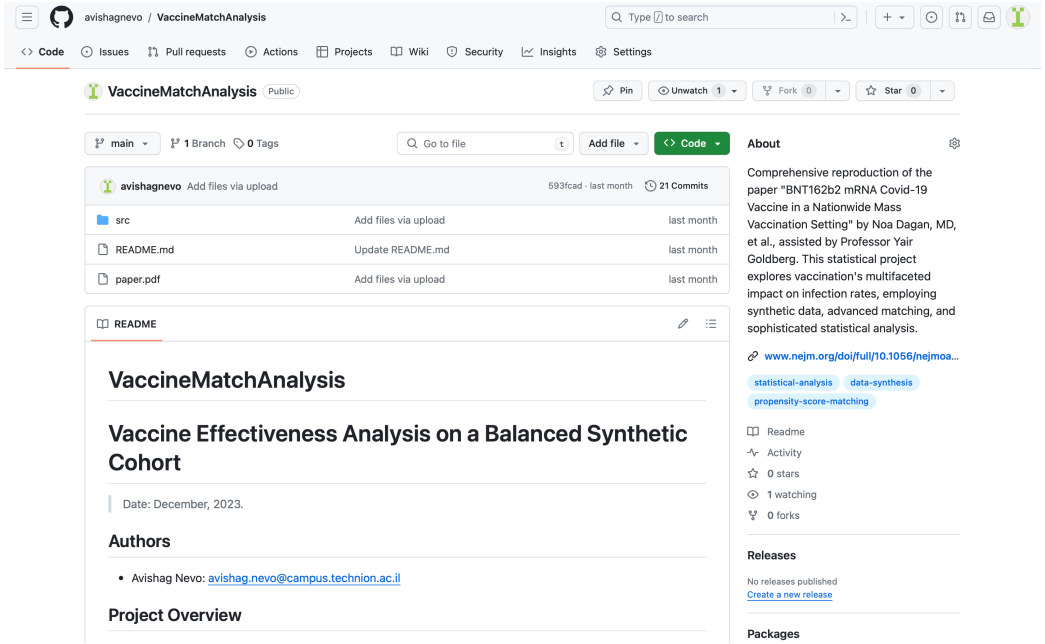
- **STATISTICAL ANALYSIS**
  - Kaplan Meier survival curves
  - Vaccine effectivmess dynamics

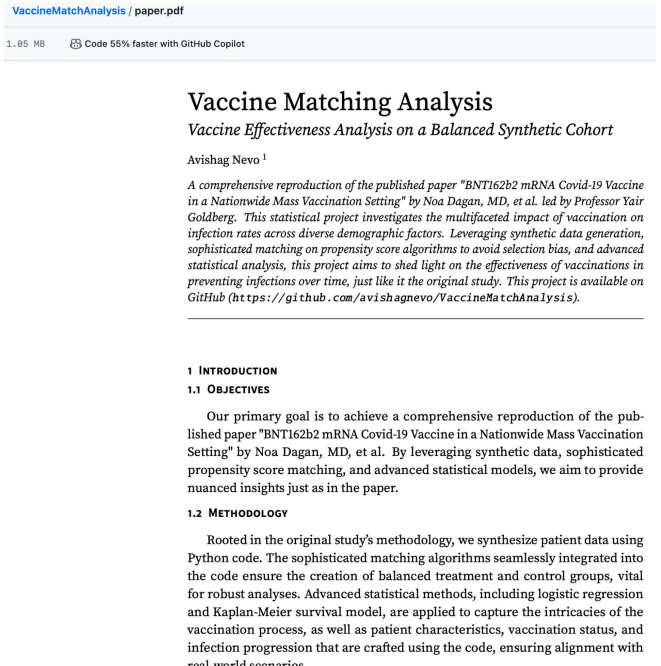
- **RESOULTS**
- **LITERATURE**

INTRODUCTION

AT HIGH LEVEL:

- Based on the study "BNT162b2 mRNA Covid-19 Vaccine in a Nationwide Mass Vaccination Setting."
- Our primary goal is to reproduce the study's key insights into vaccine effectiveness at different time intervals.
- Comprehensive analysis of vaccine effectiveness, leveraging synthetic data and advanced statistical techniques.
- Code and full paper available on GitHub: https://github.com/avishagnevo/VaccineMatchAnalysis





OVERVIEW ON PROPENSITY SCORE MATCHING

- **INTRODUCTION TO PSM**

- **CALCULATING PROPENSITY SCORES**
  - Building the Model
  - Covariate selection

- **MATCHING ON SCORES**
  - Common algorithm
  - Balance diagnostics

- **ESTIMATING T_x EFFECT**
  - Survey of approaches
  - Variance Estimation

- **Guidence for Practice**

INTRODUCTION TO PSM

- **Quasi-experimental** method for 'finding' a control group when random assignment is not possible
  - Use the propensity to be exposed as a proxy for assignment

- Based on the requirement for groups to be **"balanced"**– equivalent on average
  - Make exposure independent of observed and unobserved characteristics (covariates)

- Best suited to deal with **selection bias**
  - Cannot necessarily eliminate other confounders

INTRODUCTION TO PSM

Key Steps:

1. **Calculate the propensity:** Build a model predicting the probability of exposure (T_x) for both exposed and unexposed

2. **Use the propensity score** to create a control group matched to the exposed group

3. **Check** that the exposed group and (matched) control group are balanced

4. **Estimate** effect of exposure on the outcome of interest

PROPENSITY SCORE CALCULATION

- **Propensity score**

$$\Pr[Z_i = 1 \mid X_i]$$

  - Z_i = treatment index ($T = 1\ , C = 0$)
  - X_i = covarietes i.e., features, backround variables etc.

- Probability of being **exposed** given a set predictors

- PSs are used to find unexposed units that are the most similar to the exposed units

PROPENSITY SCORE CALCULATION

- **GOALS:**
  a. Build a model of the treatment assignment process
  b. Condence pre-treatment backround information down to a single metric
  c. Optimize balance between groups
     - Make treatment assignment **"strongly ignorable"** $T \perp (Y_0, Y_1)|X$

- We are **NOT** building the most predictive model to generalize to unseen data

- **Key concern:** properly capturing functional form of relationship between predictors and exposere
  - Are there interactionas or non-linear relationships ?

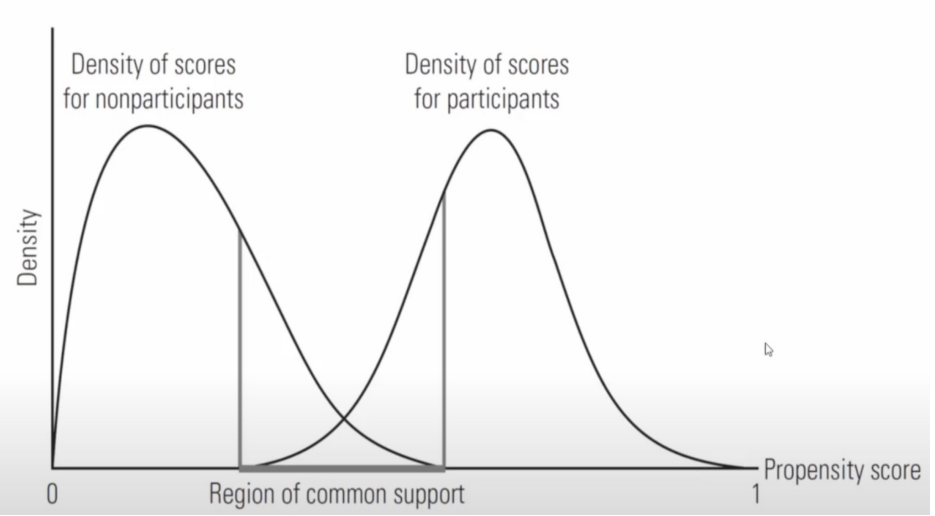
- **Logistic regression** typically used
  - Some limited research on ML models (RBF and boosting)

$$PS_i = \frac{e^{\beta^T X_i}}{1 + e^{\beta^T X_i}}$$

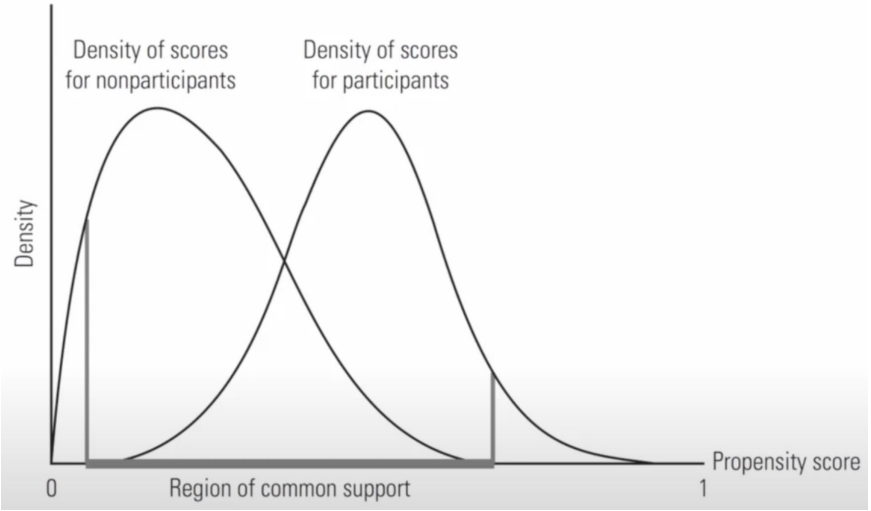
PROPENSITY SCORE CALCULATION

Common Support

- Explicitly **restrict** the analysis to individuals in the rigion of common support
- Examing the **convex hall** of the corvarietes to idenify the multidimentional space that allows **interpulation rether than extrapulation**
- **ATT** estimation may be fine while **ATE** may fail *



Extrapolation – ideal for machine leaning



Interpolation – ideal for propensity score

SELECTING COVARIATES

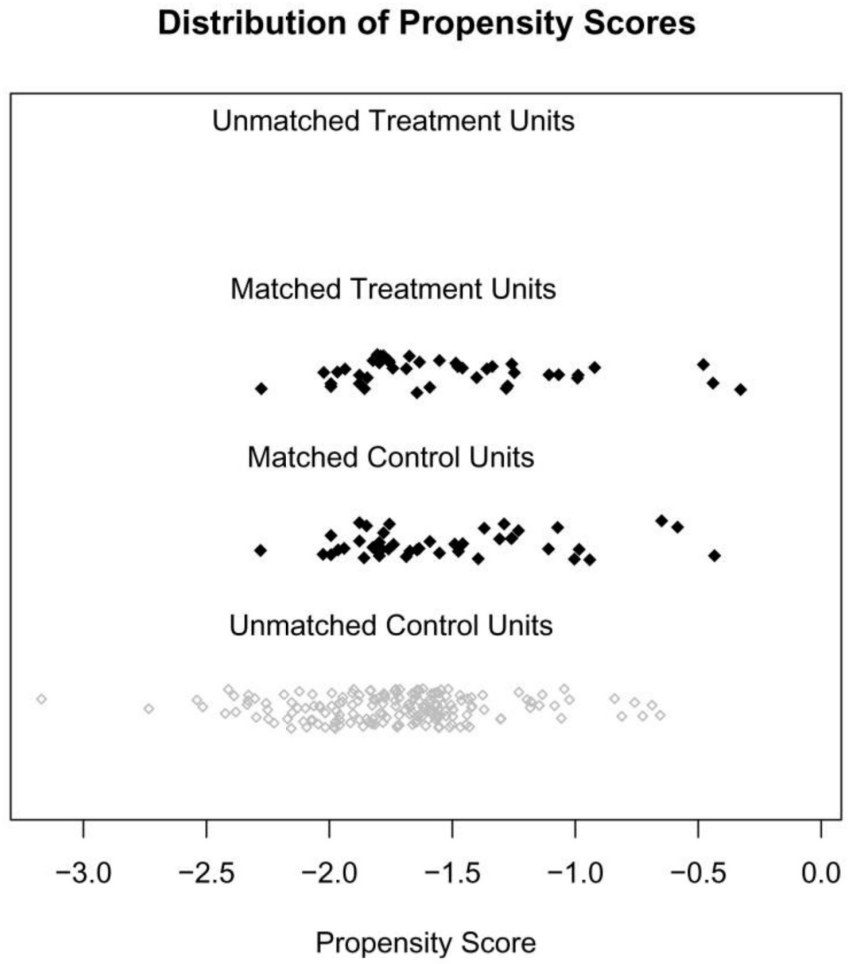
- Predictors of exposure should be related to **selection process**
  - Look for 'true' confouders, avoid instrumental variables*
  - Omitting key variables can introduce bias and increase variance
- Use a **theory** of selection process
  - Pre-exposure values on outcome
  - Past behavior, demographics, attitudes, geography
- Find, include **proxies** for unobserved variables
  - 3rd party data
- Over-including covarietes **usually** doesn't hurt
  - We aren't worried about overfitting
  - Can regularize, use variable selection, caliberate probabilities*

MATCHING

Matching algorithm dimentions:

- Partial data - Nearest Neighbor
  - *Greediness*: greedy vs. optimal
  - *Distance*: caliper & ratio matching
  - *Symmetry:* one-to-one vs. one-to-many
  - *Replacement:* with vs. without

- Whole data
  - Subclassification
  - Full matching
  - Weightning adjustment



Matches chosen using 1:1 nearest neighbor matching on propensity score. Black dots indicate matched individuals; grey unmatched individuals.

MATCHING-FINDING THE CONTROL GROUP

Matching algorithm dimentions:

- *Greediness*: greedy vs. optimal

PS_T	PS_C
0.45	0.72
0.38	**0.44**
0.41	0.60
	0.63
	0.35
	0.65
	0.47
	0.49
	0.36

1NN greedy matching 0.01+0.02+0.06=0.09

PS_T	PS_C
0.45	0.72
0.38	**0.44**
0.41	0.60
	0.63
	0.35
	0.65
	0.47
	0.49
	0.36

1NN optimal matching 0.02+0.02+0.03=0.07

MATCHING-FINDING THE CONTROL GROUP

Matching algorithm dimentions:

- *Distance*: caliper & ratio matching

PS_T	PS_C
0.45	0.72
0.38	**0.44**
0.90	0.60
	0.63
	0.35
	0.65
	0.47
	0.49
	0.36

1NN caliper (o.1) matching 0.01+0.02=0.03

$\frac{2N_T}{3}$ matches

PS_T	PS_C
0.45	**0.72**
0.38	**0.44**
0.90	0.60
	0.63
	0.35
	0.65
	0.46
	0.49
	0.36

Variable ratio matching 0.02+0+0.18=0.20

N_T matches

MATCHING-FINDING THE CONTROL GROUP

Matching algorithm dimentions:

- *Symmetry:* one-to-one vs. one-to-many

PS_T	PS_C
0.45	**0.72**
0.38	**0.44**
0.80	0.60
	0.88
	0.35
	0.65
	0.47
	0.49
	0.36

One-to-one matching
0.01+0.02+0.08=0.11

PS_T	PS_C
0.45	**0.72**
0.38	**0.44**
0.80	0.60
	0.88
	0.35
	0.65
	0.47
	0.49
	0.36

Many-to-one (2) matching
(AVG distance) 0.005+0.025+0=0.03

MATCHING-FINDING THE CONTROL GROUP

Matching algorithm dimentions:

- *Replacement:* with vs. without

PS_T	PS_C
0.45	0.72
0.48	**0.44**
0.40	0.60
	0.88
	0.35
	0.65
	0.55
	0.52
	0.36

With replacement 0.01+0.04+0.04=0.09

$\frac{1N_C}{3}$ matches

PS_T	PS_C
0.45	0.72
0.48	**0.44**
0.40	0.60
	0.88
	0.35
	0.65
	0.55
	0.52
	0.36

Without replacement 0.01+0.04+0.05=0.10

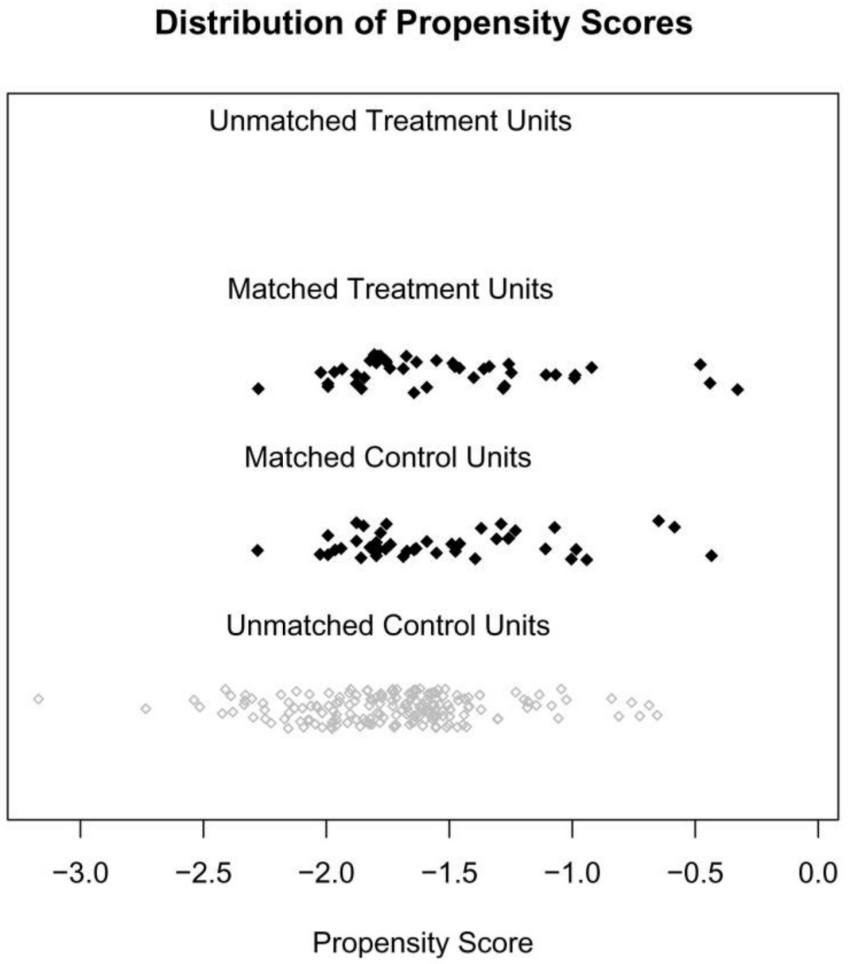
N_c matches

MATCHING

Matching algorithm dimentions:

- Partial data - Nearest Neighbor
  - *Greediness*: greedy vs. optimal
  - *Distance*: caliper & ratio matching
  - *Symmetry:* one-to-one vs. one-to-many
  - *Replacement:* with vs. without

- Whole data
  - Subclassification
  - Full matching
  - Weightning adjustment
    - Inverse prababilty of treatment weightning (IPTW)
    - Odds weightning
    - Kernel weightning

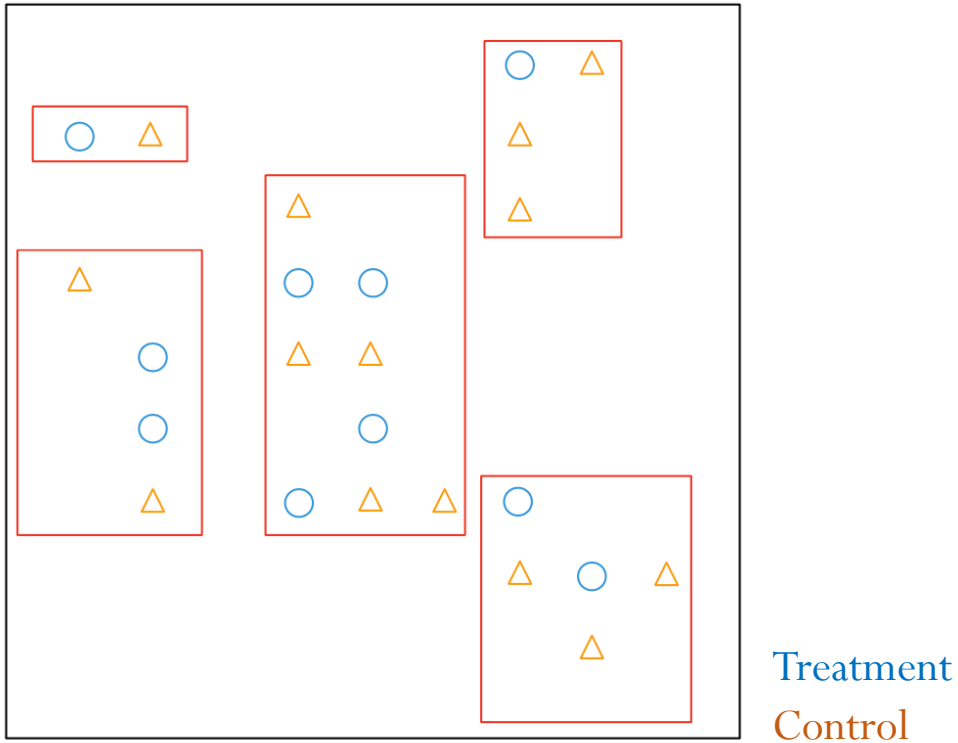


Matches chosen using 1:1 nearest neighbor matching on propensity score. Black dots indicate matched individuals; grey unmatched individuals.

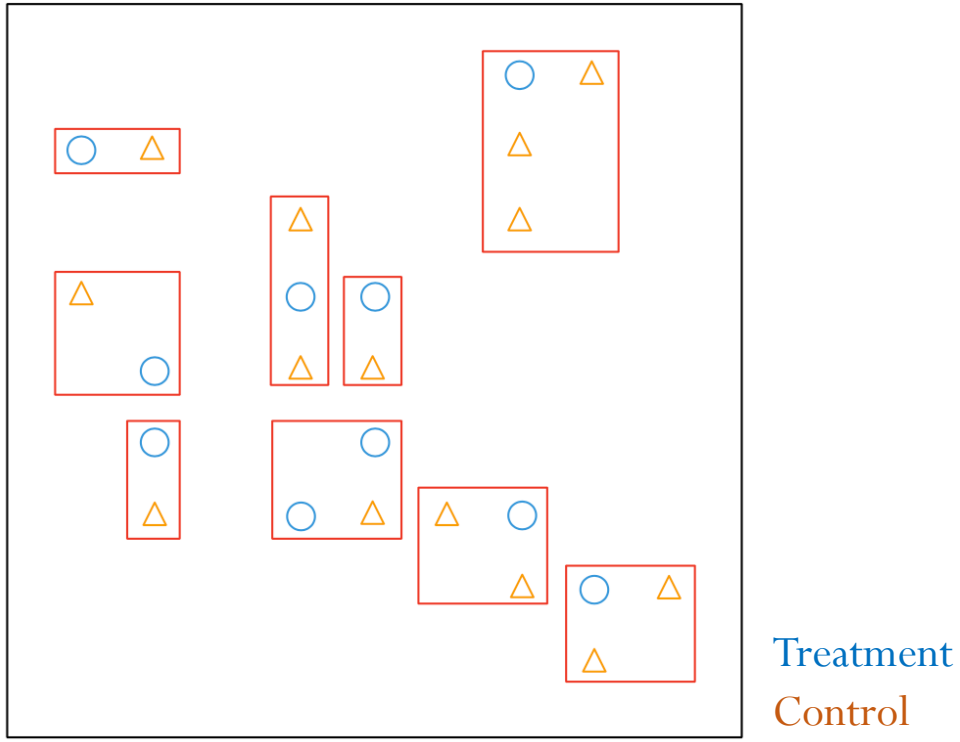
MATCHING-FINDING THE CONTROL GROUP

Matching algorithm:

- Subclassification vs. full matching



Subclassification (5)



Full matching

MATCHING-FINDING THE CONTROL GROUP

Matching algorithm:

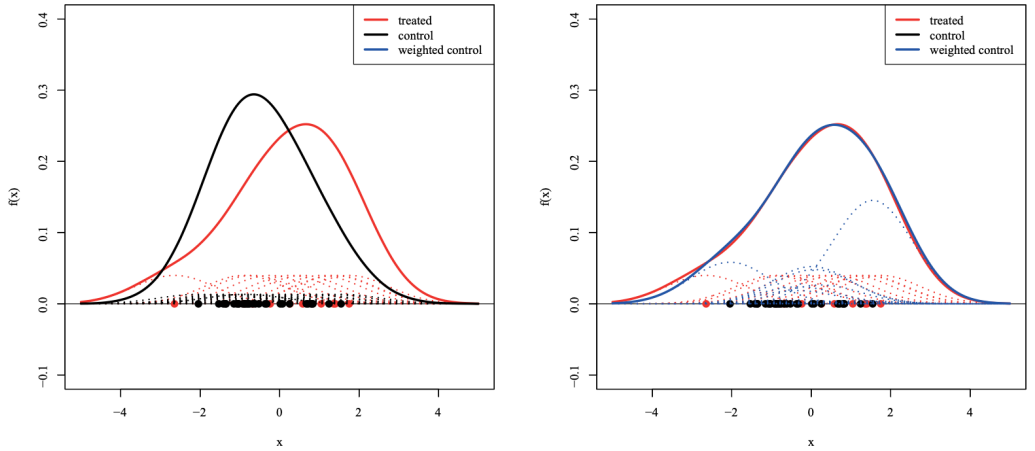
- Weightning adjustment
  - Inverse prababilty of treatment weightning
  - Odds weightning
  - Kernel weightning *

$$w_i = \frac{T_i}{\widehat{PS}_i} + \frac{1 - T_i}{1 - \widehat{PS}_i}$$

Inverse Probability

$$w_i = T_i + (1 - T_i)\frac{PS_i}{1 - \widehat{PS}_i}$$

Odds



Left: Density estimates for treated and (unweighted) controls. Red dots show the location of 10 treated units. Dashed lines show the appropriately scaled Gaussian over each observation, which sum to form the density estimator for the treated (red line) and control (black line). The L_1 imbalance is measured to be 0.32. *Right:* Weights chosen by kernel balancing effectively rescale the height of the Gaussian over each control observation (dashed blue lines). The new density estimate for the weighted controls (solid blue line) now closely matches the density of the treated at each point. The L_1 imbalance is now measured to be 0.002

Kernel

CHECKING FOR BALANCE

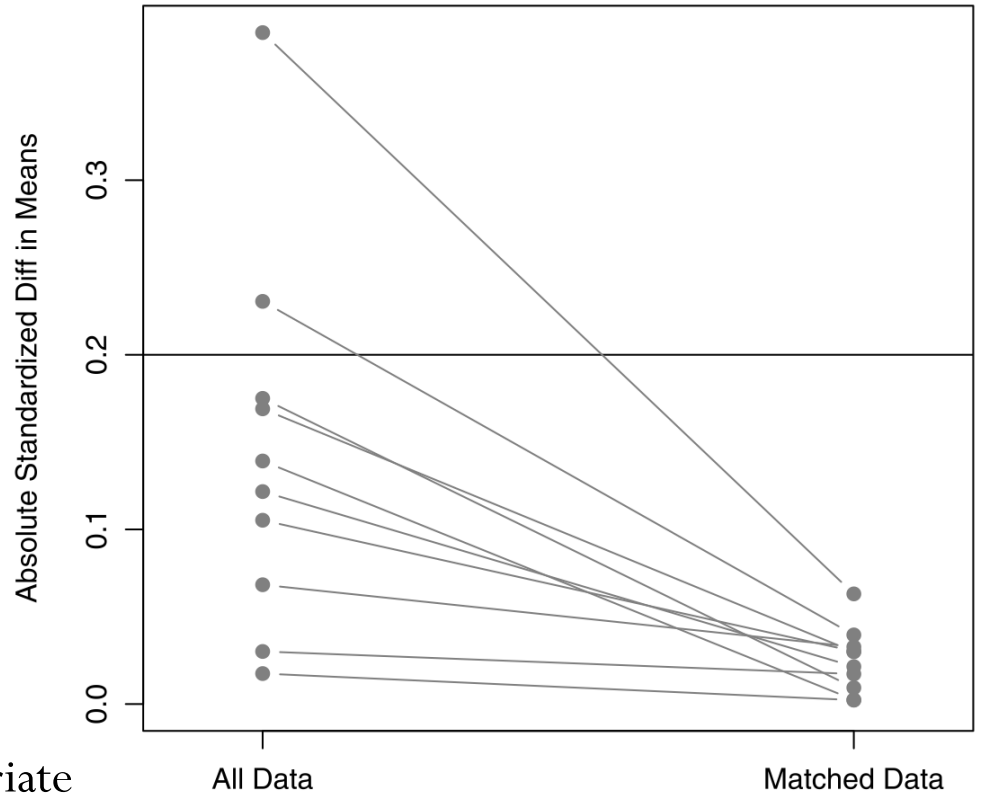
- Do matching produce **balance in the covarient space**?
  - $\tilde{\boldsymbol{p}}(\boldsymbol{X}|\boldsymbol{T}=\boldsymbol{1}) \approx \tilde{\boldsymbol{p}}(\boldsymbol{X}|\boldsymbol{T}=\boldsymbol{0})$ is hard to check if the multivariant distribution is in high dimention

- **Numerical diagnostic** by "Standardizes bias"

$$d_{contiuous} = \frac{\overline{X_T} - \overline{X_C}}{\sqrt{\frac{{S_T}^2 + {S_C}^2}{2}}}$$

$$d_{dichotomous} = \frac{\hat{p}_T - \hat{p}_C}{\sqrt{\frac{\hat{p}_T(1-\hat{p}_T) + \hat{p}_C(1-\hat{p}_C)}{2}}}$$

- **Graphical diagnostic**
  - **QQ plot** for continuous variables
  - Weighted **boxplot** for weighting methods
  - Plot of **standardized differences in means** by each corvariate



ESTIMATING TREATMENT EFFECT

- After **Subclassification** or **Full Matching**
  - Estimate within **each subclass** the aggregate across subclasses (in general)

- After **Weightning**
  - Esimation based on the **whole data**

- After **k:1 Matching**
  - Esimation based on the **whole matched data** as is it been generated through randomization

ESTIMATING TREATMENT EFFECT

After **Subclassification** or **Full Matching**

- Estimate within **each subclass** the aggregate across subclasses (in general)
  - **ATT** – weight each subclass by the ratio of treatment individuals in it

$$w_{subclass_j} = \frac{N_{j,T}}{N_j}$$

  - **ATE** – weight each subclass by its overall size compare to the overall set size

$$w_{subclass_j} = \frac{N_j}{N}$$

After **Weightning** *

- Estimation based on the **whole data**
- The weights are used directly in the reggression models

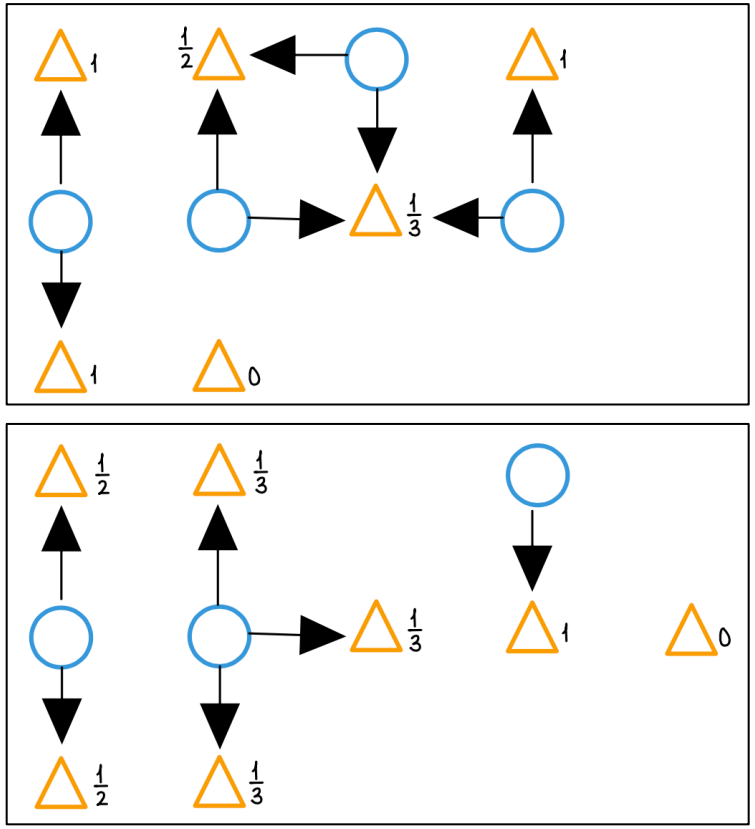
ESTIMATING TREATMENT EFFECT

After **k:1 Matching**

- Esimation based on the **whole data** as is it been generated through randomization

- Weights for matching with replacment
  - Weight each control by a freuency weight
    i.e., number of times they were selected as a match

- Weights for variable ratio matching
  - Weight each control by the proportional to the number of controls matched to "their" treated individual



Weighting control individuals

ESTIMATING TREATMENT EFFECT

Variance Estimation

- Generally, uncertainty regarding the matching process is **not** taken into accont
  - If it does, use **bootstrap**

- Do uncertainty in the propensity score estimation needs to be taken into account ?
  - **Not nesseraly**, let the models run conditional on the covarietes (treated as fixed)
  - **For sure,** not accounting for it will yield confidence intervals wider than necessary (conservative)

GUIDANCE FOR PRACTICE

Tips

- *Matching Method:*
  - If estimating **ATE**, good choices are generally IPTW or full matching.
  - If estimating **ATT and $N_C \gg N_T$**, k:1 nearest neighbor matching without replacement is a good choice for its simplicity and good performance.
  - If estimating **ATT and $N_C \approx N_T$**, appropriate choices are generally subclassification, full matching and weighting by the odds.

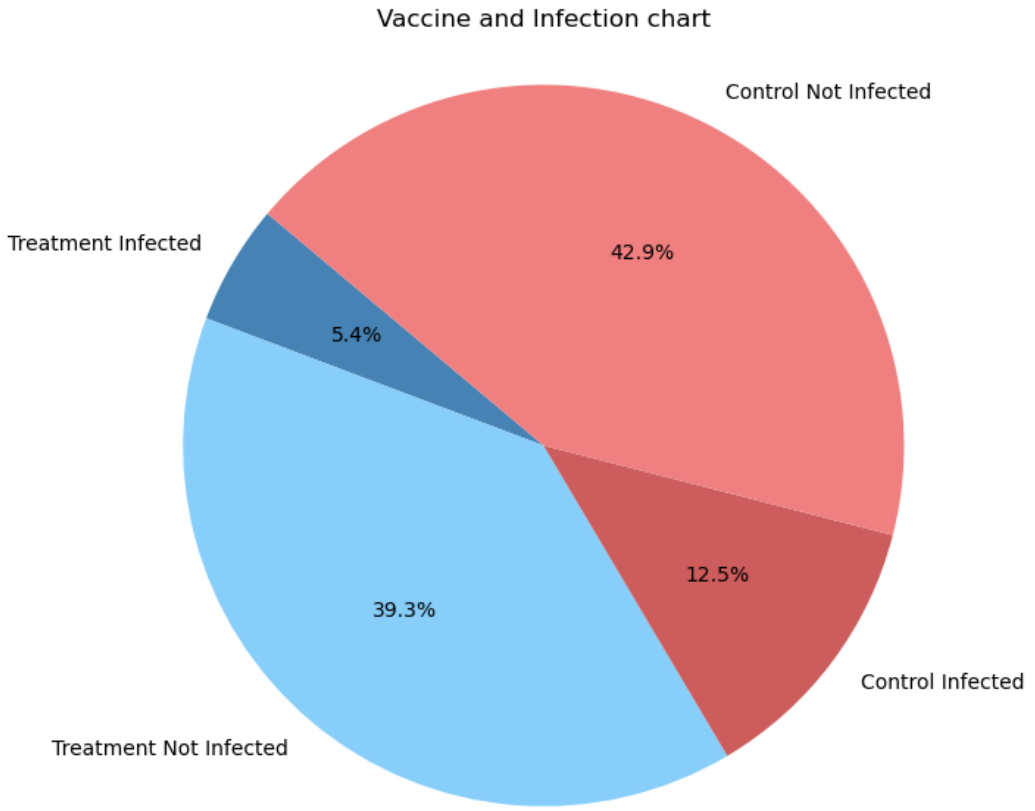
- *Balance:*
  - If adequate, move forward with treatment effect estimation, using regression adjustment on the matched samples.
  - If **imbalance on just a few covariates**, consider incorporating exact or Mahalanobis matching on those variables.
  - If **imbalance on quite a few covariates**, try another matching method or consider changing the estimand or the data.

OUTLINE

- INTRODUCTION
- OVERVIEW ON PROPENSITY SCORE MATCHING
  - Introduction
  - Calculating propencity scores
  - Matching
  - Balance diagnostic
  - Estimating treatment effect – sevral approches
  - Tips for practice
- **DATA**
  - Generating synthetic data
  - Calculating propensity scores
- **STUDY DESIGN**
  - Patient follow up
  - Matching
  - Balance checkinhg
- **STATISTICAL ANALYSIS**
  - Kaplan Meier survival curves
  - Vaccine effectivmess dynamics
- **RESOULTS**
- **LITERATURE**

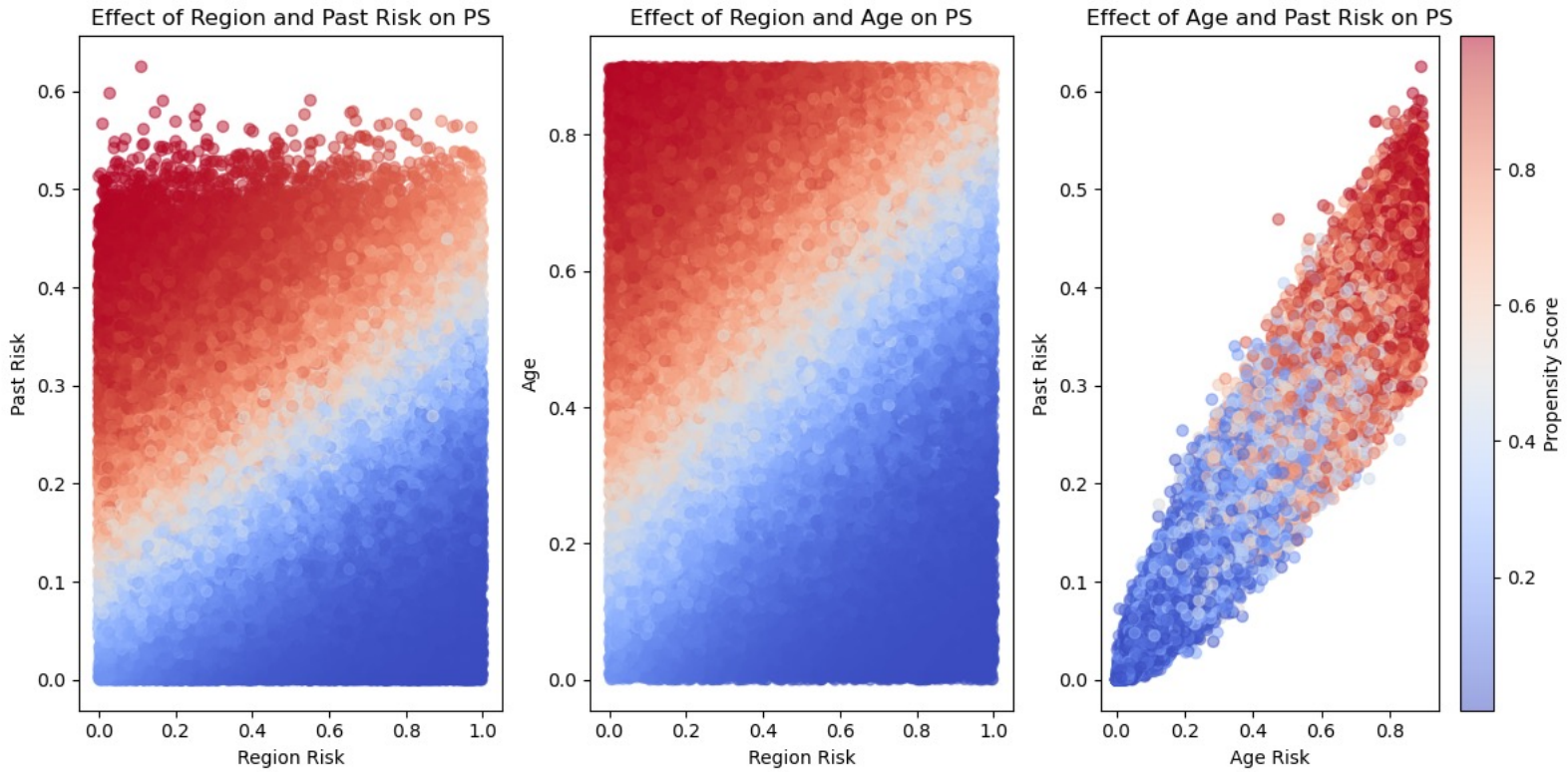
DATA

- **VARIABLE SYNTHESIS**
  - Key variables include age, past infection history, and geographical region risk level.
- **TRIAL ARM ASSIGNMENT**
  - Patients assigned to treatment or control arms based on vaccination events using Bernoulli probability parameter.
- **INFECTION GENERATION**
  - Infection events simulated based on age and region variables, modified by immunization progress.



DATA

- **PROPENSITY SCORE CALCULATION**
  - Propensity scores calculated using logistic regression to estimate the probability of vaccination based on covariates.

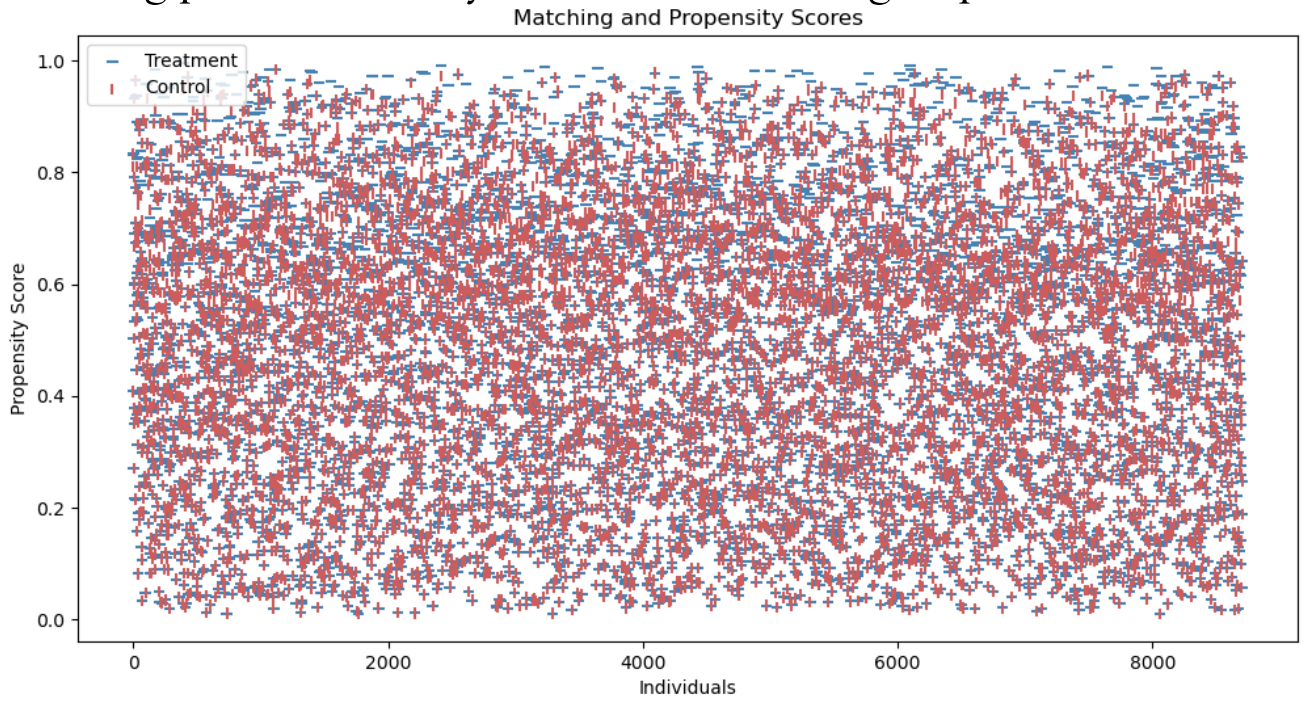


OUTLINE

- INTRODUCTION
- OVERVIEW ON PROPENSITY SCORE MATCHING
  - Introduction
  - Calculating propencity scores
  - Matching
  - Balance diagnostic
  - Estimating treatment effect – sevral approches
  - Tips for practice
- DATA
  - Generating synthetic data
  - Calculating propensity scores
- **STUDY DESIGN**
  - Patient follow up
  - Matching
  - Balance checkinhg
- **STATISTICAL ANALYSIS**
  - Kaplan Meier survival curves
  - Vaccine effectivmess dynamics
- **RESOULTS**
- **LITERATURE**

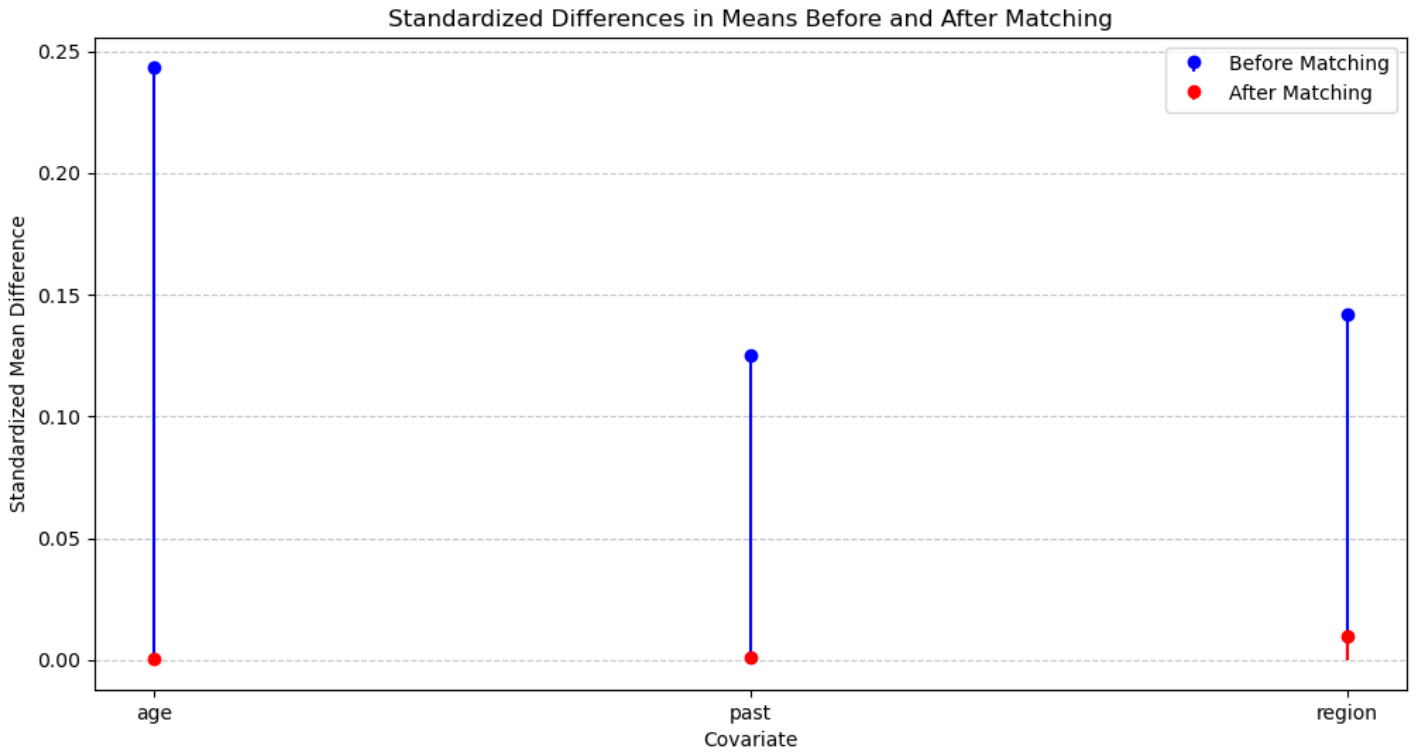
STUDY DESIGN

- **PATIENT FOLLOW UP**
  - Daily follow-up of vaccinated and unvaccinated individuals until specific events or end of the study.
- **PATIENT MATCHING**
  - Propensity score matching performed daily to ensure balanced groups.



STUDY DESIGN

- **BALANCE CHECKING**
  - Visualizations used to check balance between treatment and control groups.

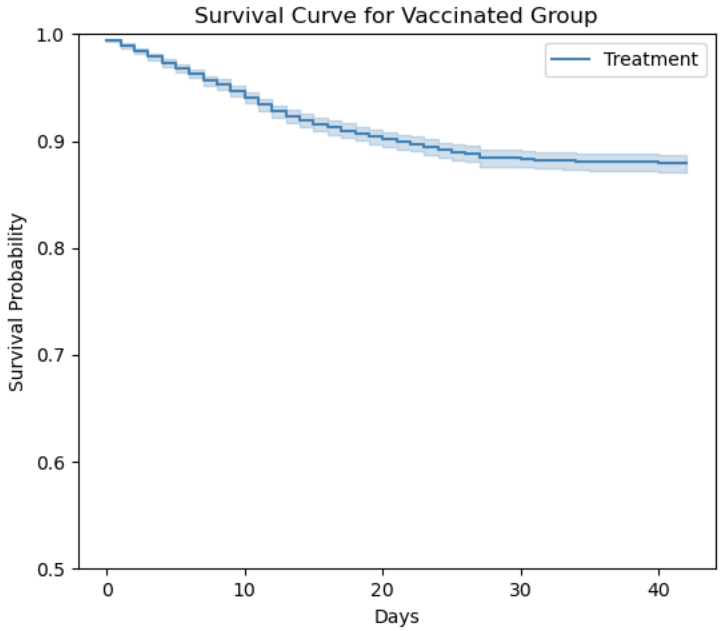


OUTLINE

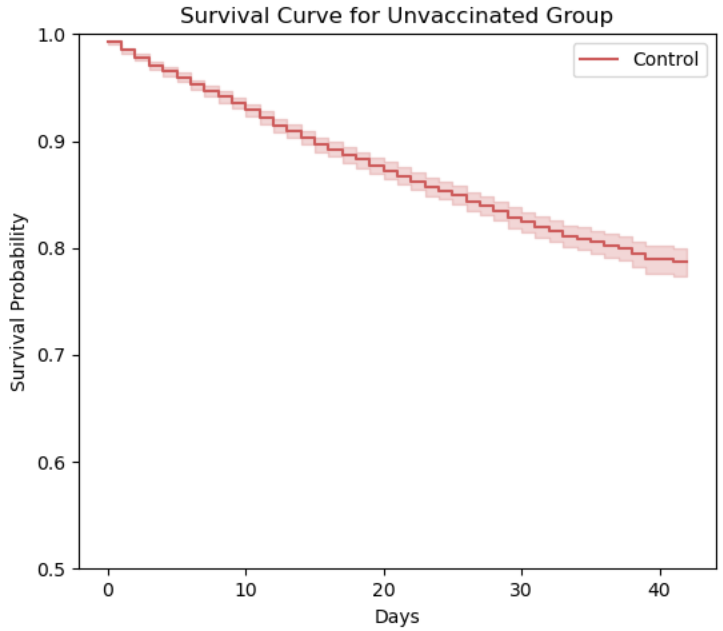
- INTRODUCTION
- OVERVIEW ON PROPENSITY SCORE MATCHING
  - Introduction
  - Calculating propencity scores
  - Matching
  - Balance diagnostic
  - Estimating treatment effect – sevral approches
  - Tips for practice
- DATA
  - Generating synthetic data
  - Calculating propensity scores
- STUDY DESIGN
  - Patient follow up
  - Matching
  - Balance checkinhg
- **STATISTICAL ANALYSIS**
  - Kaplan Meier survival curves
  - Vaccine effectivmess dynamics
- **LITERATURE**

STATISTICAL ANALYSIS

- **KAPLAN-MEIER CURVES**
  - Kaplan-Meier survival curves illustrate dynamic survival over time for both groups.



KM_estimate					
At risk	8655	6754	4808	2678	464
Censored	0	1471	3162	5207	7413
Events	50	480	735	820	828



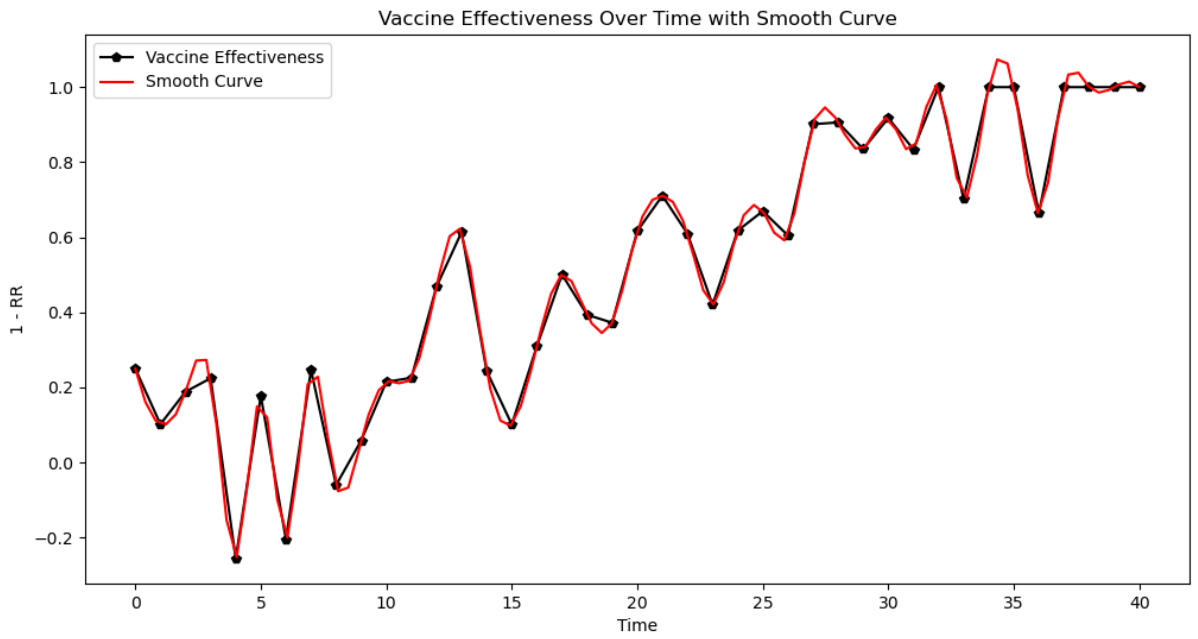
KM_estimate					
At risk	8642	6644	4626	2486	414
Censored	0	1486	3140	5070	7073
Events	63	575	939	1149	1218

STATISTICAL ANALYSIS

- **VACCINE EFFECTIVENESS ASSESSMENT**
  - Risk ratios and vaccine effectiveness presented across different time periods.
- **VISUALIZING VACCINE EFFECTIVENESS DYNAMICS**
  - Visual representation of how 1-RR changes over time.

Stage	Control Risks (N=8705)	Treatment Risks (N=8705)	Risk Ratio (T/C)	Vaccine Effect (1-RR)	CI (95%)
(0, 14)	0.096 -> 0.007	0.080 -> 0.006	0.834	0.166	[0.1616, 0.1718]
(14, 21)	0.132 -> 0.096	0.101 -> 0.080	0.58	0.42	[0.4164, 0.4232]
(21, 28)	0.165 -> 0.132	0.116 -> 0.101	0.46	0.54	[0.5353, 0.5443]
(28, 42)	0.213 -> 0.165	0.121 -> 0.116	0.0955	0.905	[0.8992, 0.9100]

Table 2: Estimated Vaccine Effectiveness against Outcome during Four Time Periods on the matched cohort



OUTLINE

- INTRODUCTION
- OVERVIEW ON PROPENSITY SCORE MATCHING
  - Introduction
  - Calculating propencity scores
  - Matching
  - Balance diagnostic
  - Estimating treatment effect – sevral approches
  - Tips for practice
- DATA
  - Generating synthetic data
  - Calculating propensity scores
- STUDY DESIGN
  - Patient follow up
  - Matching
  - Balance checkinhg
- STATISTICAL ANALYSIS
  - Kaplan Meier survival curves
  - Vaccine effectivmess dynamics
- **LITERATURE**

LITERATURE

- 'BNT162b2 mRNA Covid-19 Vaccine in a Nationwide Mass Vaccination Setting.'
  - Noa Dagan, MD, et al. (2021).
- 'Matching Methods for Causal Inference: A Review and a Look Forward'
  - Elizabeth A. Stuart (2010).
- 'An Introduction to Propensity Score Methods for Reducing the Effects of Confounding in Observational Studies'
  - Peter C. Austin (2011).

ANY QUESTIONS ?

Thank you!