Vaccine Matching Analysis

Vaccine Effectiveness Analysis on a Balanced Synthetic Cohort

Avishag Nevo [1]

A comprehensive reproduction of the published paper "BNT162b2 mRNA Covid-19 Vaccine in a Nationwide Mass Vaccination Setting" by Noa Dagan, MD, et al. led by Professor Yair Goldberg. This statistical project investigates the multifaceted impact of vaccination on infection rates across diverse demographic factors. Leveraging synthetic data generation, sophisticated matching on propensity score algorithms to avoid selection bias, and advanced statistical analysis, this project aims to shed light on the effectiveness of vaccinations in preventing infections over time, just like it the original study. This project is available on GitHub (`https://github.com/avishagnevo/VaccineMatchAnalysis`).

---

1 Introduction

1.1 Objectives

Our primary goal is to achieve a comprehensive reproduction of the published paper "BNT162b2 mRNA Covid-19 Vaccine in a Nationwide Mass Vaccination Setting" by Noa Dagan, MD, et al. By leveraging synthetic data, sophisticated propensity score matching, and advanced statistical models, we aim to provide nuanced insights just as in the paper.

1.2 Methodology

Rooted in the original study's methodology, we synthesize patient data using Python code. The sophisticated matching algorithms seamlessly integrated into the code ensure the creation of balanced treatment and control groups, vital for robust analyses. Advanced statistical methods, including logistic regression and Kaplan-Meier survival model, are applied to capture the intricacies of the vaccination process, as well as patient characteristics, vaccination status, and infection progression that are crafted using the code, ensuring alignment with real-world scenarios.

Affiliation [1] Technion Student
Correspondence avishag.nevo@campus.technion.ac.il
Version December 20, 2023

1.3 RESULTS

Estimated vaccine effectiveness for the study outcomes at days 14 through 20 after the crafted first dose is 42% (95% CI, 41.64% to 42.32%), at days 21 through 28 after the crafted first dose is 54% (95% CI, 45.53% to 54.43%), at 7 or more days after the crafted second dose is 90% (95% CI, 88.92% to 91%)[1].

Characteristic	Control (N=8705)	Treatment (N=8705)
Coefficients Statistics:		
age		
mean	0.503	0.498
std	0.224	0.189
range	(0.001, 0.899)	(0.029, 0.899)
past		
mean	0.250	0.250
std	0.121	0.099
range	(0.000, 0.572)	(0.008, 0.538)
region		
mean	0.506	0.494
std	0.302	0.252
range	(0.000, 0.999)	(0.000, 0.989)
Pearson Coefficients Correlation:		
past vs age	0.937	0.909
region vs age	0.636	0.391
region vs past	0.635	0.378

Table 1: Characteristics for matched data

[1]Experiment run with default parameters, each matched study group included 8705 persons, for fixed daily immunization level of 42%, 60%, 92% respectively - as estimated in the paper

2 Data

In this study, patient data were synthesized using custom Python classes, `Patient` and `Experiment`, designed to simulate characteristics relevant to the investigation of vaccine effectiveness. This patients simulation process generates a diverse set of patient profiles, vaccination and infection status and time of event (if occurred). The resulting data set in the input set size serves as the foundation for the subsequent statistical analysis aimed at assessing vaccine effectiveness.

2.1 Variable Synthesis

`Patient` class constructor `__init__()` initializes key attributes for each simulated patient, including vaccination-related parameters:

- **age**: Age is randomly sampled within the specified age range.
- **past**: Past infection history is generated using a beta distribution based on age (positively correlated)
- **region**: Geographical region risk level is simulated as a random value within defined parameters.

2.2 Trail Arm Assignment

The function `generate_trail_arm()` employs the sampled variables multiplied by a set base probability as a Bernoulli probability parameter to generate daily vaccination event throughout the designated vaccination window. If the patient was vaccinated at least once (sampled as True on the i'th day of the vaccination window), the follow-up starts on this day, and the patient is assigned to the treatment arm. Conversely, if the patient was never vaccinated (sampled as False throughout the entire vaccination window), they are assigned to the control arm.

2.3 Infection Generation

The function `generate_infection_process()` employs age and region variables multiplied by a set base probability as a Bernoulli probability parameter to generate daily infection event throughout follow up window. The event occurrence is modified by the immunization progress determined by `generate_immunization_process()` function. Immunization process simulates the progression of immunization in different stages for the vaccinated. Protection levels at each stage determine the probability of not being infected.

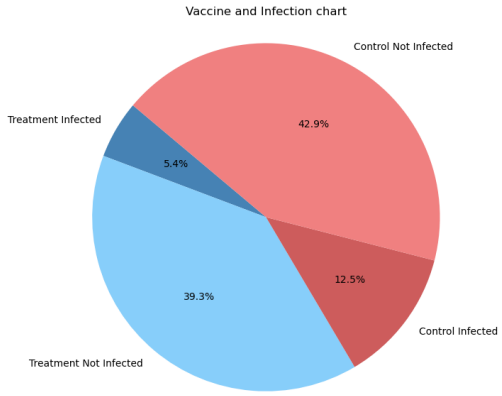

Figure 1: Full cohort distribution by arm and event

2.4 Propensity Score Calculation

In the quest to evaluate vaccine effectiveness, the calculation of propensity scores assumes a pivotal role, particularly in the forthcoming matching process designed to balance observed covariates between vaccinated and unvaccinated groups. Propensity scores represent the estimated probability of an individual receiving the treatment (vaccination) based on a set of observed (sampled) covariates. To accomplish this, a logistic regression model is used and is adeptly trained on the comprehensive dataset encompassing all patients' information.

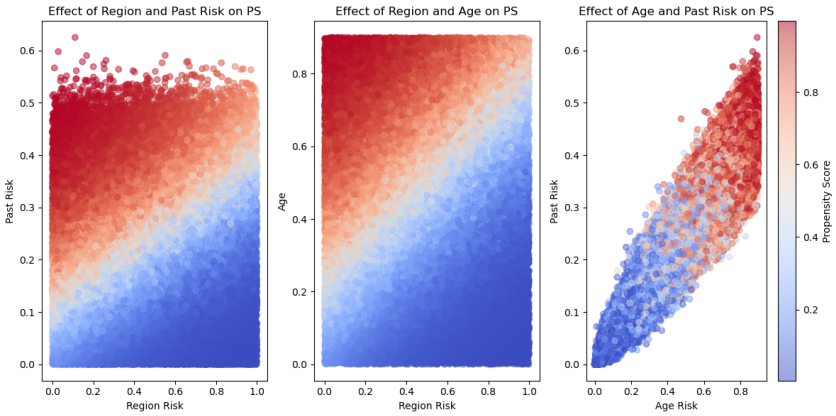

Figure 2: Logistic Regression hypothesis in feature space

3 Study Design

3.1 Patient Followup

Each day during the vaccination window, all newly vaccinated persons were matched in a 1:1 ratio to unvaccinated controls, the follow-up starts at this day. For each person, follow-up ended at the earliest of the following events: occurrence of an outcome event, vaccination (for unvaccinated controls), vaccination of the matched control (for vaccinated persons), or the end of the study period. Newly vaccinated persons were eligible for inclusion in the study (follow-up starts at this day), even if they had previously been selected as a control, their previously matched treatment is excluded from the trail for good.

3.2 Patient Matching

The patient matching process is the main step in ensuring a balanced and comparable comparison between vaccinated and unvaccinated groups. The approach utilizes the `Matching` class, integrating a carefully designed algorithm to pair patients based on propensity scores and control for relevant variables. Here's a comprehensive explanation of the approach:

- Initialization: Two sets, treatment and control, are initialized to categorize patients based on their vaccination status at the first trail day.
- Daily Update of Treatment and Control Sets: The `update_treatment_control()` method is employed for the continuous updating of treatment and control sets as the study progresses over time. Patients who receive the vaccine during the day are transferred from the control set to the treatment set.
- Daily Matching: The `run_daily_matching()` method performs 1:1 matching process on a daily basis. For each patient in the treatment group, a suitable control patient is sought based on narrow caliper matching. Successfully matched pairs are recorded, with each treated patient having a corresponding matched control patient.
- Matching Criteria: The matching criteria include ensuring that patients are still at risk, i.e., not yet infected (still at follow-up stage). Propensity scores are used to gauge the similarity between treatment and control patients. A narrow caliper is employed to limit the allowable difference in propensity scores for a valid match.
- Overall Matching Execution: The `run_matching()` method orchestrates the patient matching process for the entire follow-up period.

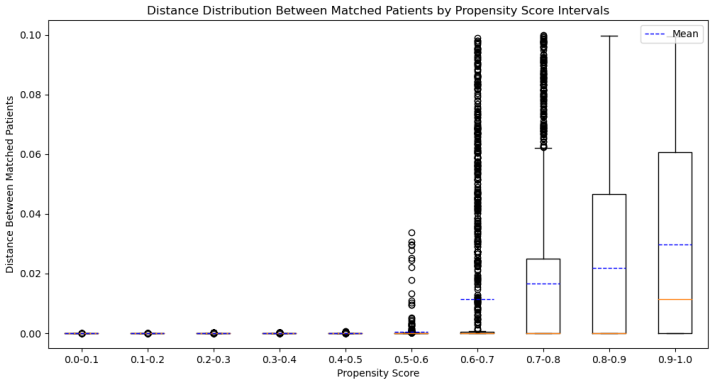

Figure 3: Distances Distribution between each matched couples distributed by buckets of size 0.1 of propensity scores

4 Statistical Analysis

4.1 Balance Checking

Covariate balance after matching was evaluated with the use of a lollipop plot of the mean differences between variables values for the vaccinated and unvaccinated groups.

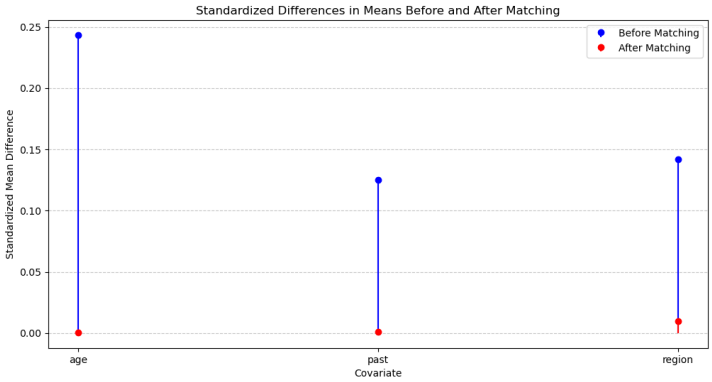

Figure 4: Matched groups' means difference by attribure

5 Statistical Analysis

The Analysis class employs Survival Analysis statistical approach to assess and analyze the outcomes of the study. Survival curves were constructed using **Kaplan-Meier** estimation, to illustrate the dynamic nature of survival over time for both groups. The Kaplan-Meier curves were fitted using the `fit_kaplan_meier()`.

5.1 Visualizing Kaplan-Meier Curves

To visually interpret the impact of vaccination, Kaplan-Meier curves were generated for both the treatment (vaccinated) and control (unvaccinated) groups. These curves provide a clear representation of the survival probabilities over the follow-up period using the `visualize_kaplan_meier_curves()` function.

5.2 Vaccine Effectiveness Assessment

Analysis of vaccine effectiveness is presented using `vaccine_effectiveness_table()` function, examining the risk ratios across different stages of the follow-up period. The calculated risk ratios, along with confidence intervals, provide a nuanced understanding of how the vaccine influences infection risk over time.

5.3 Visualizing Vaccine Effectiveness Dynamics

The dynamics of vaccine effectiveness were explored by plotting the inverse of the infection risk ratio over time using the `visualize_vaccine_effectiveness()` function.

6 Results

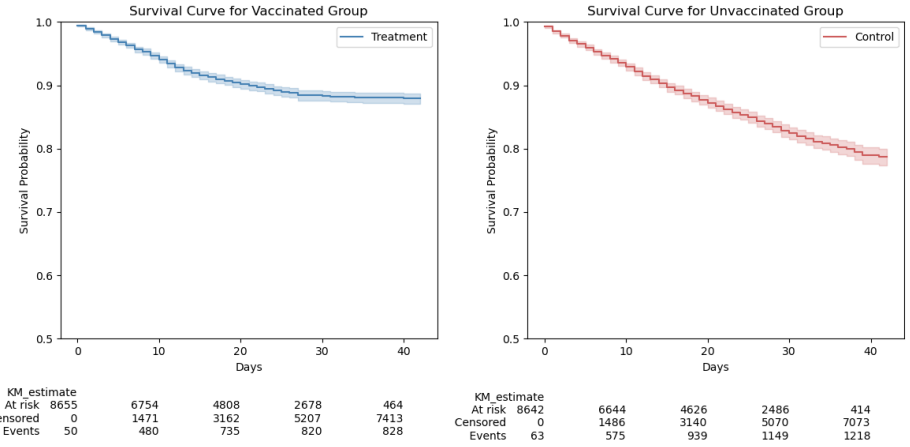

Figure 5: Kaplan-Meier Survival Curves

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

Estimated vaccine effectiveness for the study outcomes at days 14 through 20

after the crafted first dose is 42% (95% CI, 41.64% to 42.32%), at days 21 through 28 after the crafted first dose is 54% (95% CI, 45.53% to 54.43%), at 7 or more days after the crafted second dose is 90% (95% CI, 88.92% to 91%), for fixed daily immunization level of 42%, 60%, 92% respectively - as estimated in the paper. Experiment run with default parameters, each matched study group included 8705 persons.

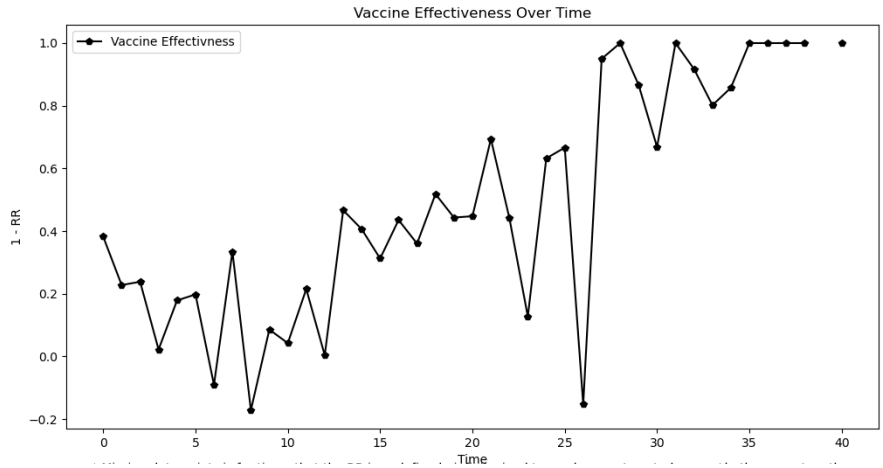

Figure 6: Vaccine Effectiveness over time
*Missing data points is for times that the risk ratio is undefined- infinitesimal to no change at control risk / both groups together

References

[1] BNT162b2 mRNA Covid-19 Vaccine in a Nationwide Mass Vaccination Setting. Noa Dagan, MD, et al. (2021).
https://www.nejm.org/doi/full/10.1056/nejmoa2101765