# Impact of Immunization and Preventive Measures on COVID-19 Reinfection Timing and Severity

data-to-paper

August 7, 2024

**Abstract**

In the ongoing battle against the COVID-19 pandemic, understanding the interplay between prior immunization and the subsequent risk and severity of reinfections is paramount. Amidst escalating concerns about vaccine efficacy over time and against various SARS-CoV-2 variants, this study fills a critical research gap by examining how different combinations of immunity influence reinfection rates and symptomatology among healthcare workers. Employing a robust dataset from a longitudinal study of 2,595 healthcare workers in Switzerland, we utilized generalized linear models to scrutinize the impact of vaccination status, prior infection, protective measures, and their synergistic effects on reinfection dynamics from August 2020 to March 2022. Our analyses reveal a marked increase in the risk and symptom severity of COVID-19 reinfections among individuals without any form of prior immunity. Conversely, hybrid immunity, a combination of infection-induced and vaccine-induced immunities, significantly reduced the frequency and severity of subsequent infections. Use of FFP2 masks and minimal patient contact were also associated with diminished risk and milder symptoms. Although this study's observational nature limits causal inferences and might be influenced by uncontrolled confounders, these findings strongly advocate for the enforcement of comprehensive vaccination programs and stringent protective measures in healthcare settings. This insight is crucial for devising effective public health strategies and safeguarding frontline workers against COVID-19 reinfections.

## Introduction

As the world continues to grapple with the COVID-19 pandemic, the scientific community faces an urgent necessity of understanding the SARS-CoV-2 virus and its implications [1]. Experts speculate that the role of immunity, specifically those arising from prior infections or vaccinations, could be crucial amidst the emergence of multiple variants of SARS-CoV-2 [2, 3]. Certain factors can influence the onset and severity of reinfections, which perpetuate the virus spread. Current literature extensively delves into the initial infection and the consequent immunity, highlighting the dynamics of reinfection risk and associated morbidity [4, 5]. Yet, limited comprehensive evidentiary material investigates how combinations of immunity, demographics, and protective measures impact the chronology and intensity of symptoms associated with COVID-19 reinfections [6].

Addressing this knowledge gap, this study probes the influence of varying immunization statuses and associated factors on the timing and symptom severity of COVID-19 reinfections. Leveraging a robust dataset harvested from a longitudinal study covering 2,595 healthcare workers in Switzerland [7], this research illuminates the dynamics of reinfections among individuals exposed to both the delta and omicron variants between August 2020 and March 2022. The dataset comprises specific variables, including individual demographics, vaccination and infection statuses at different time intervals, and protective measures employed by healthcare workers, facilitating an enriching analysis of symptom severity among reinfection cases [8, 9].

This research employed popular and sophisticated statistical techniques, concretely generalized linear models, to scrutinize the correlation between vaccination and infection status as well as demographic factors and protective measures [10]. Our models were carefully configured to individually assess these effects on the time and symptom severity during reinfections, mimicking the analysis structure utilized in similar influential works [11].

The initial results underscore the multidimensional role of immunization and protective measures in attenuating the frequency and severity of COVID-19 reinfections [12, 13]. Notably, the findings indicate that this role is significant in the context of healthcare workers who are continually exposed to the virus. These findings could potentially extend existing knowledge on managing COVID-19 spread and its implications, hence contributing to the broader scientific and healthcare communities' effort to navigate the ongoing pandemic [1].

## Results

First, to understand the impact of immunization status and other factors on the time until reinfection with COVID-19, we conducted an analysis captured in Table 1. This analysis utilized generalized linear models to assess the influence of vaccination status, demographic factors, and protective measures on reinfection timing. The model identified significant effects of non-immunization (-0.647, $P < 10^{-6}$), age, and protective mask usage. Individuals lacking any immunization experienced reinfections significantly sooner. Using FFP2 masks extended the time until reinfection significantly (-0.12, $P = 0.00805$). Additionally, each year of increase in age slightly delayed reinfections (0.00367, $P = 0.0305$). Vaccination without booster did not show a statistically significant delay in reinfection time ($P = 0.245$).

Table 1: Association between various factors and time until reinfection

| | Coef | SE | Z-Score | P-value | [0.025 | 0.975] |
|---|---|---|---|---|---|---|
| **Intercept** | 4.56 | 0.113 | 40.3 | $<10^{-6}$ | 4.34 | 4.78 |
| **Not Immun** | -0.647 | 0.0831 | -7.78 | $<10^{-6}$ | -0.81 | -0.484 |
| **Vacc** | -0.0559 | 0.0481 | -1.16 | 0.245 | -0.15 | 0.0384 |
| **Inf** | -0.103 | 0.0952 | -1.08 | 0.278 | -0.29 | 0.0833 |
| **Fem** | 0.0346 | 0.0436 | 0.795 | 0.427 | -0.0508 | 0.12 |
| **BMI < 30** | -0.0881 | 0.0527 | -1.67 | 0.0944 | -0.191 | 0.0151 |
| **Age** | 0.00367 | 0.0017 | 2.16 | 0.0305 | 0.000345 | 0.007 |
| **P. Contact** | 0.0586 | 0.0437 | 1.34 | 0.179 | -0.027 | 0.144 |
| **Use FFP2** | -0.12 | 0.0455 | -2.65 | 0.00805 | -0.21 | -0.0314 |

Statistical analyses performed using Generalized Linear Models (GLM).
**Fem**: 1: Yes, 0: No
**BMI < 30**: 1: <30, 0: >=30
**P. Contact**: 1: Yes, 0: No
**Use FFP2**: 1: Yes, 0: No
**Age**: years

Then, to further understand the severity of symptoms at reinfection, we analyzed the number of reported symptoms and their association with various factors, as detailed in Table 2. Non-immunized individuals reported significantly more symptoms than their counterparts (1.26, $P < 10^{-6}$). The analysis also highlighted that previously infected individuals displayed more symptoms (0.953, $P = 1.33\ 10^{-5}$). Vaccinated individuals had somewhat fewer symptoms than non-immunized individuals but still substantial (0.784, $P < 10^{-6}$). Interestingly, females reported more symptoms than males (0.298, $P = 0.00292$), and older age was associated with fewer symptoms (-0.00853, $P = 0.0288$).

Our analyses capitalized on a robust dataset with 2947 observations which bolstered the statistical power and reliability of the outcomes. This

Table 2: Association between vaccination status and number of symptoms at reinfection

| | Coef | SE | P-value |
|---|---|---|---|
| **Intercept** | 3.21 | 0.26 | $<10^{-6}$ |
| **Not Immun** | 1.26 | 0.191 | $<10^{-6}$ |
| **Vacc** | 0.784 | 0.111 | $<10^{-6}$ |
| **Inf** | 0.953 | 0.219 | $1.33\ 10^{-5}$ |
| **Fem** | 0.298 | 0.1 | 0.00292 |
| **BMI $<$ 30** | -0.232 | 0.121 | 0.0551 |
| **Age** | -0.00853 | 0.0039 | 0.0288 |
| **P. Contact** | 0.158 | 0.1 | 0.116 |
| **Use FFP2** | -0.108 | 0.104 | 0.299 |

Statistical analyses performed using Generalized Linear Models (GLM).
**Fem**: 1: Yes, 0: No
**BMI $<$ 30**: 1: <30, 0: >=30
**P. Contact**: 1: Yes, 0: No
**Use FFP2**: 1: Yes, 0: No
**Age**: years

dataset allowed for precise estimates of the time to reinfection and symptom severity across different immunization status categories.

In summary, these results underscore that while previous immunization, whether from vaccines or infections, reduces symptom severity at reinfection, individuals without immunization not only experience faster reinfection but also endure more severe symptoms. Protective measures such as the usage of FFP2 masks extend the period before reinfection. This collection of findings highlights the significant role of active immunization and protective behaviors in managing the impact of COVID-19, especially in preventing severe reinfection symptoms.

## Discussion

Our study sheds light on the interplay of prior immunization, preventive measures, and demographic characteristics, and their subsequent effect on the risk and severity of COVID-19 reinfections [1]. These issues, critical for managing the pandemic, were investigated within a cohort of 2,595 healthcare workers in Switzerland, a high-risk group continually exposed to the virus [14]. We adopted an analysis structure using generalized linear models similar to recent influential works, aiming for a deep understanding of the dynamics of reinfection [10, 12].

Significantly, our analysis revealed that non-immunized individuals experienced a more rapid timeline to reinfection and more severe symptoms, aligning with recent findings suggesting a protective effect of prior immunization [15]. Interestingly, those with hybrid immunity–prior infection and immunization–had significantly fewer reinfections, corroborating previous research highlighting the strength of hybrid immunity [3]. Use of protective measures, particularly FFP2 masks, delayed reinfection, augmenting the scarce empirical evidence on personal protective equipment effectiveness [16].

Although these findings contribute significantly to the literature, they should be interpreted in light of several limitations. The observational nature of our study limits our ability to make causal inferences and there may have been unmeasured confounding factors not controlled for in our analyses. Variables such as comorbidities, which could influence both susceptibility to and severity of COVID-19 [17], were not included in the analysis. Additionally, our sample consisted exclusively of healthcare workers, who differ from the general population in their exposure risk, behavior patterns, and use of preventive measures. Therefore, the generalizability to other populations may not be straightforward. As highlighted by recent research, demographic factors such as age and gender can significantly affect the risk and severity of reinfection [10, 17, 18], reaffirming our findings. This highlights the value in further research within diverse populations, thus ensuring wider applicability of the findings.

Future research prospects are abundant and critical for a holistic understanding of COVID-19 reinfections. Experimental studies utilizing randomized controlled trials could shed further light on the causal relationship between immunization, preventive measures, and COVID-19 reinfections. Additionally, the complications of vaccine protocols, such as the use of vaccine boosters and vaccine types [19], warrant focused scientific attention. Together, these approaches would help refine our understanding of optimal protective strategies.

Ultimately, our findings underscore the crucial role of comprehensive immunization and preventive measures in mitigating COVID-19 reinfections, particularly within high exposure environments like healthcare settings. The notable reduction in reinfection timing and symptom severity among vaccinated individuals compared to their counterparts emphasizes the effectiveness of vaccines as a primary preventive measure [4]. Furthermore, the study underscores the relevance of PPE, such as FFP2 masks, in healthcare settings for combating infection spread. Future public health strategies can greatly benefit from insights into managing the pandemic more effectively, which will be crucial until global herd immunity is achieved.

# Methods

### Data Source

The data for this study were collected from a prospective, multicentre cohort of hospital employees across ten healthcare networks in Eastern and Northern Switzerland, between August 2020 and March 2022. Our dataset included a total of 2,595 participants with a median follow-up duration of 171 days. The dataset comprises two main files. The first file details vaccination and infection events of all healthcare workers, capturing diverse variables such as demographic details, vaccination status, previous infections, and protective measures during different time intervals. The second file provides data on symptoms for those healthcare workers who tested positive for SARS-CoV-2, listing information pertinent to their infection episodes and symptomatology. Our study includes records from both the delta and omicron variants periods.

### Data Preprocessing

The dataset required preliminary cleaning and integration. Initially, missing data points were replaced with appropriate indicators to standardize the dataset. Following this, the two main datasets were merged based on multiple key identifiers to maintain continuity between the vaccination/infection data and the symptom data. This merge operation was performed to create a unified dataset which facilitated comprehensive analyses. Post-merge, the dataset underwent transformation to accommodate the statistical modeling needs. This included encoding categorical variables such as vaccination group, sex, and BMI into dummy variables to be used as predictors in the regression models.

### Data Analysis

The principal statistical approach employed was generalized linear modeling, aiming to elucidate the effects of various factors on the timing and severity of COVID-19 reinfections. The analysis was executed in two parts. Firstly, we assessed the relationship between demographic factors, vaccination status, protective measures, and the time until reinfection. Secondly, we analyzed the data to determine the associated factors influencing the number of symptoms at the point of reinfection. The models incorporated variables such as age, vaccination status, sex, BMI, patient contact frequency, and the usage of protective respiratory masks. These analyses facilitated a detailed understanding of how these variables interact to influence susceptibility to and severity of reinfection among healthcare workers.

### Code Availability

Custom code used to perform the data preprocessing and analysis, as well as the raw code outputs, are provided in Supplementary Methods.

## References

[1] J. Pulliam, C. van Schalkwyk, N. Govender, A. von Gottberg, C. Cohen, M. Groome, J. Dushoff, K. Mlisana, and H. Moultrie. Increased risk of sars-cov-2 reinfection associated with emergence of omicron in south africa. *Science (New York, N.y.)*, 2022.

[2] A. Mensah, J. Lacy, J. Stowe, Giulia Seghezzo, R. Sachdeva, R. Simmons, A. Bukasa, S. OBoyle, N. Andrews, M. Ramsay, H. Campbell, and K. Brown. Disease severity during sars-cov-2 reinfection: a nationwide study. *The Journal of Infection*, 84:542 – 550, 2022.

[3] J. Dan, J. Mateus, Y. Kato, K. Hastie, E. Yu, Caterina E. Faliti, A. Grifoni, S. Ramirez, Sonya Haupt, A. Frazier, Catherine Nakao, V. Rayaprolu, Stephen A. Rawlings, Bjoern Peters, F. Krammer, V. Simon, E. Saphire, Davey M. Smith, D. Weiskopf, A. Sette, and S. Crotty. Immunological memory to sars-cov-2 assessed for up to 8 months after infection. *Science (New York, N.y.)*, 2021.

[4] Xiang Ren, Jie Zhou, J. Guo, Chunmei Hao, Mengxue Zheng, Rong Zhang, Qiao Huang, Xiaomei Yao, Ruiling Li, and Yinghui Jin. Reinfection in patients with covid-19: a systematic review. *Global Health Research and Policy*, 7, 2022.

[5] D. Cromer, J. Juno, D. Khoury, A. Reynaldi, A. Wheatley, S. Kent, and M. Davenport. Prospects for durable immune control of sars-cov-2 and prevention of reinfection. *Nature Reviews. Immunology*, 21:395 – 404, 2021.

[6] Jingzhou Wang, Christopher Kaperak, Toshiro Sato, and A. Sakuraba. Covid-19 reinfection: a rapid systematic review of case reports and case series. *Journal of Investigative Medicine*, 69:1253 – 1255, 2021.

[7] Rachel Baskin and R. Bartlett. Healthcare worker resilience during the covid19 pandemic: An integrative review. *Journal of Nursing Management*, 29:2329 – 2342, 2021.

[8] A. Pataka, S. Kotoulas, A. Tzinas, Nectaria Kasnaki, E. Sourla, E. Chatzopoulos, I. Grigoriou, and P. Argyropoulou. Sleep disorders and mental stress of healthcare workers during the two first waves of covid-19 pandemic: Separate analysis for primary care. *Healthcare*, 10, 2022.

[9] Faisal Mohammed Nafie Ali and Abdelmoneim Ali Mohamed Hamed. Usage apriori and clustering algorithms in weka tools to mining dataset of traffic accidents. *Journal of Information and Telecommunication*, 2:231 – 245, 2018.

[10] Sahil Loomba, A. de Figueiredo, S. Piatek, K. de Graaf, and H. Larson. Measuring the impact of covid-19 vaccine misinformation on vaccination intent in the uk and usa. *Nature Human Behaviour*, 5:337 – 348, 2021.

[11] M. Filbin, Arnav Mehta, Alexis M. Schneider, Kyle R. Kays, Jamey Guess, M. Gentili, Bank G. Fenyves, Nicole C. Charland, Anna L. K. Gonye, Irena Gushterova, Hargun K. Khanna, Thomas J. LaSalle, Kendall M. Lavin-Parsons, B. M. Lilley, Carl L. Lodenstein, Kasidet Manakongtreecheep, J. Margolin, Brenna N. Mckaig, Maricarmen Rojas-Lopez, Brian C. Russo, Nihaarika Sharma, J. Tantivit, M. Thomas, R. Gerszten, Graham S. Heimberg, P. Hoover, David J. Lieb, B. Lin, D. Ngo, K. Pelka, Miguel Reyes, C. Smillie, Avinash Waghray, Thomas E. Wood, Amanda S. Zajac, L. Jennings, I. Grundberg, R. Bhattacharyya, B. Parry, A. Villani, Moshe Sade-Feldman, N. Hacohen, and M. Goldberg. Longitudinal proteomic analysis of severe covid-19 reveals survival-associated signatures, tissue-specific cell death, and cell-cell interactions. *Cell Reports Medicine*, 2, 2021.

[12] Yeen Huang and N. Zhao. Generalized anxiety disorder, depressive symptoms and sleep quality during covid-19 outbreak in china: a web-based cross-sectional survey. *Psychiatry Research*, 288:112954 – 112954, 2020.

[13] M. Johansson, Talia M. Quandelacy, S. Kada, Pragati V Prasad, M. Steele, J. Brooks, R. Slayton, M. Biggerstaff, and J. Butler. Sars-cov-2 transmission from people without covid-19 symptoms. *JAMA Network Open*, 4, 2021.

[14] J. Slezak, K. Bruxvoort, Heidi Fischer, Benjamin I Broder, B. Ackerson, and S. Tartof. Rate and severity of suspected sars-cov-2 reinfection in a cohort of pcr-positive covid-19 patients. *Clinical Microbiology and Infection*, 27:1860.e7 – 1860.e10, 2021.

[15] Benjamin Bowe, Yan Xie, and Z. AlAly. Acute and postacute sequelae associated with sars-cov-2 reinfection. *Nature Medicine*, 28:2398 – 2405, 2022.

[16] H. kik, Y. a, M. Sezerol, Aral Surmeli, Y. Ta, . ahin, and I. Maral. Sociodemographic and clinical features of covid-19 reinfection cases among healthcare workers in turkey. 2021.

[17] Letcia Adrielle dos Santos, P. Filho, Ana Maria Fantini Silva, Joo Vasco Santos, D. Santos, M. Aquino, Rafaela Mota de Jesus, M. L. Almeida, J. S. da Silva, D. Altmann, R. Boyton, Cliomar Alves dos Santos, C. N. O. Santos, J. Alves, Ianaline Lima Santos, L. Magalhes, E. Belitardo, Danilo J P G Rocha, J. Almeida, L. G. Pacheco, E. Aguiar, G. Campos, S. Sardi, Rejane H. Carvalho, A. D. de Jesus, K. Rezende, and R. P. de Almeida. Recurrent covid-19 including evidence of reinfection and enhanced severity in thirty brazilian healthcare workers. *The Journal of Infection*, 82:399 – 406, 2021.

[18] Ali Tavakoli, F. Lotfi, Mehrzad Lotfi, Mohsen Bayati, M. Seif, M. Salesi, Mehrnoosh Emadi, K. Keshavarz, and Sajad Delavari. Covid-19 reinfection rate and related risk factors in fars province, iran: A retrospective cohort study. *Iranian Journal of Medical Sciences*, 48:302 – 312, 2023.

[19] Humeyra Aslaner, H. Aslaner, Yasemin Savranlar, and A. Benli. Covid-19 relapse and reinfection frequency, clinical features of cases. *Ahi Evran Medical Journal*, 2022.

# A Data Description

Here is the data description, as provided by the user:

```
\#\# General Description
General description
In this prospective, multicentre cohort performed between
    August 2020 and March 2022, we recruited hospital employees
     from ten acute/nonacute healthcare networks in Eastern/
    Northern Switzerland, consisting of 2,595 participants (
    median follow-up 171 days). The study comprises infections
    with the delta and the omicron variant. We determined
    immune status in September 2021 based on serology and
    previous SARS-CoV-2 infections/vaccinations: Group N (no
    immunity); Group V (twice vaccinated, uninfected); Group I
    (infected, unvaccinated); Group H (hybrid: infected and $\
    geq$1 vaccination). Participants were asked to get tested
    for SARS-CoV-2 in case of compatible symptoms, according to
     national recommendations. SARS-CoV-2 was detected by
    polymerase chain reaction (PCR) or rapid antigen diagnostic
     (RAD) test, depending on the participating institutions.
    The dataset is consisting of two files, one describing
    vaccination and infection events for all healthworkers, and
     the secone one describing the symptoms for the
    healthworkers who tested positive for SARS-CoV-2.
\#\# Data Files
The dataset consists of 2 data files:

\#\#\# File 1: "TimeToInfection.csv"
Data in the file "TimeToInfection.csv" is organised in time
    intervals, from day\_interval\_start to day\_interval\_stop
    . Missing data is shown as "" for not indicated or not
    relevant (e.g. which vaccine for the non-vaccinated group).
     It is very important to note, that per healthworker (=ID
    number), several rows (time intervals) can exist, and the
    length of the intervals can vary (difference between day\
    _interval\_start and day\_interval\_stop). This can lead to
     biased results if not taken into account, e.g. when
    running a statistical comparison between two columns. It
    can also lead to biases when merging the two files, which
    therefore should be avoided. The file contains 16 columns:

ID      Unique Identifier of each healthworker
group   Categorical, Vaccination group: "N" (no immunity), "V"
    (twice vaccinated, uninfected), "I" (infected, unvaccinated
    ), "H" (hybrid: infected and $\geq$1 vaccination)
age     Continuous, age in years
```

~~~
sex     Categorical, female", "male" (or "" for not indicated)

BMI     Categorical, "o30" for over 30  or "u30" for below 30

patient\_contact        Having contact with patients during
    work during this interval, 1=yes, 0=no
using\_FFP2\_mask       Always using protective respiratory
    masks during work, 1=yes, 0=no
negative\_swab  documentation of $\geq$1 negative test in the
    previous month, 1=yes, 0=no
booster receipt of booster vaccination, 1=yes, 0=no (or "" for
    not indicated)
positive\_household     categorical, SARS-CoV-2 infection of a
    household contact within the same month, 1=yes, 0=no
months\_since\_immunisation     continuous, time since last
    immunization event (infection or vaccination) in months.
    Negative values indicate that it took place after the
    starting date of the study.
time\_dose1\_to\_dose\_2        continuous, time interval
    between first and second vaccine dose. Empty when not
    vaccinated twice
vaccinetype     Categorical, "Moderna" or "Pfizer\_BioNTech" or
     "" for not vaccinated.
day\_interval\_start    day since start of study when the
    interval starts
day\_interval\_stop     day since start of study when the
    interval stops
infection\_event       If an infection occured during this
    time interval, 1=yes, 0=no

Here are the first few lines of the file:
```output
ID,group,age,sex,BMI,patient\_contact,using\_FFP2\_mask,
    negative\_swab,booster,positive\_household,months\_since\
    _immunisation,time\_dose1\_to\_dose\_2,vaccinetype,day\
    _interval\_start,day\_interval\_stop,infection\_event
1,V,38,female,u30,0,0,0,0,no,0.8,1.2,Moderna,0,87,0
1,V,38,female,u30,0,0,0,0,no,0.8,1.2,Moderna,87,99,0
1,V,38,female,u30,0,0,0,0,no,0.8,1.2,Moderna,99,113,0

```


\#\#\# File 2: "Symptoms.csv"
Data in the file "Symptoms.csv" is organised per infection
    event, consisting in total of 764 events. Each worker is
    only indicated once. It contains 11 columns:
ID      Unique Identifier, same in both files
~~~

```
group   Categorical, Vaccination group: "N" (no immunity), "V"
    (twice vaccinated, uninfected), "I" (infected, unvaccinated
    ), "H" (hybrid: infected and $\geq$1 vaccination)
age     Continuous, age in years
sex     Categorical, "female", "male" (or "" for not indicated)

BMI     Categorical, "o30" for $>$30 or "u30" for under 30

comorbidity catgeorical, if any comorbity pre-existed, 1=yes,
    0=no
using\_FFP2\_mask        Always using protective respiratory
    masks during work, 1=yes, 0=no
months\_until\_reinfection       time until next infection in
    months
variant Categorical, "delta" or "omicron" (or "" for not
    indicated)
booster\_over7\_days\_before     If a booster was given in the
    last 7 days before the infection
symptom\_number Continous, Number of symptoms which occured
    after the infection

Here are the first few lines of the file:
‘‘‘output
ID,group,age,sex,BMI,comorbidity,using\_FFP2\_mask,months\
    _until\_reinfection,variant,booster\_over7\_days\_before,
    symptom\_number
2,N,45,female,u30,0,0,2.5,delta,0,11
3,V,58,female,u30,1,0,4.2,omicron,0,6
7,V,32,female,u30,0,1,4.5,omicron,1,5

‘‘‘
```

# B Data Exploration

## B.1 Code

The Data Exploration was carried out using the following custom code:

```python
# Import required packages
import pandas as pd
import numpy as np

# Load datasets
df_time_to_infection = pd.read_csv('TimeToInfection.csv')
df_symptoms = pd.read_csv('Symptoms.csv')
```

```
# Creating an output text file to print the summary
with open(’data_exploration.txt’, ’w’) as f:
    # Data Size
    f.write("# Data Size\n")
    f.write(f"Time to Infection Data size: {
       ↪ df_time_to_infection.shape[0]} rows, {
       ↪ df_time_to_infection.shape[1]} columns\n")
    f.write(f"Symptoms Data size: {df_symptoms.shape[0]} rows,
       ↪ {df_symptoms.shape[1]} columns\n\n")

    # Summary Statistics
    f.write("# Summary Statistics\n")
    f.write("Time to Infection Data:\n")
    f.write(str(df_time_to_infection.describe(include=’all’)) +
       ↪  "\n\n")
    f.write("Symptoms Data:\n")
    f.write(str(df_symptoms.describe(include=’all’)) + "\n\n")

    # Categorical Variables
    f.write("# Categorical Variables\n")
    f.write("Time to Infection Data:\n")
    for col in df_time_to_infection.select_dtypes(include=[’
       ↪ object’]).columns:
        f.write(f"Most common value on {col}: {
           ↪ df_time_to_infection[col].mode()[0]}\n")
    f.write("Symptoms Data:\n")
    for col in df_symptoms.select_dtypes(include=[’object’]).
       ↪ columns:
        f.write(f"Most common value on {col}: {df_symptoms[col
           ↪ ].mode()[0]}\n")
    f.write("\n")

    # Missing Values
    f.write("# Missing Values\n")
    f.write("Time to Infection Data:\n")
    f.write(str(df_time_to_infection.isnull().sum()) + "\n")
    f.write("Symptoms Data:\n")
    f.write(str(df_symptoms.isnull().sum()) + "\n\n")

f.close()
```

### B.2 Code Description

The provided Python code conducts data exploration on two datasets, ’TimeToInfection’ and ’Symptoms’. Firstly, it calculates and writes the data sizes of both datasets. Next, it generates summary statistics for each dataset, including mean, median, min, and max values for numerical columns. The code then identifies and prints the most common values for each categorical variable in the datasets. Moreover, it detects and reports any missing values in the datasets. Lastly, all the exploration results are written into a text file 'data_exploration.txt' for reference and further analysis. This file includes data sizes, summary statistics, common values in categorical variables, and information on missing values for both datasets.

### B.3 Code Output

#### data_exploration.txt

```
\# Data Size
Time to Infection Data size: 12086 rows, 16 columns
Symptoms Data size: 764 rows, 11 columns

\# Summary Statistics
Time to Infection Data:
          ID  group   age     sex    BMI  patient\_contact
             using\_FFP2\_mask  negative\_swab  booster
             positive\_household  months\_since\_immunisation
              time\_dose1\_to\_dose\_2      vaccinetype  day\
             _interval\_start  day\_interval\_stop   infection\
             _event
count  12086  12086 12065   11987  12086            11686
              11686          12086    12086              12086
                         11459                  9332
              10035               12086              12086
              12086
unique   NaN      4   NaN       2      2              NaN
                NaN            NaN      NaN                  2
                           NaN                 NaN
                  3                NaN                NaN
                NaN
top      NaN      V   NaN  female    u30              NaN
                NaN            NaN      NaN                 no
                           NaN                 NaN  Pfizer\
    _BioNTech                  NaN              NaN
                NaN
freq     NaN   8157   NaN    9617  10625              NaN
                NaN            NaN      NaN              10584
                           NaN                 NaN
               7816                NaN                NaN
                NaN
mean    1300    NaN 44.03     NaN    NaN           0.7941
             0.2014         0.4933   0.5007                 NaN
                         5.015                1.026
                NaN             81.21             113.2
            0.06321
std    748.2    NaN 11.01     NaN    NaN           0.4044
             0.4011            0.5      0.5                NaN
                         2.344                0.4213
                NaN               47.03               32.1
             0.2434
min        1    NaN    17     NaN    NaN                0
                  0              0        0                NaN
                          -5.3                     0
                NaN                   0                  1
                  0
25\%      648    NaN    35     NaN    NaN                1
                  0              0        0                NaN
                           3.8                   0.9
                NaN                  75                 88
                  0
50\%     1310    NaN    44     NaN    NaN                1
                  0              0        1                NaN
                           5.5                     1
                NaN                  99                106
                  0
75\%     1942    NaN    53     NaN    NaN                1
                  0              1        1                NaN
                           6.6                   1.2
                NaN                 113                142
                  0
max     2595    NaN    73     NaN    NaN                1
                  1              1        1                NaN
                          17.8                   5.1
                NaN                 171                178
                  1

Symptoms Data:
          ID group   age     sex  BMI  comorbidity  using\_FFP2
             \_mask  months\_until\_reinfection  variant
             booster\_over7\_days\_before  symptom\_number
count    764   764   764     759  764          719
                734                       764      764
                          764             764
unique   NaN     4   NaN       2    2          NaN
                NaN                       NaN        2
                          NaN             NaN
top      NaN     V   NaN  female  u30          NaN
                NaN                       NaN  omicron
                          NaN             NaN
freq     NaN   550   NaN     620  679          NaN
                NaN                       NaN      591
                          NaN             NaN
```

```
mean    1315   NaN 41.45     NaN  NaN       0.3825
    0.1839                    4.1      NaN
                       0.5209           3.806
std    742.7   NaN 10.69     NaN  NaN       0.4863
    0.3877                  1.268      NaN
                       0.4999           2.177
min        2   NaN    17     NaN  NaN           0
                  0                         0      NaN
                            0               0
25\%    694.8   NaN    33     NaN  NaN            0
                  0                       3.4      NaN
                            0               2
50\%     1354   NaN    41     NaN  NaN            0
                  0                       4.4      NaN
                            1               4
75\%     1942   NaN 49.25     NaN  NaN            1
                  0                       5.2      NaN
                            1               5
max     2588   NaN    68     NaN  NaN            1
                  1                       5.8      NaN
                            1              11

# Categorical Variables
Time to Infection Data:
Most common value on group: V
Most common value on sex: female
Most common value on BMI: u30
Most common value on positive_household: no
Most common value on vaccinetype: Pfizer_BioNTech
Symptoms Data:
Most common value on group: V
Most common value on sex: female
Most common value on BMI: u30
Most common value on variant: omicron

# Missing Values
Time to Infection Data:
ID                              0
group                           0
age                            21
sex                            99
BMI                             0
patient_contact                400
using_FFP2_mask                 400
negative_swab                    0
booster                         0
positive_household               0
months_since_immunisation       627
time_dose1_to_dose_2             2754
```

```
vaccinetype                  2051
day\_interval\_start              0
day\_interval\_stop               0
infection\_event                0
dtype: int64
Symptoms Data:
ID                            0
group                         0
age                           0
sex                           5
BMI                           0
comorbidity                  45
using\_FFP2\_mask               30
months\_until\_reinfection        0
variant                       0
booster\_over7\_days\_before        0
symptom\_number                 0
dtype: int64
```

# C Data Analysis

## C.1 Code

The Data Analysis was carried out using the following custom code:

```
# IMPORT
import pandas as pd
import numpy as np
from statsmodels.formula.api import glm
import statsmodels.api as sm
import pickle

# LOAD DATA
time_to_infection_data = pd.read_csv('TimeToInfection.csv')
symptoms_data = pd.read_csv('Symptoms.csv')

# DATASET PREPARATIONS
time_to_infection_data.fillna(np.nan, inplace=True)
symptoms_data.fillna(np.nan, inplace=True)

merged_data = pd.merge(time_to_infection_data, symptoms_data,
    ↪ how = 'inner', on = ['ID','group','age','sex','BMI'],
    ↪ suffixes= ('', '_symptoms'))

# DESCRIPTIVE STATISTICS
# No descriptive statistics are needed.

# PREPROCESSING
```

```
merged_data = pd.concat([merged_data,pd.get_dummies(merged_data
    ↪ ['group'], prefix='group')], axis=1)
merged_data = pd.concat([merged_data,pd.get_dummies(merged_data
    ↪ ['sex'], prefix='sex')], axis=1)
merged_data = pd.concat([merged_data,pd.get_dummies(merged_data
    ↪ ['BMI'], prefix='BMI')], axis=1)
merged_data = merged_data.drop(['group', 'sex', 'BMI', '
    ↪ vaccinetype', 'group_H'], axis = 1)

# ANALYSIS
## Table 1: "Association between vaccination status and time
    ↪ until reinfection"
formula1 = 'months_until_reinfection ~ age + group_N + group_V
    ↪ + group_I + sex_female + BMI_u30 + patient_contact +
    ↪ using_FFP2_mask'
model1 = glm(formula1, data=merged_data, family=sm.families.
    ↪ Gaussian()).fit()
df1 = pd.DataFrame(model1.summary2().tables[1])
df1.to_pickle('table_1.pkl')

## Table 2: "Association between vaccination status and number
    ↪ of symptoms at reinfection"
formula2 = 'symptom_number ~ age + group_N + group_V + group_I
    ↪ + sex_female + BMI_u30 + patient_contact +
    ↪ using_FFP2_mask'
model2 = glm(formula2, data=merged_data, family=sm.families.
    ↪ Gaussian()).fit()
df2 = pd.DataFrame(model2.summary2().tables[1])
df2.to_pickle('table_2.pkl')

# SAVE ADDITIONAL RESULTS
additional_results = {
    'Total number of observations': len(merged_data),
}
with open('additional_results.pkl', 'wb') as f:
    pickle.dump(additional_results, f)
```

## C.2 Code Description

The provided code conducts a data analysis on the merged dataset containing information on health workers' time to reinfection and symptoms during reinfection based on their vaccination status and other demographic variables.

After loading and merging the data, the code preprocesses the dataset by creating dummy variables for categorical features and dropping unnecessary columns.

Two generalized linear models (GLMs) are then fitted to analyze the association between vaccination status and different outcomes: 1. Table 1 examines the association between vaccination status and time until reinfection, considering factors such as age, sex, BMI, patient contact, and FFP2 mask usage. 2. Table 2 investigates the association between vaccination status and the number of symptoms experienced during reinfection, controlling for similar demographic variables.

The results of these analyses, presented as summary tables, are saved as pickled files "table_1.pkl" and "table_2.pkl", respectively.

Additionally, the code saves the total number of observations in the dataset as 'Total number of observations' in the "additional_results.pkl" file for further reference or reporting.

## C.3 Code Output

### table_1.pkl

```
                     Coef.  Std.Err.      z     P$>$\textbar{}
                        z\textbar{}    [0.025   0.975]
Intercept            4.558     0.113  40.33          0
   4.336    4.779
group\_N[T.True]     -0.6467   0.08314 -7.779  7.32e-15
   -0.8097  -0.4838
group\_V[T.True]    -0.05593   0.04811 -1.163     0.245
   -0.1502  0.03837
group\_I[T.True]     -0.1033   0.09519 -1.085     0.278
   -0.2898   0.0833
sex\_female[T.True]  0.03462   0.04357 0.7945     0.427
   -0.05078      0.12
BMI\_u30[T.True]    -0.08809   0.05267 -1.672    0.0944
   -0.1913  0.01514
age                0.003675  0.001699  2.163    0.0305
   0.0003454 0.007004
patient\_contact     0.05861   0.04366  1.342     0.179
   -0.02696   0.1442
using\_FFP2\_mask     -0.1205   0.04546  -2.65   0.00805
   -0.2096 -0.03137
```

### table_2.pkl

```
                      Coef.  Std.Err.      z     P$>$\textbar
                         {}z\textbar{}   [0.025     0.975]
Intercept             3.211    0.2597  12.36  4.14e-35
   2.702        3.72
group\_N[T.True]       1.263     0.191  6.609  3.87e-11
   0.8882      1.637
```

```
group\_V[T.True]        0.7837    0.1106   7.089  1.35e-12
    0.5671          1
group\_I[T.True]        0.9527    0.2187   4.355  1.33e-05
    0.5239      1.381
sex\_female[T.True]      0.298    0.1001   2.976   0.00292
    0.1017     0.4942
BMI\_u30[T.True]       -0.2321     0.121  -1.918    0.0551
    -0.4693   0.005109
age                  -0.008532  0.003903  -2.186    0.0288
    -0.01618  -0.0008819
patient\_contact        0.1576    0.1003    1.57     0.116
    -0.03908     0.3542
using\_FFP2\_mask      -0.1084    0.1045  -1.038     0.299
    -0.3131    0.09635
```

### additional_results.pkl

```
{
    'Total number of observations': 2947,
}
```

# D LaTeX Table Design

## D.1 Code

The LaTeX Table Design was carried out using the following custom code:

```python
# IMPORT
import pandas as pd
from my_utils import to_latex_with_note, is_str_in_df,
    ↪ split_mapping, AbbrToNameDef

# PREPARATION FOR ALL TABLES
# Shared mapping for labels that are common to both tables
shared_mapping: AbbrToNameDef = {
    'group_N[T.True]': ('Not Immun', None),
    'group_V[T.True]': ('Vacc', None),
    'group_I[T.True]': ('Inf', None),
    'sex_female[T.True]': ('Fem', '1: Yes, 0: No'),
    'BMI_u30[T.True]': ('BMI < 30', '1: <30, 0: >=30'),
    'patient_contact': ('P. Contact', '1: Yes, 0: No'),
    'using_FFP2_mask': ('Use FFP2', '1: Yes, 0: No'),
}

# TABLE 1:
df1 = pd.read_pickle('table_1.pkl')

# RENAME ROWS AND COLUMNS
```

```
mapping1 = dict((k, v) for k, v in shared_mapping.items() if
    ↪ is_str_in_df(df1, k))
mapping1.update({'Coef.': ('Coef', None),
                 'Std.Err.': ('SE', None),
                 'z': ('Z-Score', None),
                 'P>|z|': ('P-value', None),
                 'age': ('Age', 'years')})

abbrs_to_names1, legend1 = split_mapping(mapping1)
df1 = df1.rename(columns=abbrs_to_names1, index=abbrs_to_names1
    ↪ )

# SAVE AS LATEX
to_latex_with_note(
    df1, 'table_1.tex',
    caption="Association between various factors and time until
        ↪  reinfection",
    label='table:time_until_reinfection',
    note="Statistical analyses performed using Generalized
        ↪ Linear Models (GLM).",
    legend=legend1)

# TABLE 2:
df2 = pd.read_pickle('table_2.pkl')

# RENAME ROWS AND COLUMNS
mapping2 = dict((k, v) for k, v in shared_mapping.items() if
    ↪ is_str_in_df(df2, k))
mapping2.update({'Coef.': ('Coef', None),
                 'Std.Err.': ('SE', None),
                 'z': ('Z-Score', None),
                 'P>|z|': ('P-value', None),
                 'age': ('Age', 'years')})

abbrs_to_names2, legend2 = split_mapping(mapping2)
df2 = df2.rename(columns=abbrs_to_names2, index=abbrs_to_names2
    ↪ )

# SAVE AS LATEX
to_latex_with_note(
    df2, 'table_2.tex',
    caption="Association between vaccination status and number
        ↪ of symptoms at reinfection",
    label='table:symptoms_at_reinfection',
    note="Statistical analyses performed using Generalized
        ↪ Linear Models (GLM).",
    legend=legend2,
    columns=['Coef', 'SE', 'P-value'])
```

## D.2 Provided Code

The code above is using the following provided functions:

```
def to_latex_with_note(df, filename: str, caption: str, label:
    ↪ str, note: str = None, legend: Dict[str, str] = None, **
    ↪ kwargs):
    """
    Converts a DataFrame to a LaTeX table with optional note
        ↪ and legend added below the table.

    Parameters:
    - df, filename, caption, label: as in ‘df.to_latex‘.
    - note (optional): Additional note below the table.
    - legend (optional): Dictionary mapping abbreviations to
        ↪ full names.
    - **kwargs: Additional arguments for ‘df.to_latex‘.
    """

def is_str_in_df(df: pd.DataFrame, s: str):
    return any(s in level for level in getattr(df.index, '
        ↪ levels', [df.index]) + getattr(df.columns, 'levels',
        ↪ [df.columns]))

AbbrToNameDef = Dict[Any, Tuple[Optional[str], Optional[str]]]

def split_mapping(abbrs_to_names_and_definitions: AbbrToNameDef
    ↪ ):
    abbrs_to_names = {abbr: name for abbr, (name, definition)
        ↪ in abbrs_to_names_and_definitions.items() if name is
        ↪ not None}
    names_to_definitions = {name or abbr: definition for abbr,
        ↪ (name, definition) in abbrs_to_names_and_definitions.
        ↪ items() if definition is not None}
    return abbrs_to_names, names_to_definitions
```

## D.3 Code Output

### table_1.tex

```
\% This latex table was generated from: ‘table\_1.pkl‘
\begin{table}[h]
\caption{Association between various factors and time until
   reinfection}
\label{table:time\_until\_reinfection}
\begin{threeparttable}
\renewcommand{\TPTminimum}{\linewidth}
\makebox[\linewidth]{\%
\begin{tabular}{lrrrlrr}
\toprule
 \& Coef \& SE \& Z-Score \& P-value \& [0.025 \& 0.975] \\
\midrule
\textbf{Intercept} \& 4.56 \& 0.113 \& 40.3 \& \$$<$\$1e-06 \&
    4.34 \& 4.78 \\
\textbf{Not Immun} \& -0.647 \& 0.0831 \& -7.78 \& \$$<$\$1e-06
     \& -0.81 \& -0.484 \\
\textbf{Vacc} \& -0.0559 \& 0.0481 \& -1.16 \& 0.245 \& -0.15
    \& 0.0384 \\
\textbf{Inf} \& -0.103 \& 0.0952 \& -1.08 \& 0.278 \& -0.29 \&
    0.0833 \\
\textbf{Fem} \& 0.0346 \& 0.0436 \& 0.795 \& 0.427 \& -0.0508
    \& 0.12 \\
\textbf{BMI \$$<$\$ 30} \& -0.0881 \& 0.0527 \& -1.67 \& 0.0944
     \& -0.191 \& 0.0151 \\
\textbf{Age} \& 0.00367 \& 0.0017 \& 2.16 \& 0.0305 \& 0.000345
     \& 0.007 \\
\textbf{P. Contact} \& 0.0586 \& 0.0437 \& 1.34 \& 0.179 \&
    -0.027 \& 0.144 \\
\textbf{Use FFP2} \& -0.12 \& 0.0455 \& -2.65 \& 0.00805 \&
    -0.21 \& -0.0314 \\
\bottomrule
\end{tabular}}
\begin{tablenotes}
\footnotesize
\item Statistical analyses performed using Generalized Linear
    Models (GLM).
\item \textbf{Fem}: 1: Yes, 0: No
\item \textbf{BMI \$$<$\$ 30}: 1: \$$<$\$30, 0: \$$>$\$=30
\item \textbf{P. Contact}: 1: Yes, 0: No
\item \textbf{Use FFP2}: 1: Yes, 0: No
\item \textbf{Age}: years
\end{tablenotes}
\end{threeparttable}
\end{table}
```

### table_2.tex

```
\% This latex table was generated from: `table\_2.pkl`
\begin{table}[h]
\caption{Association between vaccination status and number of
    symptoms at reinfection}
\label{table:symptoms\_at\_reinfection}
\begin{threeparttable}
\renewcommand{\TPTminimum}{\linewidth}
\makebox[\linewidth]{\%
\begin{tabular}{lrrl}
\toprule
```

```
 \& Coef \& SE \& P-value \\
\midrule
\textbf{Intercept} \& 3.21 \& 0.26 \& \$$<$\$1e-06 \\
\textbf{Not Immun} \& 1.26 \& 0.191 \& \$$<$\$1e-06 \\
\textbf{Vacc} \& 0.784 \& 0.111 \& \$$<$\$1e-06 \\
\textbf{Inf} \& 0.953 \& 0.219 \& 1.33e-05 \\
\textbf{Fem} \& 0.298 \& 0.1 \& 0.00292 \\
\textbf{BMI \$$<$\$ 30} \& -0.232 \& 0.121 \& 0.0551 \\
\textbf{Age} \& -0.00853 \& 0.0039 \& 0.0288 \\
\textbf{P. Contact} \& 0.158 \& 0.1 \& 0.116 \\
\textbf{Use FFP2} \& -0.108 \& 0.104 \& 0.299 \\
\bottomrule
\end{tabular}}
\begin{tablenotes}
\footnotesize
\item Statistical analyses performed using Generalized Linear
   Models (GLM).
\item \textbf{Fem}: 1: Yes, 0: No
\item \textbf{BMI \$$<$\$ 30}: 1: \$$<$\$30, 0: \$$>$\$=30
\item \textbf{P. Contact}: 1: Yes, 0: No
\item \textbf{Use FFP2}: 1: Yes, 0: No
\item \textbf{Age}: years
\end{tablenotes}
\end{threeparttable}
\end{table}
```