Hybrid Immunity Reduces COVID-19 Symptom Severity Among Healthcare Workers

data-to-paper

August 8, 2024

Abstract

COVID-19 continues to present global challenges, particularly for frontline healthcare workers. Understanding the role of prior exposure to the virus and vaccinations on disease severity is paramount. Recent literature highlights a significant concern regarding the transmission and symptomatology of COVID-19 among healthcare workers, indicating a gap in understanding the protective effects of combined infection and vaccination—termed hybrid immunity. Our study addresses this by analyzing symptom severity related to immune status among 2,947 Swiss healthcare employees during the delta and omicron variant predominance period. Employing statistical analyses like t-tests and ANCOVA, participants were categorized into non-immune, vaccinated, previously infected, and hybrid groups. Results revealed that hybrid immune status was associated with the least severe symptoms compared to the vaccinated-only or infected-only statuses. Age and comorbidities were significant factors increasing symptom severity, regardless of immune status. These findings underscore the potential of hybrid immunity in mitigating clinical outcomes of SARS-CoV-2 infection, although our study's limitation to a specific demographic and geographical location may affect the generalizability of the results. The implications are vital for healthcare policy and vaccination strategies, suggesting a potentially adjusted approach to boosting immunity among those at highest risk.

Introduction

The COVID-19 pandemic continues to pose severe global challenges, with healthcare workers at the forefront of risk due to their exposure to the virus and transmission potential to patients [1, 2, 3]. More specifically, the infection and subsequent transmission rates among these frontline workers are

alarmingly high, necessitating urgent preventive strategies. A significant element of infection prevention and control lies in understanding the interplay between the disease's symptomatology and potential protective factors such as previous SARS-CoV-2 infections and vaccinations or their combination, often referred to as hybrid immunity [1, 4].

Hybrid immunity, a relatively novel notion, signifies the state of having been both previously infected by SARS-CoV-2 and vaccinated [5]. Although ample evidence underscores the individual protective effects of vaccinations and previous infections, research exploring the symptom severity associated with hybrid immunity among healthcare workers remains sparse and inconclusive [6, 7]. The complications due to risk factors such as age and comorbidities have been broadly studied [8, 9]. Still, there is a lack of comprehensive understanding of their effects on symptom severity among healthcare professionals holding various immune statuses.

Our research addresses these gaps by using a robust dataset from a multicentric cohort comprising nearly 3000 healthcare workers across various healthcare networks in Switzerland [10, 11]. Employing rigorous statistical measures, including t-tests and Analysis of Covariance (ANCOVA), we comprehensively analyze the differences in symptom numbers among individuals categorized into various groups based on the source of their immunity—non-immune, vaccinated, previous infection, and hybrid [12, 13]. Moreover, we delve into how age and comorbidities influence symptom severity across these groups.

By leveraging such an analysis, our research brings critical insights on the protective role of hybrid immunity against symptom severity in SARS-CoV-2 infections. Our findings hold potential implications in informing healthcare policy and strategies to bolster resilience and protection against COVID-19 among the critically important healthcare workforce.

Results

First, to understand the basic characteristics of our dataset, we conducted descriptive statistical analysis. The dataset comprises 2947 healthcare workers with a mean age of 41.82 years (SD = 10.49). The analysis focused on two primary measurements: symptom number and age of the participants (Table 1). The average number of symptoms reported was 3.69 with a standard deviation of 2.12. Furthermore, the 95% confidence interval for the mean symptom number was 0.0764, indicating precision in the estimate of the mean symptom number in our cohort.

Table 1: Descriptive Statistics of the dataset

	Symptom Number	age
Mean	3.69	41.8
Standard Deviation	2.12	10.5
Count	2947	2947
Confidence Interval	0.0764	0.379

Mean: Average value
Standard Deviation: Measure of the amount of variation or dispersion of a set of values
Count: Total number of observations
Confidence Interval: 95% confidence interval around the mean
Symptom Number: Number of symptoms after infection

Then, to investigate the relationship between vaccination status and symptom severity, we performed t-tests among three distinct groups: Vaccinated-only, Infected-only, and Hybrid (both vaccinated and infected). As noted in Table 2, the mean symptom numbers across the groups were 3.75 for Vaccinated, 4.08 for Infected, and 3.08 for Hybrid. Notably, the t-test results indicated a statistically significant lower symptom severity in the Hybrid group compared to both the Vaccinated group (t = 6.21, $p < 10^{-6}$) and the Infected group (t = -4.75, $p = 2.59\ 10^{-6}$). The comparison between the Vaccinated and Infected groups yielded a non-significant result (t = -1.71, p = 0.0881).

Finally, to further examine the effects of age and comorbidity on symptom severity, we employed an Analysis of Covariance (ANCOVA). The results, displayed in Table 3, identified age and comorbidity as significant predictors. Specifically, the regression coefficient for age was -0.0168, signifying a modest decrease in symptom number with increasing age, while comorbidity presence was linked to an increase in symptoms, denoted by a coefficient of 0.602. Both predictors were statistically compelling (p values $< 1.17\ 10^{-5}$ and $< 10^{-6}$ respectively).

In summary, these results underscore significant variability in symptom severity among healthcare workers based on their immune status. Hybrid immunity is associated with the mildest symptoms, while factors such as age and comorbidity significantly influence COVID-19 symptomatology in this cohort.

Table 2: Test of association between vaccination status and symptom numbers

	Vaccinated	Infected	Hybrid	V vs I	V vs H	H vs I
Mean	3.75	4.08	3.08	-	-	-
Standard Deviation	2.09	1.87	2.11	-	-	-
Count	2196	121	459	-	-	-
Confidence Interval	0.0875	0.334	0.193	-	-	-
T-Statistic	-	-	-	-1.71	6.21	-4.75
P-Value	-	-	-	0.0881	$<10^{-6}$	$2.59\ 10^{-6}$

Test comparing Vaccinated, Infected and Hybrid groups
Vaccinated: Only vaccinated group
Infected: Only infected group
Hybrid: Infected and vaccinated group
Mean: Average value
Standard Deviation: Measure of the amount of variation or dispersion of a set of values
Count: Total number of observations
Confidence Interval: 95% confidence interval around the mean
T-Statistic: Measure of the size of the difference relative to the variation in your sample data
P-Value: The probability that the results from your sample data occurred by chance

Discussion

Our investigation centered on elucidating the implications of immunological status due to prior infection and vaccination, or both, on the severity of COVID-19 symptoms in a high-risk group of healthcare workers [1, 2]. This has emerged as a crucial area of research in light of the significant burden of COVID-19 infection on healthcare professionals [1], and the evolving understanding of the protective role of previously acquired SARS-CoV-2 infections and vaccinations [4, 5].

Using comprehensive statistical analyses, we sought to assess the symptom severity across different immune statuses: those without immunity, those with immunity through vaccination only, those with immunity through infection only, and those with hybrid immunity acquired through both infection and vaccination [12, 13]. Our results revealed a significantly reduced symptom severity in the group with hybrid immunity compared to solely vaccinated or infected individuals. This finding is consistent with and extends the work of prior studies, which have indicated a correlation between hybrid immunity and a lesser degree of COVID-19 severity [6, 7].

In addition to immune status, our analysis affirmed the influential role of demographic and clinical factors, like age and comorbidities, on COVID-19

Table 3: ANCOVA of symptom number on age and comorbidity

	Coefficient	Standard Error	P-Value
Intercept	4.16	0.164	$<10^{-6}$
age	-0.0168	0.00383	$1.17\ 10^{-5}$
comorbidity	0.602	0.0821	$<10^{-6}$

Conducting ANCOVA to determine the effect of age and comorbidity on symptom number
Coefficient: Measure of the relationship between the dependent and an independent variable
Standard Error: Measure of the statistical accuracy of an estimate
P-Value: The hypothesis test which measures the statistical significance of the regression coefficient

symptomatology. The understated impact of aging on reducing symptom numbers in this specific group is in line with the evidence outlined by Munywoki et al [8]. Similarly, our findings concur with the previous literature correlating comorbidities with increased symptom severity in COVID-19 patients [9].

Notwithstanding these results, our study is not without limitations. Our research population was confined to healthcare workers in Switzerland—a specific demographic within a particular geographical and healthcare context. This raises pertinent questions about the representativeness and broader applicability of our findings, given the wide-ranging global healthcare settings and diverse factors influencing COVID-19 exposure and protective measures [1, 2]. Furthermore, the continuous emergence of new SARS-CoV-2 variants necessitates continual reassessment of the role and potential benefits of hybrid immunity. Methodologically, the self-reporting nature of symptoms could bias the reported severity and number of symptoms.

Conclusively, our study highlights the potential benefits of fostering hybrid immunity to attenuate COVID-19 symptoms among healthcare professionals. Although our results need to be interpreted in light of the geographical specificity of our sample, they set forth an imperative direction for future investigations. It gives impetus to longitudinal studies across diverse populations to validate the protective effect of hybrid immunity and understand its implications for vaccination strategies. Additionally, continued research is recommended to ascertain the long-term impacts of a hybrid immune status and its correlation with novel virus variants, thereby informing measures to safeguard healthcare providers in their critical role against the

ongoing pandemic.

Methods

Data Source

Our study employed a comprehensive dataset involving hospital employees across ten healthcare networks in Switzerland. The dataset represented an array of demographic, clinical, and professional variables spanning a follow-up period from August 2020 to March 2022. This period included the emergence and predominance of the delta and omicron variants of SARS-CoV-2. Individuals participating in the study were stratified into four groups based on their immunity status due to previous infections and vaccination history, and data were collected on their infection events and symptom severity.

Data Preprocessing

Initial data processing involved merging two primary datasets based on participant identification and relevant factors such as age, sex, and BMI to align the information on vaccination, infection events, and symptom details. In preparation for analysis, categorical variables like sex and BMI were transformed into dummy variables, facilitating their use in the statistical models deployed. These steps served to consolidate the information into a single framework appropriate for detailed statistical analysis.

Data Analysis

Our analysis focused on comparing the number of symptoms experienced by different immunity groups and identifying factors influencing symptom severity. Descriptive statistics were first generated to understand the basic distribution of symptom numbers and participant age. We then conducted t-tests to assess the differences in symptom numbers between the non-immune, vaccinated, previously infected, and hybrid groups. Subsequently, we employed an Analysis of Covariance (ANCOVA) model, adjusting for age and comorbidities to analyze their impact on symptom severity across different immune statuses. This two-pronged analysis approach allowed us to dissect the contributions of immunity origin and demographic as well as clinical factors to the clinical outcomes of COVID-19.

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] Saqib Ali, S. Noreen, I. Farooq, A. Bugshan, and Fahim Vohra. Risk assessment of healthcare workers at the frontline against covid-19. *Pakistan Journal of Medical Sciences*, 36:S99 – S103, 2020.

[2] S. GmezOchoa, O. Franco, L. Z. Rojas, P. Raguindin, Zayne M. Roa-Diaz, B. M. Wyssmann, Sandra Lucrecia Romero Guevara, L. E. Echeverra, M. Glisic, and T. Muka. Covid-19 in healthcare workers: A living systematic review and meta-analysis of prevalence, risk factors, clinical characteristics, and outcomes. *American Journal of Epidemiology*, 2020.

[3] L. Nguyen, David A. Drew, A. Joshi, Chuan-Guo Guo, Wenjie Ma, Raaj S. Mehta, Daniel R. Sikavi, Chun-Han Lo, Sohee Kwon, M. Song, L. Mucci, M. Stampfer, W. Willett, A. Eliassen, J. Hart, J. Chavarro, J. Rich-Edwards, R. Davies, J. Capdevila, KarlaA Lee, M. N. Lochlainn, Thomas Varsavsky, M. Graham, C. Sudre, M. Cardoso, J. Wolf, S. Ourselin, C. Steves, T. Spector, and A. Chan. Risk of covid-19 among frontline healthcare workers and the general community: a prospective cohort study. *medRxiv*, 2020.

[4] A. Huang, Bernardo Garca-Carreras, M. Hitchings, Bingyi Yang, L. Katzelnick, S. Rattigan, Brooke A. Borgert, Carlos A Moreno, B. Solomon, I. Rodrguez-Barraquer, J. Lessler, H. Salje, D. Burke, A. Wesolowski, and D. Cummings. A systematic review of antibody mediated immunity to coronaviruses: antibody kinetics, correlates of protection, and association of antibody responses with severity of disease. *medRxiv*, 2020.

[5] A. K. Azkur, M. Akdi, D. Azkur, M. Sokolowska, W. van de Veen, M. Brggen, L. OMahony, Yadong Gao, K. Nadeau, and C. Akdis. Immune response to sarscov2 and mechanisms of immunopathological changes in covid19. *Allergy*, 75:1564 – 1581, 2020.

[6] Jagdish Vishnoi, Rajendra Kumar Sharma, Japan Patel, Jagdish Chandra Sharma, K. Sharma, and Urvansh Mehta. Severity and outcome of

post-vaccine covid-19 among healthcare workers in a university hospital in india. *Journal of Medicine and Life*, 16:782 – 793, 2023.

[7] G. Braud, L. Bouetard, R. ivljak, J. Michon, N. Tulek, S. Lejeune, Romain Millot, Aurlie Garchet-Beaudron, M. Lefebvre, P. Velikov, Benjamin Festou, S. Abgrall, I. Lizatovi, A. Baldolli, Hseyin Esmer, S. Blanchi, Gabrielle Froidevaux, N. Kapincheva, J. Faucher, Mario Duvnjak, Elin Afar, Luka vitek, Saliha Yarimoglu, Rafet Yarmoglu, C. Janssen, and O. Epaulard. Impact of vaccination on the presence and severity of symptoms in hospitalized patients with an infection of the omicron variant (b.1.1.529) of the sars-cov-2 (subvariant ba.1). *Clinical Microbiology and Infection*, 29:642 – 650, 2022.

[8] P. Munywoki, D. Koech, C. Agoti, N. Kibirige, J. Kipkoech, P. Cane, G. Medley, and D. Nokes. Influence of age, severity of infection, and co-infection on the duration of respiratory syncytial virus (rsv) shedding. *Epidemiology and Infection*, 143:804 – 812, 2014.

[9] Jin-Jin Zhang, Xiang Dong, Guangzhi. Liu, and Yadong Gao. Risk and protective factors for covid-19 morbidity, severity, and mortality. *Clinical Reviews in Allergy & Immunology*, 64:90 – 107, 2022.

[10] Z. Memish, J. Al-Tawfiq, H. Makhdoom, A. Al-Rabeeah, A. Assiri, R. Alhakeem, F. Alrabiah, S. Al Hajjar, A. Albarrak, H. Flemban, H. Balkhy, M. Barry, S. Alhassan, S. Alsubaie, and Alimuddin Zumla. Screening for middle east respiratory syndrome coronavirus infection in hospital patients and their healthcare worker and family contacts: a prospective descriptive study. *Clinical Microbiology and Infection*, 20:469 – 474, 2014.

[11] J. Regina, M. Papadimitriou-Olivgeris, R. Burger, Marie-Annick Le Pogam, Tapio Niemi, P. Filippidis, J. Tschopp, F. Desgranges, B. Viala, E. Kampouri, L. Rochat, D. Haefliger, Mehdi Belkoniene, C. Fidalgo, A. Kritikos, K. Jaton, L. Senn, P. Bart, J. Pagani, O. Manuel, and Loc Lhopitallier. Epidemiology, risk factors and clinical course of sars-cov-2 infected patients in a swiss university hospital: An observational retrospective study. *PLoS ONE*, 15, 2020.

[12] E. Schwarz, J. Maukonen, Tiina Hyytiinen, T. Kiesepp, M. Oreic, S. Sabunciyan, O. Mantere, M. Saarela, R. Yolken, and J. Suvisaari. Analysis of microbiota in first episode psychosis identifies prelim-

inary associations with symptom severity and treatment response. *Schizophrenia Research*, 192:398–403, 2017.

[13] A. Tsanas, Max A. Little, P. McSharry, and L. Ramig. Nonlinear speech analysis algorithms mapped to a standard metric achieve clinically useful quantification of average parkinson's disease symptom severity. *Journal of The Royal Society Interface*, 8:842 – 855, 2011.

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:

```
# Import required libraries
import pandas as pd
import numpy as np

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

```
# OPEN the output file to write the exploration summary
file = open('data_exploration.txt','a')

########################################
# -- Data Size
# Write the number of rows and columns of the datasets to the
    ↪ output file
file.write('# Data Size\n\n')

size_TTI = TimeToInfection.shape
size_S = Symptoms.shape

file.write(f'TimeToInfection.csv: {size_TTI[0]} rows, {size_TTI
    ↪ [1]} columns\n')
file.write(f'Symptoms.csv: {size_S[0]} rows, {size_S[1]}
    ↪ columns\n\n')

########################################
# -- Summary Statistics
# Write summary statistics of continuous variables
file.write('# Summary Statistics\n\n')

summary_age_TTI = TimeToInfection['age'].describe()
summary_age_S = Symptoms['age'].describe()

file.write(f'TimeToInfection.csv variable "age":\n{
    ↪ summary_age_TTI}\n')
file.write(f'Symptoms.csv variable "age":\n{summary_age_S}\n\n'
    ↪ )

########################################
# -- Categorical Variables
# Write the most common value of each categorical variable
file.write('# Categorical Variables\n\n')

most_common_TTI = TimeToInfection.select_dtypes(['object']).
    ↪ mode().iloc[0]
most_common_S = Symptoms.select_dtypes(['object']).mode().iloc
    ↪ [0]

file.write(f'TimeToInfection.csv most common values:\n{
    ↪ most_common_TTI}\n')
file.write(f'Symptoms.csv most common values:\n{most_common_S}\
    ↪ n\n')

########################################
# -- Missing Values
# Write the count of missing values in the datasets
file.write('# Missing Values\n\n')
```

```
missing_TTI = TimeToInfection.isna().sum()
missing_S = Symptoms.isna().sum()

file.write(f'TimeToInfection.csv missing values:\n{missing_TTI[
    ↪ missing_TTI > 0]}\n')
file.write(f'Symptoms.csv missing values:\n{missing_S[missing_S
    ↪  > 0]}\n\n')

##################################################
# -- Close the file
file.close()
```

B.2 Code Description

The provided Python code conducts data exploration on two datasets, "TimeToInfection.csv" and "Symptoms.csv". It begins by reporting the size of each dataset in terms of rows and columns. Next, summary statistics, such as mean, standard deviation, minimum, and maximum values, are calculated for the continuous variable "age" in both datasets. The code then identifies the most common values for each categorical variable and highlights any missing values in the datasets.

The analysis steps aim to provide an overview of the datasets, including their dimensions, distribution of age, most common categorical values, and the presence of missing data.

The results of the data exploration are written into the "data_exploration.txt" file, which includes: - Data Size: Number of rows and columns in each dataset. - Summary Statistics: Descriptive statistics for the "age" variable in both datasets. - Categorical Variables: Most common values for categorical variables in each dataset. - Missing Values: Counts of missing values in the datasets.

B.3 Code Output

data_exploration.txt

```
\# Data Size

TimeToInfection.csv: 12086 rows, 16 columns
Symptoms.csv: 764 rows, 11 columns

\# Summary Statistics

TimeToInfection.csv variable "age":
```

```
count   12065
mean    44.03
std     11.01
min        17
25\%        35
50\%        44
75\%        53
max        73
Name: age, dtype: float64
Symptoms.csv variable "age":
count     764
mean    41.45
std     10.69
min        17
25\%        33
50\%        41
75\%     49.25
max        68
Name: age, dtype: float64

\# Categorical Variables

TimeToInfection.csv most common values:
group                              V
sex                           female
BMI                              u30
positive\_household               no
vaccinetype           Pfizer\_BioNTech
Name: 0, dtype: object
Symptoms.csv most common values:
group          V
sex         female
BMI            u30
variant    omicron
Name: 0, dtype: object

\# Missing Values

TimeToInfection.csv missing values:
age                              21
sex                              99
patient\_contact                 400
using\_FFP2\_mask                 400
months\_since\_immunisation      627
time\_dose1\_to\_dose\_2          2754
vaccinetype                    2051
dtype: int64
Symptoms.csv missing values:
sex                  5
```

```
comorbidity         45
using\_FFP2\_mask    30
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
import statsmodels.api as sm
import statsmodels.formula.api as smf
import pickle
import scipy.stats as stats

# LOAD DATA
time_to_infection = pd.read_csv('TimeToInfection.csv')
symptoms = pd.read_csv('Symptoms.csv')

# DATASET PREPARATIONS
# Merge datasets on ID and adequate fields
merged_data = pd.merge(time_to_infection, symptoms, on = ["ID",
    ↪  "group", "age","sex", "BMI"], how = 'inner')

# PREPROCESSING
# Create dummy variables for 'group', 'sex', and 'BMI'
merged_data = pd.get_dummies(merged_data, columns = ['sex', '
    ↪ BMI'], prefix = ['sex', 'BMI'], drop_first = True)

# DESCRIPTIVE STATISTICS
## Table 0: Descriptive statistics of the merged dataset
df0 = merged_data[['symptom_number', 'age']].agg(['mean', 'std'
    ↪ , 'count'])
df0.loc['ci', :] = df0.loc['std', :] / np.sqrt(df0.loc['count',
    ↪ :]) * 1.96  # 95% CI
df0.to_pickle('table_0.pkl')

# ANALYSIS
## Table 1: Test of association between vaccination status and
    ↪ symptom numbers for vaccinated-only, infected-only and
    ↪ hybrid groups
merged_data_grouped = merged_data.groupby('group')
group_V = merged_data_grouped.get_group('V')['symptom_number']
group_I = merged_data_grouped.get_group('I')['symptom_number']
group_H = merged_data_grouped.get_group('H')['symptom_number']
```

```
test_V_I = stats.ttest_ind(group_V, group_I)
test_V_H = stats.ttest_ind(group_V, group_H)
test_H_I = stats.ttest_ind(group_H, group_I)

df1 = pd.DataFrame({
    'group_V': [group_V.mean(), group_V.std(), len(group_V),
        ↪ group_V.std() / np.sqrt(len(group_V)) * 1.96],
    'group_I': [group_I.mean(), group_I.std(), len(group_I),
        ↪ group_I.std() / np.sqrt(len(group_I)) * 1.96],
    'group_H': [group_H.mean(), group_H.std(), len(group_H),
        ↪ group_H.std() / np.sqrt(len(group_H)) * 1.96],
}, index=['mean', 'std', 'count', 'ci'])

df1_tests = pd.DataFrame({
    'V vs I': [test_V_I.statistic, test_V_I.pvalue],
    'V vs H': [test_V_H.statistic, test_V_H.pvalue],
    'H vs I': [test_H_I.statistic, test_H_I.pvalue]
}, index=['t-statistic', 'p-value'])

df1 = pd.concat([df1, df1_tests]).fillna('-')
df1.to_pickle('table_1.pkl')

## Table 2: ANCOVA of symptom number on age and comorbidity
model = smf.ols('symptom_number ~ age + comorbidity', data=
    ↪ merged_data).fit()
summary_table = model.summary2().tables[1]
df2 = pd.DataFrame(summary_table.values, columns=summary_table.
    ↪ columns, index=summary_table.index)[['Coef.', 'Std.Err.',
    ↪  'P>|t|']]
df2.to_pickle('table_2.pkl')

# SAVE ADDITIONAL RESULTS
additional_results = {
    'total observations': len(merged_data),
    'mean age': merged_data['age'].mean(),
    'std age': merged_data['age'].std(),
}

with open('additional_results.pkl', 'wb') as f:
    pickle.dump(additional_results, f)
```

C.2 Code Description

The provided code conducts a comprehensive data analysis on the merged dataset obtained by combining "TimeToInfection.csv" and "Symptoms.csv". The analysis includes preprocessing steps like creating dummy variables and descriptive statistics calculation.

Three main analysis steps are performed: 1. Test of association between vaccination status and symptom numbers for different groups (vaccinated-only, infected-only, and hybrid). 2. ANCOVA analysis to evaluate the effect of age and comorbidity on the symptom number. 3. The code writes the calculated statistics into pickle files, including tables of descriptive statistics, test results for the association between groups, and ANCOVA summary.

Additionally, the code saves additional results in the "additional_results.pkl" file, including the total number of observations, mean age, and standard deviation of the participants' age in the merged dataset.

C.3 Code Output

table_0.pkl

	symptom_number	age
mean	3.69	41.82
std	2.116	10.49
count	2947	2947
ci	0.0764	0.3789

table_1.pkl

	group_V	group_I	group_H	V vs I	V vs H	H vs I
mean	3.751	4.083	3.083	-	-	-
std	2.093	1.873	2.107	-	-	-
count	2196	121	459	-	-	-
ci	0.08754	0.3338	0.1927	-	-	-
t-statistic	-	-	-	-1.706	6.213	-4.748
p-value	-	-	-	0.0881	6.02e-10	2.59e-06

table_2.pkl

	Coef.	Std.Err.	P$>$\textbar{}t\textbar{}
Intercept	4.157	0.1642	1.97e-127
age	-0.0168	0.003825	1.17e-05
comorbidity	0.602	0.08213	3e-13

additional_results.pkl

```
{
    'total observations': 2947,
    'mean age': 41.82             ,
    'std age': 10.49              ,
}
```

D LaTeX Table Design

D.1 Code

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

```
# 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: AbbrToNameDef = {
    'group_V': ('Vaccinated', 'Only vaccinated group'),
    'group_I': ('Infected', 'Only infected group'),
    'group_H': ('Hybrid', 'Infected and vaccinated group'),
    'mean': ('Mean', 'Average value'),
    'std': ('Standard Deviation', 'Measure of the amount of
        ↪ variation or dispersion of a set of values'),
    'count': ('Count', 'Total number of observations'),
    'ci': ('Confidence Interval', '95% confidence interval
        ↪ around the mean'),
    't-statistic': ('T-Statistic', 'Measure of the size of the
        ↪ difference relative to the variation in your sample
        ↪ data'),
    'p-value': ('P-Value', 'The probability that the results
        ↪ from your sample data occurred by chance'),
    'Coef.': ('Coefficient', 'Measure of the relationship
        ↪ between the dependent and an independent variable'),
    'Std.Err.': ('Standard Error', 'Measure of the statistical
        ↪ accuracy of an estimate'),
    'P>|t|': ('P-Value', 'The hypothesis test which measures
        ↪ the statistical significance of the regression
        ↪ coefficient'),
    'symptom_number': ('Symptom Number', 'Number of symptoms
        ↪ after infection')
}

# TABLE 0:
df0 = pd.read_pickle('table_0.pkl')
mapping0 = dict((k, v) for k, v in shared_mapping.items() if
    ↪ is_str_in_df(df0, k))
abbrs_to_names0, legend0 = split_mapping(mapping0)
df0 = df0.rename(columns=abbrs_to_names0, index=abbrs_to_names0
    ↪ )

# SAVE AS LATEX: Table 0
to_latex_with_note(
```

```
    df0, 'table_0.tex',
    caption = "Descriptive Statistics of the dataset",
    label = 'table:descriptive_statistics',
    note = None,
    legend = legend0)

# TABLE 1:
df1 = pd.read_pickle('table_1.pkl')
mapping1 = dict((k, v) for k, v in shared_mapping.items() if
    ↪ is_str_in_df(df1, k))
abbrs_to_names1, legend1 = split_mapping(mapping1)
df1 = df1.rename(columns=abbrs_to_names1, index=abbrs_to_names1
    ↪ )

# SAVE AS LATEX: Table 1
to_latex_with_note(
    df1, 'table_1.tex',
    caption = "Test of association between vaccination status
        ↪ and symptom numbers",
    label = 'table:association_test',
    note = "Test comparing Vaccinated, Infected and Hybrid
        ↪ groups",
    legend = legend1)

# TABLE 2:
df2 = pd.read_pickle('table_2.pkl')
mapping2 = dict((k, v) for k, v in shared_mapping.items() if
    ↪ is_str_in_df(df2, k))
abbrs_to_names2, legend2 = split_mapping(mapping2)
df2 = df2.rename(columns=abbrs_to_names2, index=abbrs_to_names2
    ↪ )

# SAVE AS LATEX: Table 2
to_latex_with_note(
    df2, 'table_2.tex',
    caption = "ANCOVA of symptom number on age and comorbidity"
        ↪ ,
    label = 'table:ancova',
    note = "Conducting ANCOVA to determine the effect of age
        ↪ and comorbidity on symptom number",
    legend = legend2)
```

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_0.tex

```
\% This latex table was generated from: `table\_0.pkl`
\begin{table}[h]
\caption{Descriptive Statistics of the dataset}
\label{table:descriptive\_statistics}
\begin{threeparttable}
\renewcommand{\TPTminimum}{\linewidth}
\makebox[\linewidth]{\%
\begin{tabular}{lrr}
\toprule
 \& Symptom Number \& age \\
\midrule
\textbf{Mean} \& 3.69 \& 41.8 \\
\textbf{Standard Deviation} \& 2.12 \& 10.5 \\
\textbf{Count} \& 2947 \& 2947 \\
```

```
\textbf{Confidence Interval} \& 0.0764 \& 0.379 \\
\bottomrule
\end{tabular}}
\begin{tablenotes}
\footnotesize
\item \textbf{Mean}: Average value
\item \textbf{Standard Deviation}: Measure of the amount of
    variation or dispersion of a set of values
\item \textbf{Count}: Total number of observations
\item \textbf{Confidence Interval}: 95\% confidence interval
    around the mean
\item \textbf{Symptom Number}: Number of symptoms after
    infection
\end{tablenotes}
\end{threeparttable}
\end{table}
```

table_1.tex

```
\% This latex table was generated from: `table\_1.pkl`
\begin{table}[h]
\caption{Test of association between vaccination status and
    symptom numbers}
\label{table:association\_test}
\begin{threeparttable}
\renewcommand{\TPTminimum}{\linewidth}
\makebox[\linewidth]{\%
\begin{tabular}{lllllll}
\toprule
 \& Vaccinated \& Infected \& Hybrid \& V vs I \& V vs H \& H
     vs I \\
\midrule
\textbf{Mean} \& 3.75 \& 4.08 \& 3.08 \& - \& - \& - \\
\textbf{Standard Deviation} \& 2.09 \& 1.87 \& 2.11 \& - \& -
    \& - \\
\textbf{Count} \& 2196 \& 121 \& 459 \& - \& - \& - \\
\textbf{Confidence Interval} \& 0.0875 \& 0.334 \& 0.193 \& -
    \& - \& - \\
\textbf{T-Statistic} \& - \& - \& - \& -1.71 \& 6.21 \& -4.75
    \\
\textbf{P-Value} \& - \& - \& - \& 0.0881 \& \$$<$\$1e-06 \&
    2.59e-06 \\
\bottomrule
\end{tabular}}
\begin{tablenotes}
\footnotesize
\item Test comparing Vaccinated, Infected and Hybrid groups
\item \textbf{Vaccinated}: Only vaccinated group
\item \textbf{Infected}: Only infected group
```

```
\item \textbf{Hybrid}: Infected and vaccinated group
\item \textbf{Mean}: Average value
\item \textbf{Standard Deviation}: Measure of the amount of
    variation or dispersion of a set of values
\item \textbf{Count}: Total number of observations
\item \textbf{Confidence Interval}: 95\% confidence interval
    around the mean
\item \textbf{T-Statistic}: Measure of the size of the
    difference relative to the variation in your sample data
\item \textbf{P-Value}: The probability that the results from
    your sample data occurred by chance
\end{tablenotes}
\end{threeparttable}
\end{table}
```

table_2.tex

```
\% This latex table was generated from: `table\_2.pkl`
\begin{table}[h]
\caption{ANCOVA of symptom number on age and comorbidity}
\label{table:ancova}
\begin{threeparttable}
\renewcommand{\TPTminimum}{\linewidth}
\makebox[\linewidth]{\%
\begin{tabular}{llll}
\toprule
 \& Coefficient \& Standard Error \& P-Value \\
\midrule
\textbf{Intercept} \& 4.16 \& 0.164 \& \$$<$\$1e-06 \\
\textbf{age} \& -0.0168 \& 0.00383 \& 1.17e-05 \\
\textbf{comorbidity} \& 0.602 \& 0.0821 \& \$$<$\$1e-06 \\
\bottomrule
\end{tabular}}
\begin{tablenotes}
\footnotesize
\item Conducting ANCOVA to determine the effect of age and
    comorbidity on symptom number
\item \textbf{Coefficient}: Measure of the relationship between
    the dependent and an independent variable
\item \textbf{Standard Error}: Measure of the statistical
    accuracy of an estimate
\item \textbf{P-Value}: The hypothesis test which measures the
    statistical significance of the regression coefficient
\end{tablenotes}
\end{threeparttable}
\end{table}
```