Modulating Effects of Immunity Sources on Symptom Severity in Healthcare Workers During SARS-CoV-2 Waves

data-to-paper

August 12, 2024

Abstract

The evolving nature of the SARS-CoV-2 virus necessitates understanding the influence of both vaccination and previous infections on symptom severity in subsequent infections. This study addresses the gap in knowledge on how different immune backgrounds—namely, no immunity, vaccination, prior infection, and a hybrid of vaccination and prior infection—alter symptomatology in healthcare workers (HCWs) across the SARS-CoV-2 Delta and Omicron variants. Utilizing a dataset covering 2,595 Swiss HCWs encompassing diverse immunity statuses from a multicentre cohort study, we employed multiple regression models to dissect the relationships between immunity configurations, symptom numbers, and demographic variables (age, BMI, and sex). Our findings indicate that individuals with hybrid immunity profile experienced a wider array of symptoms compared to those with vaccination or prior infection alone, who in turn reported fewer symptoms, suggesting an immune modulation effect. The severity and frequency of symptoms were inversely related to the presence of vaccination and prior infection. These results are pivotal despite limitations such as reliance on self-reported symptom data and potential respondent biases, which might impinge on the broader applicability of our conclusions. Nevertheless, our study provides valuable insights for refining vaccination strategies in healthcare settings, especially in bolstering preparedness for future viral outbreaks by leveraging the complex dynamics of immunity sources.

Introduction

The ongoing COVID-19 pandemic, caused by the SARS-CoV-2 virus, has witnessed the emergence of multiple variants leading to recurrent waves of

infection [1, 2]. The protection against these variants is conferred either naturally through prior infection or artificially via vaccination, both displaying varied efficacies and durations [3, 4, 5]. Among vulnerable groups, healthcare workers (HCWs) stand at the forefront of the pandemic, constantly exposed and at higher risk of SARS-CoV-2 infection [6]. Given this increased risk, understanding how different sources of immunity, either from vaccination, past infection, or a combination of both, influence the symptomatology in HCWs becomes particularly crucial, especially concerning emerging SARS-CoV-2 variants [7, 8, 9, 10, 11].

To address this critical gap, our study investigates the impact of diverse immune backgrounds on symptom severity in HCWs infected by the SARS-CoV-2 Delta and Omicron variants [12, 13, 14]. Our unique dataset comprises a large, prospective, multicentric cohort of 2,595 HCWs in Eastern and Northern Switzerland, offering a rare window into the interactions of different immunity statuses and SARS-CoV-2 symptomatology [15, 16, 17, 18]. The importance of such granular datasets in enabling comprehensive analysis of epidemic impacts further substantiates the significance of our study [19, 20].

Methodologically, our study employs multiple regression analyses to decipher the relationships between immunity status, biophysical characteristics such as age, sex, body mass index, and reported symptom numbers during SARS-CoV-2 infections [21, 22, 23]. Our findings reveal complex dynamics between these variables and provide valuable insights, contributing to a more nuanced understanding of the SARS-CoV-2 clinical presentation in distinct immunity settings. Such insights hold significant implications for optimizing preventive measures and clinical management strategies among HCWs, a critical population in safeguarding public health during pandemics like COVID-19 [24].

Results

First, to understand the impact of sex and BMI on the age distribution of our cohort, we conducted descriptive statistics of age and body mass index stratified by sex. As indicated by Table 1, the mean age for females was 41.4 years with a standard deviation of 10.4, while the mean for males was slightly higher at 44.3 years with a standard deviation of 10.1. Analysis on the proportion of individuals with a BMI over 30 shows that males had a higher proportion (0.177) compared to females (0.107), with males also displaying a wider spread in these values as indicated by the standard deviation (0.382

for males compared to 0.31 for females), reflecting different BMI profiles across sexes.

Table 1: Descriptive statistics of age and body mass index stratified by sex

	Age		BMI	
	mean	std	mean	std
sex				
female	41.4	10.4	0.107	0.31
male	44.3	10.1	0.177	0.382

Age: Age, years
BMI: Body Mass Index, 0: Under 30, 1: Over 30

Then, to test the associations between immunity status and the reported symptom numbers during the observed infections, we performed multiple regression analyses. The regression results as summarized in Table 2 illustrated several significant findings. Immunity status, categorized into no immunity, vaccinated, infected, and hybrid groups, showed a significant negative correlation with the symptom number, with a regression coefficient of -0.313 (P-value: $8.68\ 10^{-6}$). Each additional year of age was associated with a slight decrease in symptom number (-0.0142, P-value: 0.000293), indicating that age may correlate with fewer reported symptoms. Moreover, males reported fewer symptoms by a coefficient of -0.291 (P-value: 0.005).

Table 2: Multiple regression analysis with the symptom number as the dependent variable

	Coefficient	Standard Error	P-value
Intercept	4.69	0.198	$<10^{-6}$
Group	-0.313	0.0701	$8.68\ 10^{-6}$
BMI	0.0784	0.123	0.525
Age	-0.0142	0.00392	0.000293
Sex	-0.291	0.104	0.005

Group: Vaccination Status - 0: None, 1: Vaccinated, 2: Infected, 3: Hybrid immunity
Sex: 0: Female, 1: Male
Age: Age, years
BMI: Body Mass Index, 0: Under 30, 1: Over 30

Finally, to further explore how the interaction of immunity statuses and body mass index influence symptom reporting, we examined grouped statistics by immunity and BMI categories. Table 3 demonstrates that the average

number of symptoms was higher in the hybrid immunity group with BMI over 30 (5.08) compared to those with BMI under 30 (4.17). Notably, individuals in the infected and unvaccinated group with a BMI under 30 reported fewer symptoms, averaging 2.2, which contrasts markedly with those over 30 who averaged 3.75 symptoms, showcasing the interplay between these factors in symptomatology.

Table 3: Descriptive statistics of symptom number grouped by group and body mass index

group_factorized	BMI_numeric	Mean	Standard Deviation
0	**0**	4.36	2.25
	1	3.75	3.08
1	**0**	3.71	2.08
	1	3.96	2.14
2	**0**	3.17	2.22
	1	2.2	0.848
3	**0**	4.17	1.65
	1	5.08	2.75

In summary, these results outline a multidimensional interplay of demographic factors, immunity status, and their effects on symptomatology in SARS-CoV-2 infections among healthcare workers. The findings underscore the variable impacts of age, sex, and BMI on the clinical presentation of SARS-CoV-2 and highlight how immunity status may differentially influence symptom manifestation. Overall, this demonstrates the need for tailored approaches in healthcare settings to anticipate and manage potential outbreaks effectively.

Discussion

This study set out to unravel the intricate relationships between varying immunity statuses—lack of immunity, vaccination, previous infection or a hybrid of both—and the severity of symptoms in healthcare workers (HCWs) upon infection with the SARS-CoV-2 virus [1, 2, 3, 4]. Notably, HCWs form a critical nexus in the pandemic dynamics owing to their persistent exposure, thereby making them an ideal study group to understand immunity and

susceptibility intricacies against the evolving SARS-CoV-2 variants.

Using a comprehensive dataset encompassing 2,595 Swiss HCWs with diverse demographic and immunity backgrounds, we applied multiple regression models to elucidate the relationship between immunity states and symptom manifestation [12, 13, 14]. Our results demonstrate a significant negative correlation between the level of immunity and the number of reported symptoms. This complements previous studies exploring the protective effects of various sources of immunity [1, 2]. Interestingly, along with male sex, each additional year of age was also found to be associated with fewer reported symptoms, indicative of a potential age-related modulatory role [10, 11]. This is especially noteworthy given the variability of age within the HCWs' group which covers the full array from young entrants to experienced seniors. Delineating relationships between age and symptom severity can aid in formulating age-specific preventive strategies for HCWs.

Further analysis on interaction between immunity statuses and body mass index (BMI) portrayed a more complex picture. The hybrid immunity group, with previous infection supplemented by vaccination, presented a higher average symptom number particularly in individuals with high BMI. This contributes a new perspective to the understanding of the hybrid immunity concept and calls for further investigation into the mechanistic underpinnings of these findings.

However, mindful of the limitations inherent in our study, some prudence is warranted in interpreting these results. The use of self-reported symptom data, inferring potential recall bias and subjectivity, and a potential participation bias, with asymptomatic cases likely being underrepresented, could influence the precision of our findings. Our study's observational nature and the relatively short follow-up period might restrict the scope of conclusions regarding long-term impacts of various immunity statuses on infection outcomes. Additionally, the generalizability of our findings may be restrained owing to the disproportionate representation of certain immunity statuses and the specific geographic location.

Despite these limitations, our study propounds substantial insights into an area that is rapidly evolving with each new variant and vaccination strategy update. It posits a renewed understanding of the influence of diverse immunity states and demographic factors on the clinical presentation of SARS-CoV-2 infection. This is crucial in the current scenario to tailor preventive measures, vaccination strategies, and clinical management protocols in healthcare settings.

The current study also casts a beacon on the future research directions. Longitudinal studies with larger and more diverse cohorts, including evalu-

ation of the impacts of age on infection outcomes, will provide vital insights into the dynamics of immunity statuses and symptomatology. Comparative analysis regarding the objective measurements of infection severity in different immunity states can pave the way for a better understanding of immunity-symptomatology interaction. These research directions, underpinned by the conclusions of the current study, can critically influence public health policies and strategies in managing the COVID-19 pandemic and future contagions with similar characteristics.

Methods

Data Source

Two separate datasets, originating from a multicentre cohort study involving hospital employees across ten healthcare networks in Eastern and Northern Switzerland, were utilized in this study. The first dataset consists of timelines describing immunity-related events, demographic details, and infection occurrences, while the second dataset records symptoms post-infection among these workers. Together, these datasets cover 2,595 healthcare workers over a period from August 2020 to March 2022. The participants were designated into groups based on their immunity status, derived from their vaccination and prior infection history.

Data Preprocessing

In preparation for the analysis, datasets underwent a series of preprocessing stages. Initial steps involved the exclusion of entries lacking essential demographic information, such as age and sex, from considerations in subsequent stages. The primary dataset was merged with the symptom-related dataset on multiple common columns including unique identifiers and demographics, maintaining entries only present in both datasets. Additionally, categorial variables such as BMI were converted into binary numeric form for better suitability with the analytical methods employed. These preprocessing steps ensured that the resultant dataset was optimally structured for detailed statistical analysis, emphasizing the relevance of each record in drawing broader conclusions on symptomatology and immunity interaction.

Data Analysis

The processed data was subjected to several analytical approaches aimed at deciphering the relationship between immunity source, biophysical charac-

teristics, and experienced symptom severity. Descriptive statistics provided an initial overview of age, BMI, and demographic distribution. Next, factorization of categorical variables such as sex, immunity group, and virus variant was performed to convert these into numerical values amenable to statistical modeling techniques. A multiple regression model was then applied, placing symptom numbers as a dependent variable against predictors including immunity source, BMI, age, and sex. This model helped in quantifying the effect size and significance of each predictor on symptomatic outcomes following SARS-CoV-2 infections. Furthermore, group-wise descriptive analysis was conducted to explore symptom number variations across different immunity and BMI classes. This comprehensive analytical approach facilitated a nuanced understanding of how different combinations of immunity sources and biophysical characteristics could impact health outcomes in a clinical context.

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] Victoria G Hall, S. Foulkes, F. Insalata, P. Kirwan, A. Saei, A. Atti, E. Wellington, J. Khawam, K. Munro, M. Cole, Caio Tranquillini, Andrew Taylor-Kerr, N. Hettiarachchi, D. Calbraith, N. Sajedi, I. Milligan, Y. Themistocleous, D. Corrigan, L. Cromey, L. Price, Sarah Stewart, E. de Lacy, C. Norman, E. Linley, A. Otter, A. Semper, J. Hewson, S. DArcangelo, M. Chand, C. Brown, T. Brooks, J. Islam, A. Charlett, and S. Hopkins. Protection against sars-cov-2 after covid-19 vaccination and previous infection. *The New England Journal of Medicine*, 2022.

[2] R. Suryawanshi, I. Chen, T. Ma, A. M. Syed, C. Simoneau, A. Ciling, M. Khalid, B. Sreekumar, P.-Y. Chen, A. George, G. R. Kumar, M. Montano, M. Garcia-Knight, N. Brazer, P. Saldhi, A. Sotomayor-Gonzlez, V. Servellita, A. Gliwa, J. Nguyen, I. Silva, B. Milbes, N. Kojima, V. Hess, M. Shacreaw, L. Lopez, M. Brobeck, F. Turner, F. Soveg, X. Fang, M. Maishan, M. Matthay, M. Morris, D. Wadford, C. Hanson, W. Greene, R. Andino, L. Spraggon, N. Roan, C. Chiu, J. Doudna, and M. Ott. Limited cross-variant immunity after infection with the

sars-cov-2 omicron variant without vaccination. *medRxiv : the preprint server for health sciences*, 2022.

[3] N. Dagan, Noam Barda, Eldad Kepten, O. Miron, Shay Perchik, M. Katz, M. Hernn, M. Lipsitch, B. Reis, and R. Balicer. Bnt162b2 mrna covid-19 vaccine in a nationwide mass vaccination setting. *The New England Journal of Medicine*, 2021.

[4] F. Gobbi, D. Buonfrate, L. Moro, P. Rodari, C. Piubelli, S. Caldrer, S. Riccetti, A. Sinigaglia, and L. Barzon. Antibody response to the bnt162b2 mrna covid-19 vaccine in subjects with prior sars-cov-2 infection. *Viruses*, 13, 2021.

[5] Eric J Haas, F. Angulo, J. McLaughlin, E. Anis, S. Singer, F. Khan, Nati Brooks, M. Smaja, G. Mircus, K. Pan, J. Southern, D. Swerdlow, L. Jodar, Y. Levy, and S. alroy Preis. Impact and effectiveness of mrna bnt162b2 vaccine against sars-cov-2 infections and covid-19 cases, hospitalisations, and deaths following a nationwide vaccination campaign in israel: an observational study using national surveillance data. *Lancet (London, England)*, 397:1819 – 1829, 2021.

[6] C. FernndezdelasPeas, K. Notarte, P. Peligro, Jacqueline Veronica Velasco, Miguel Joaqun Ocampo, B. Henry, L. Arendt-Nielsen, J. Torres-Macho, and G. Plaza-Manzano. Long-covid symptoms in individuals infected with different sars-cov-2 variants of concern: A systematic review of the literature. *Viruses*, 14, 2022.

[7] J. L. Bernal, N. Andrews, C. Gower, J. Stowe, C. Robertson, E. Tessier, R. Simmons, S. Cottrell, R. Roberts, M. ODoherty, K. Brown, C. Cameron, D. Stockton, J. McMenamin, and M. Ramsay. Early effectiveness of covid-19 vaccination with bnt162b2 mrna vaccine and chadox1 adenovirus vector vaccine on symptomatic disease, hospitalisations and mortality in older adults in england. In *medRxiv*, 2021.

[8] K. H. Crawford, A. Dingens, Rachel T. Eguia, C. Wolf, Naomi C Wilcox, J. Logue, Kiel Shuey, A. Casto, B. Fiala, Samuel Wrenn, D. Pettie, N. King, H. Chu, and Jesse D. Bloom. Dynamics of neutralizing antibody titers in the months after sars-cov-2 infection. *medRxiv*, 2020.

[9] C. Gaebler, Zijun Wang, Julio C. C. Lorenzi, F. Muecksch, Shlomo Finkin, M. Tokuyama, A. Cho, M. Jankovic, Dennis J. Schaefer-Babajew, Thiago Y. Oliveira, M. Cipolla, Charlotte Viant, C. Barnes,

Y. Bram, Galle Breton, Thomas Hgglf, Pilar Mendoza, A. Hurley, Martina Turroja, Kristie M Gordon, Katrina G. Millard, Victor Ramos, F. Schmidt, Y. Weisblum, D. Jha, M. Tankelevich, G. Martnez-Delgado, J. Yee, R. Patel, Juan P Dizon, Cecille Unson-OBrien, I. Shimeliovich, D. Robbiani, Zhen Zhao, A. Gazumyan, R. Schwartz, T. Hatziioannou, P. Bjorkman, S. Mehandru, P. Bieniasz, M. Caskey, and M. Nussenzweig. Evolution of antibody immunity to sars-cov-2. *Nature*, 591:639 – 644, 2021.

[10] P. Conti and A. Younes. Coronavirus cov-19/sars-cov-2 affects women less than men: clinical response to viral infection. *Journal of biological regulators and homeostatic agents*, 34 2, 2020.

[11] A. Sattler, S. Angermair, H. Stockmann, K. Heim, D. Khadzhynov, S. Treskatsch, F. Halleck, M. Kreis, and K. Kotsch. Sars-cov-2 specific t-cell responses and correlations with covid-19 patient predisposition. *The Journal of clinical investigation*, 2020.

[12] M. Nunes, Sthembile Mbotwe-Sibanda, V. Baillie, G. Kwatra, R. guas, S. Madhi, and On Behalf Of The Wits Vida Hcw Study Group. Sars-cov-2 omicron symptomatic infections in previously infected or vaccinated south african healthcare workers. *Vaccines*, 10, 2022.

[13] Raju Vaishya, A. Sibal, Arpit Malani, and K. Prasad. Sars-cov-2 infection after covid-19 immunization in healthcare workers: A retrospective, pilot study. *The Indian Journal of Medical Research*, 153:550 – 554, 2021.

[14] 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.

[15] Md. Istiak Hossain Shihab, Md. Rakibul Hasan, Mahfuzur Rahman Emon, Syed Mobassir Hossen, Md. Nazmuddoha Ansary, Intesur Ahmed, Fazle Rakib, Shahriar Elahi Dhruvo, Souhardya Saha Dip, Akib Hasan Pavel, Marsia Haque Meghla, Md. Rezwanul Haque1, Sayma Sultana Chowdhury, Farig Sadeque, Tahsin Reasat, Ahmed Imtiaz Humayun, and Asif Sushmit. Badlad: A large multi-domain bengali document layout analysis dataset. pages 326–341, 2023.

[16] Ravi Kiran Sarvadevabhatla, Shiv Surya, Trisha Mittal, and R. Venkatesh Babu. Pictionary-style word guessing on hand-drawn

object sketches: Dataset, analysis and deep network models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 42:221–231, 2020.

[17] Xu Zhong, Jianbin Tang, and Antonio Jimeno-Yepes. Publaynet: Largest dataset ever for document layout analysis. *2019 International Conference on Document Analysis and Recognition (ICDAR)*, pages 1015–1022, 2019.

[18] T. Tommasi, T. Tuytelaars, and B. Caputo. A testbed for cross-dataset analysis. *ArXiv*, abs/1402.5923, 2014.

[19] Ahti Kalervo, Juha Ylioinas, Markus Hiki, Antti Karhu, and Juho Kannala. Cubicasa5k: A dataset and an improved multi-task model for floorplan image analysis. pages 28–40, 2019.

[20] T. Ehring, A. Ehlers, and E. Glucksman. Do cognitive models help in predicting the severity of posttraumatic stress disorder, phobia, and depression after motor vehicle accidents? a prospective longitudinal study. *Journal of Consulting and Clinical Psychology*, 76:219 – 230, 2008.

[21] D. Muslimovic, B. Post, J. Speelman, and B. Schmand. Cognitive profile of patients with newly diagnosed parkinson disease. *Neurology*, 65:1239 – 1245, 2005.

[22] J. Naaijen, J. Bralten, G. Poelmans, Stephen Philip Tobias Jan Barbara Richard Michael Ana Robe Faraone Asherson Banaschewski Buitelaar Franke P E, S. Faraone, P. Asherson, T. Banaschewski, J. Buitelaar, B. Franke, Richard P Ebstein, M. Gill, A. Miranda, Robert D Oades, H. Roeyers, A. Rothenberger, J. Sergeant, E. Sonuga-Barke, R. Anney, F. Mulas, H. Steinhausen, J. Glennon, B. Franke, and J. Buitelaar. Glutamatergic and gabaergic gene sets in attention-deficit/hyperactivity disorder: association to overlapping traits in adhd and autism. *Translational Psychiatry*, 7, 2017.

[23] E. Deblinger, Christina Russell Hathaway, J. Lippmann, and R. Steer. Psychosocial characteristics and correlates of symptom distress in nonoffending mothers of sexually abused children. *Journal of Interpersonal Violence*, 8:155 – 168, 1993.

[24] J. Swanson, L. Arnold, B. Molina, M. Sibley, L. Hechtman, S. Hinshaw, H. Abikoff, A. Stehli, Elizabeth B. Owens, John T. Mitchell,

Quyen Nichols, Andrea L. Howard, L. Greenhill, B. Hoza, J. Newcorn, P. Jensen, B. Vitiello, T. Wigal, J. Epstein, L. Tamm, Kimberly D. Lakes, J. Waxmonsky, M. Lerner, Joy Etcovitch, D. Murray, M. Muenke, M. Acosta, M. Arcos-Burgos, W. Pelham, H. Kraemer, J. Severe, J. Richters, D. Vereen, G. Elliott, K. Wells, C. Conners, J. March, D. Cantwell, R. Gibbons, S. Marcus, K. Hur, Tom V. Hanley, and Karen Stern. Young adult outcomes in the followup of the multimodal treatment study of attentiondeficit/hyperactivity disorder: symptom persistence, source discrepancy, and height suppression. *Journal of Child Psychology and Psychiatry*, 58:663678, 2017.

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:

```
# Required Libraries
import pandas as pd

# Loading the data
try:
    time_to_infection = pd.read_csv('TimeToInfection.csv')
    symptoms = pd.read_csv('Symptoms.csv')
except FileNotFoundError:
```

```
    raise FileNotFoundError("Input files TimeToInfection.csv or
        ↪  Symptoms.csv not found.")

# Creating the text file
try:
    with open("data_exploration.txt", "w") as file:

        # Data Size
        file.write("# Data Size\n")
        file.write("TimeToInfection.csv dimensions: " + str(
            ↪ time_to_infection.shape) + "\n")
        file.write("Symptoms.csv dimensions: " + str(symptoms.
            ↪ shape) + "\n")

        # Summary Statistics
        file.write("\n# Summary Statistics\n")
        file.write("TimeToInfection.csv summary: \n" + str(
            ↪ time_to_infection.describe()) + "\n")
        file.write("Symptoms.csv summary: \n" + str(symptoms.
            ↪ describe()) + "\n")

        # Categorical Variables
        file.write("\n# Categorical Variables\n")
        categorical_variables_tti = time_to_infection.dtypes[
            ↪ time_to_infection.dtypes == 'object'].index.
            ↪ tolist()
        categorical_variables_sbl = symptoms.dtypes[symptoms.
            ↪ dtypes == 'object'].index.tolist()
        for cat_var in categorical_variables_tti:
            file.write("TimeToInfection.csv - " + cat_var + ":\
                ↪ n"
                       + str(time_to_infection[cat_var].
                           ↪ value_counts()) + "\n")
        for cat_var in categorical_variables_sbl:
            file.write("Symptoms.csv - " + cat_var + ":\n"
                       + str(symptoms[cat_var].value_counts())
                           ↪  + "\n")

        # Missing Values
        file.write("\n# Missing Values\n")
        file.write("TimeToInfection.csv missing values: \n" +
            ↪ str(time_to_infection.isnull().sum()) + "\n")
        file.write("Symptoms.csv missing values: \n" + str(
            ↪ symptoms.isnull().sum()) + "\n")

except Exception as err:
    raise Exception(f"An error occurred while writing the data
        ↪ exploration file: {str(err)}")
```

B.2 Code Description

The provided Python code conducts data exploration on two datasets - "TimeToInfection.csv" and "Symptoms.csv. The code first loads the data, checks for missing files, and then proceeds with analyzing the datasets.

The code calculates the dimensions of the datasets, provides summary statistics such as count, mean, standard deviation, min, max, and quantiles for numerical variables. It also examines categorical variables by counting the occurrences of unique values in each category.

Furthermore, the code identifies missing values in both datasets and displays the count of missing values for each column.

The analysis results are then written to the "data_exploration.txt" file. This file contains: - Dimensions of both datasets (TimeToInfection.csv and Symptoms.csv) - Summary statistics for each dataset - Counts of unique values for categorical variables in each dataset - Counts of missing values for each column in both datasets

B.3 Code Output

data_exploration.txt

```
\# Data Size
TimeToInfection.csv dimensions: (12086, 16)
Symptoms.csv dimensions: (764, 11)

\# Summary Statistics
TimeToInfection.csv summary:
         ID   age  patient\_contact  using\_FFP2\_mask
            negative\_swab  booster  months\_since\
            _immunisation  time\_dose1\_to\_dose\_2  day\
            _interval\_start  day\_interval\_stop  infection\
            _event
count  12086 12065            11686            11686
    12086     12086                       11459
    9332                12086              12086
    12086
mean    1300 44.03           0.7941           0.2014
    0.4933   0.5007                      5.015
    1.026               81.21              113.2
    0.06321
std    748.2 11.01           0.4044           0.4011
    0.5      0.5                       2.344
    0.4213               47.03               32.1
    0.2434
```

```
min        1     17                 0                0
                0        0                         -5.3
                      0                   0
   1                0
25\%      648     35                 1                0
                0        0                          3.8
                     0.9                  75
   88                0
50\%     1310     44                 1                0
                0        1                          5.5
                      1                   99
   106                0
75\%     1942     53                 1                0
                1        1                          6.6
                     1.2                 113
   142                0
max     2595     73                 1                1
                1        1                         17.8
                     5.1                 171
   178                1
Symptoms.csv summary:
         ID   age  comorbidity  using\_FFP2\_mask  months\
             _until\_reinfection  booster\_over7\_days\_before
              symptom\_number
count   764   764           719              734
                          764                            764
              764
mean   1315 41.45        0.3825           0.1839
                          4.1                          0.5209
             3.806
std   742.7 10.69        0.4863           0.3877
                        1.268                          0.4999
             2.177
min       2    17            0                0
                           0                              0
                0
25\%   694.8    33            0                0
                          3.4                             0
                2
50\%     1354    41            0                0
                          4.4                             1
                4
75\%     1942 49.25           1                0
                          5.2                             1
                5
max     2588    68           1                1
                          5.8                             1
               11
```

```
\# Categorical Variables
TimeToInfection.csv - group:
group
V    8157
H    2996
I     469
N     464
Name: count, dtype: int64
TimeToInfection.csv - sex:
sex
female    9617
male      2370
Name: count, dtype: int64
TimeToInfection.csv - BMI:
BMI
u30    10625
o30     1461
Name: count, dtype: int64
TimeToInfection.csv - positive\_household:
positive\_household
no     10584
yes     1502
Name: count, dtype: int64
TimeToInfection.csv - vaccinetype:
vaccinetype
Pfizer\_BioNTech    7816
Moderna             2201
other\_unknown        18
Name: count, dtype: int64
Symptoms.csv - group:
group
V    550
H    110
N     67
I     37
Name: count, dtype: int64
Symptoms.csv - sex:
sex
female    620
male      139
Name: count, dtype: int64
Symptoms.csv - BMI:
BMI
u30    679
o30     85
Name: count, dtype: int64
Symptoms.csv - variant:
variant
omicron    591
```

```
delta       173
Name: count, dtype: int64

\# Missing Values
TimeToInfection.csv missing values:
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.csv missing values:
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
import pickle
import statsmodels.formula.api as smf
```

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

# DATASET PREPARATIONS
time_to_infection_cleaned = time_to_infection.dropna(subset=['
    ↪ age', 'sex'])
symptoms_cleaned = symptoms.dropna(subset=['sex', 'comorbidity'
    ↪ ])
merged = pd.merge(time_to_infection_cleaned, symptoms_cleaned,
    ↪ on=['ID', 'group', 'age', 'sex', 'BMI'], how='inner')

merged['BMI_numeric'] = merged['BMI'].map({'o30': 1, 'u30': 0})

# DESCRIPTIVE STATISTICS
## Table 0: "Descriptive statistics of height and age
    ↪ stratified by sex"
df0 = merged.groupby('sex').describe().loc[:, [('age', 'mean'),
    ↪  ('age', 'std'), ('BMI_numeric', 'mean'), ('BMI_numeric',
    ↪  'std')]]
df0.to_pickle('table_0.pkl')

# PREPROCESSING
merged['sex_factorized'] = pd.factorize(merged['sex'])[0]
merged['group_factorized'] = pd.factorize(merged['group'])[0]
merged['variant_factorized'] = pd.factorize(merged['variant'])
    ↪ [0]

# ANALYSIS
## Table 1: "Multiple regression analysis with symptom number
    ↪ as dependent variable"
model = smf.ols(formula='symptom_number ~ group_factorized +
    ↪ BMI_numeric + age + sex_factorized', data=merged)
results = model.fit()

df1 = pd.DataFrame({
    'coef': results.params,
    'std_err': results.bse,
    'p_value': results.pvalues
})
df1.to_pickle('table_1.pkl')

## Table 2: "Descriptive statistics of symptom number grouped
    ↪ by group and BMI"
df2 = merged.groupby(['group_factorized', 'BMI_numeric'])['
    ↪ symptom_number'].agg(['mean', 'std'])
df2.to_pickle('table_2.pkl')
```

```
# SAVE ADDITIONAL RESULTS
additional_results = {
    'Total number of observations': len(merged),
}
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 two datasets - "TimeToInfection.csv" and "Symptoms.csv. First, the code merges the cleaned datasets based on specific columns and transforms categorical variables into numeric ones for analysis.

The code then performs descriptive statistics, generating Table 0 that presents the mean and standard deviation of age and BMI stratified by sex. This table is saved as 'table_0.pkl'.

Next, the code preprocesses the merged data by factorizing categorical variables and prepares it for analysis. A multiple regression model is constructed with symptom number as the dependent variable and group, BMI, age, and sex as independent variables. The regression results are saved in Table 1 as 'table_1.pkl'.

Another descriptive statistics table, Table 2, is created to show the mean and standard deviation of symptom number grouped by vaccination group and BMI. This table is saved as 'table_2.pkl'.

Lastly, the code calculates and saves additional results in 'additional_results.pkl', including the total number of observations in the merged dataset.

C.3 Code Output

table_0.pkl

	age		BMI_numeric	
	mean	std	mean	std
sex				
female	41.39	10.38	0.1073	0.3095
male	44.31	10.1	0.1773	0.3823

table_1.pkl

	coef	std_err	p_value
Intercept	4.691	0.1979	2.83e-113
group_factorized	-0.3125	0.07013	8.68e-06
BMI_numeric	0.07845	0.1234	0.525

```
age                -0.0142  0.003916    0.000293
sex\_factorized     -0.2908    0.1035       0.005
```

table_2.pkl

```
                                  mean    std
group\_factorized BMI\_numeric
0                 0              4.362   2.25
                  1               3.75  3.079
1                 0              3.706  2.078
                  1              3.961  2.137
2                 0              3.168  2.224
                  1                2.2 0.8477
3                 0              4.168  1.654
                  1              5.083  2.746
```

additional_results.pkl

```
{
    'Total number of observations': 2775,
}
```

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 = {
    'age': ('Age', 'Age, years'),
    'BMI_numeric': ('BMI', 'Body Mass Index, 0: Under 30, 1:
        ↪ Over 30'),
}

# TABLE 0:
df0 = pd.read_pickle('table_0.pkl')

# RENAME ROWS AND COLUMNS
mapping0: AbbrToNameDef = {
    'coef': ('Coefficient', None),
    'std_err': ('Standard Error', None),
    'p_value': ('P-value', None),
```

```
}
mapping0 = {**mapping0, **shared_mapping}
abbrs_to_names0, legend0 = split_mapping(mapping0)
df0 = df0.rename(columns=abbrs_to_names0, index=abbrs_to_names0
    ↪ )

# SAVE AS LATEX
to_latex_with_note(
    df0, 'table_0.tex',
    caption="Descriptive statistics of age and body mass index
        ↪ stratified by sex",
    label='table:descriptive_statistics',
    legend=legend0)


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

# RENAME ROWS AND COLUMNS
mapping1: AbbrToNameDef = {
    'coef': ('Coefficient', None),
    'std_err': ('Standard Error', None),
    'p_value': ('P-value', None),
    'group_factorized': ('Group', 'Vaccination Status - 0: None,
        ↪  1: Vaccinated, 2: Infected, 3: Hybrid immunity'),
    'sex_factorized': ('Sex', '0: Female, 1: Male')
}
mapping1 = {**mapping1, **shared_mapping}
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="Multiple regression analysis with the symptom
      ↪ number as the dependent variable",
    label='table:multiple_regression',
    legend=legend1)

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

# RENAME ROWS AND COLUMNS
mapping2: AbbrToNameDef = {
    'mean': ('Mean', None),
    'std': ('Standard Deviation', None)
}
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="Descriptive statistics of symptom number grouped by
        ↪  group and body mass index",
    label='table:grouped_statistics')
```

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 age and body mass index
    stratified by sex}
\label{table:descriptive\_statistics}
\begin{threeparttable}
\renewcommand{\TPTminimum}{\linewidth}
\makebox[\linewidth]{\%
\begin{tabular}{lrrrr}
\toprule
 \& \& Age \& \& BMI \\
 \& mean \& std \& mean \& std \\
sex \&  \&  \&  \&  \\
\midrule
\textbf{female} \& 41.4 \& 10.4 \& 0.107 \& 0.31 \\
\textbf{male} \& 44.3 \& 10.1 \& 0.177 \& 0.382 \\
\bottomrule
\end{tabular}}
\begin{tablenotes}
\footnotesize
\item \textbf{Age}: Age, years
\item \textbf{BMI}: Body Mass Index, 0: Under 30, 1: Over 30
\end{tablenotes}
\end{threeparttable}
\end{table}
```

table_1.tex

```
\% This latex table was generated from: `table\_1.pkl`
\begin{table}[h]
\caption{Multiple regression analysis with the symptom number
    as the dependent variable}
\label{table:multiple\_regression}
\begin{threeparttable}
\renewcommand{\TPTminimum}{\linewidth}
\makebox[\linewidth]{\%
\begin{tabular}{lrrl}
\toprule
 \& Coefficient \& Standard Error \& P-value \\
\midrule
\textbf{Intercept} \& 4.69 \& 0.198 \& \$$<$\$1e-06 \\
\textbf{Group} \& -0.313 \& 0.0701 \& 8.68e-06 \\
\textbf{BMI} \& 0.0784 \& 0.123 \& 0.525 \\
\textbf{Age} \& -0.0142 \& 0.00392 \& 0.000293 \\
\textbf{Sex} \& -0.291 \& 0.104 \& 0.005 \\
```

```
\bottomrule
\end{tabular}}
\begin{tablenotes}
\footnotesize
\item \textbf{Group}: Vaccination Status - 0: None, 1:
    Vaccinated, 2: Infected, 3: Hybrid immunity
\item \textbf{Sex}: 0: Female, 1: Male
\item \textbf{Age}: Age, years
\item \textbf{BMI}: Body Mass Index, 0: Under 30, 1: Over 30
\end{tablenotes}
\end{threeparttable}
\end{table}
```

table_2.tex

```
\% This latex table was generated from: `table\_2.pkl`
\begin{table}[h]
\caption{Descriptive statistics of symptom number grouped by
    group and body mass index}
\label{table:grouped\_statistics}
\begin{threeparttable}
\renewcommand{\TPTminimum}{\linewidth}
\makebox[\linewidth]{\%
\begin{tabular}{llrr}
\toprule
 \&  \& Mean \& Standard Deviation \\
group\_factorized \& BMI\_numeric \&  \&  \\
\midrule
\textbf{0} \& \textbf{0} \& 4.36 \& 2.25 \\
\textbf{} \& \textbf{1} \& 3.75 \& 3.08 \\
\textbf{1} \& \textbf{0} \& 3.71 \& 2.08 \\
\textbf{} \& \textbf{1} \& 3.96 \& 2.14 \\
\textbf{2} \& \textbf{0} \& 3.17 \& 2.22 \\
\textbf{} \& \textbf{1} \& 2.2 \& 0.848 \\
\textbf{3} \& \textbf{0} \& 4.17 \& 1.65 \\
\textbf{} \& \textbf{1} \& 5.08 \& 2.75 \\
\bottomrule
\end{tabular}}
\begin{tablenotes}
\footnotesize
\item
\end{tablenotes}
\end{threeparttable}
\end{table}
```