scientific reports

Check for updates

OPEN

Clustering and time series analyses of hybrid immunity to SARS-COV-2 using data from the BQC19 biobank

Jean-Frédéric Boulianne[1,2], Denis Larocque[1], Simon Rousseau[3,4] & Delphine Bosson-Rieutort[2,5]✉

The SARS-CoV-2 pandemic revealed that immunity after infection was temporary, with reinfections occurring. As the pandemic progressed, individuals encountered infection and vaccination in varying sequences and at different time intervals, resulting in heterogeneous patterns of infection, reinfection and vaccination, so-called hybrid immunity. This study analyzed these patterns by grouping individuals based on their infection, reinfection, and vaccination sequences using data from the Biobanque québécoise de la COVID-19 (BQC19). We applied agglomerative and divisive hierarchical clustering on time series representing patients' COVID-19 episodes, using Dynamic Time Warping to compute distances. Their characterization revealed that clusters followed a temporal progression depending on the timing of infection and its positioning across the pandemic waves. On the other hand, reinfections occurred from the fifth wave onward. The most highly vaccinated groups appear to have been infected and consequently reinfected later in the pandemic. Some groups featured a higher proportion of healthcare workers, while for others, the trajectory and their timeframes were decisive. This study highlights the role of vaccination, which is in line with current knowledge. It also shows that, beyond the sequence of events, it is rather their temporality and the delays between them that are of greatest importance. In terms of hybrid immunity, the results of this study suggest that an infection between two vaccines could offer greater immunity.

Keywords SARS-CoV-2, COVID-19, Immunity, Cluster analysis, BQC19, Time series

Abbreviations

BMI	Body mass index
BQC19	Biobanque québécoise de la COVID-19
SARS-CoV-2	Severe acute respiratory syndrome coronavirus 2
PCR	Polymerase chain reaction
SD	Standard deviation
WHO	World Health Organization

Background

In December 2019, a novel coronavirus (SARS-CoV-2) was identified in Wuhan, China. A few weeks later, in January 2020, the World Health Organization (WHO) confirmed interhuman transmission. Rapidly, the virus spread across the planet, leading to the COVID-19 pandemic. By April 2024, this pandemic had resulted in more than 7 million deaths and 770 million infections worldwide[1,2], making it the most significant pandemic since the 1918 Spanish flu[3].

[1]HEC Montréal, Montréal, Québec, Canada. [2]Centre de Recherche en Santé Publique (CReSP), Montréal, Québec, Canada. [3]The Meakins-Christie Laboratories, Research Institute of the McGill University Health Centre (RI-MUHC), Montreal, QC, Canada. [4]Department of Medicine, McGill University, Montréal, Québec, Canada. [5]École de Santé Publique de L'Université de Montréal, Montréal, Québec, Canada. ✉email: delphine.bosson-rieutort@umontreal.ca

Coronaviruses can infect different animals and cause moderate to severe respiratory infections in humans. Although most symptoms resembled those of a typical respiratory disease, including fever, fatigue and cough[4,5], a significant proportion of infections progressed to a more severe, even critical, form of the disease, potentially involving dyspnea, acute respiratory distress syndrome and multiple organ failure. In addition to the acute form of the disease, the infection could lead to persistent health problems (e.g., fatigue, shortness of breath, and cognitive problems) known as post-COVID-19 syndrome, also known as long COVID[3,6].

The pandemic revealed that individuals who contracted the disease were not permanently immunized and could be reinfected after a relatively short period of 7 to 12 months[7,8]. This variability in reinfection time can be attributed to individual sensitivities, infection severity, and the evolution of the SARS-CoV-2 virus through various variants. These various mutations generate different immune responses that impact the immunity of individuals who have contracted the disease.

In this context, we aimed to study hybrid immunity, the immunity conferred by a combination of infection, reinfection and vaccination, by identifying and characterizing SARS-CoV-2 reinfection profiles. To specifically tackle this temporal and complex aspect of the hybrid immunity, we proposed to use machine learning techniques on data from *Biobanque québécoise de la COVID-19* (BQC19) to group individuals according to their temporal pattern of vaccination, infection, and reinfection only, and then, characterize the groups obtained in terms of sociodemographic and clinical factors to highlight any hidden patterns or characteristics that could lead to similar temporal sequences across our population.

Methods

BQC19 is a multicenter database involving a network of 11 Quebec hospitals with five partner academic institutions and has been described elsewhere[9]. In brief, this panprovincial initiative collects, stores and shares data and blood samples from COVID-19 patients, both severe and non-severe cases. The biobank contains several datasets about participants' characteristics, events and certain biological data (e.g., RNA, DNA, serum, plasma and peripheral blood mononuclear cells). It also includes longitudinal follow-up for 24 months following hospitalization (inpatient) or PCR testing (outpatient). This study was approved by Centre universitaire de santé McGill's ethics committee within the framework of the project *Determining the impact of hybrid immunity on the evolving landscape of host responses to SARS-CoV-2 in the Biobanque Québécoise de la COVID-19 (BQC19)* and all methods were performed in accordance with the relevant guidelines and regulations. Each BQC19 enrolling site has established a consent process that reflects the BQC19's standard operating procedures and all the participants provided informed consent before the start of the study. Full details are available in a previously published article[9]. The source population consisted of 6,272 participants included between March 2020 and August 2023. To be included in the study cohort, participants had to 1) have reached 18 years of age, 2) have a documented primary infection with a date, and 3) have a documented secondary infection (reinfection) with a date. For each individual, the longitudinal data available for the study ranged between two months and three years. As each participant's follow-up period may vary during the study, the number of events differed from one individual to another.

Data management was performed through an iterative process of data exploration and processing of the highlighted elements. Data exploration included frequency and mode for categorical variables, measures of central tendency (mean and median) and measures of dispersion (minimum, maximum, standard deviation (SD) and variance) for numerical variables. Stratified analyses were also performed (e.g., sex, occupation, etc.) with parametric (t-test or ANOVA) or nonparametric (Wilcoxon or Kruskal–Wallis) statistical test as required. Post-hoc tests (Tukey's range test or Dunn's test), when appropriate, were also conducted. Data management iterations also included standardization of values domain and cross-validation between variables to enable consistency validations between data correlated or linked by context. The various BQC19 data collections were merged to reduce missing values in the main dataset, mostly for event details. Process mining analysis was applied to support visualization of event sequences and flows[10].

To achieve the aim of grouping individuals according to their specific event patterns, variables of interest about infection, reinfection and vaccination were used for the clustering (temporal variables). Firstly, as the date of primary infection did not exist in the initial dataset, it was reconstructed from the date of each participant's first positive PCR test. Secondly, the reinfection variable had to be reconstructed since the dataset contained two variables with conflicting information (missing or different data). The variable was redefined as follows: 1) the value of the two variables when they were identical, 2) the nonmissing value when one of the variables was null, and 3) the earliest of the dates when the two variables did not contain the same value. In some cases (N = 96), more than one reinfection event was documented, and only the first documented reinfection for each individual was systematically retained. When subsequent observations were not explicitly identified as longitudinal reinfection follow-ups, they were compared with the other available dates and were not considered as a new reinfection if the date of a subsequent observation was within 14 days of the previous reinfection. This period corresponds to the internal delay set by BQC19 to consider a reinfection. Finally, vaccination dates were directly extracted from the dataset.

Participants were grouped using an agglomerative hierarchical clustering analysis based on the dissimilarity between their temporal trajectories. Specifically, we first generated a dissimilarity matrix using Dynamic Time Warping (DTW). To do so, we applied DTW on multidimensional time series constructed using the temporal variables of interest (infection, vaccinations and reinfections). Each patient was therefore represented by a matrix containing a time series for each event of interest. These individual matrices served as inputs to the DTW algorithm. A visual representation of the input data and the analytical sequence can be found as Supplementary Figure S1 online. Dynamic time warping was chosen as it was the most suitable method for the type of data in this study, especially as it takes into account temporal deformations in order to align time sequences[11].

Hierarchical clustering was then performed using Ward's minimum variance method on the DTW-based dissimilarity matrix obtained to group participants by minimizing within-cluster variance in their time-aligned trajectories. The number of groups was determined following the average silhouette statistic[12]. Multiple numbers of groups were compared, and the optimal number of groups was selected based on the highest average silhouette statistic, while ensuring that each group contained enough individuals to allow meaningful interpretation. For each group, a process map has been generated to support the visualization of their specific temporal sequence of events (infections, vaccinations and reinfections)[13].

To describe the different groups obtained from the cluster analysis based on the temporal sequence and highlight potential characteristics leading to similar sequences, we used variables related to sociodemographic characteristics (e.g., age, sex at birth, BMI), participants' condition, habits and environment (e.g., smoking and drug use status, occupation, household information) and the context of the contagion (e.g., number of reinfections, number of vaccines received when infection or reinfection occurs, wave of the pandemic, predominant variant). However, as variant sequencing data were only available for a small proportion of the cohort (n = 20), the actual variant was consequently not included in the variables of interest. We instead used waves defined according to an internal BQC19 timeline to ensure comparability with other related work. This timeline was visually compared to publicly available data (e.g., Our World In Data[14,15]), and no substantial differences were observed (see Supplementary Figure S2 online). Finally, delays between each pair of events were calculated and used to describe each cluster. All statistical analyses previously mentioned were used to describe each variable stratified by cluster.

Data cleaning and analyses were performed using R, version 4.3.0[16] with *Rstudio,* version 2023.12.1.402[17] supported by package *tidyverse*, version 2.0.0[18]. Hierarchical clustering and dynamic time warping were performed using packages *dtw*, version 1.23–1[13], *dwtclust*, version 5.5.12[19] and *proxy,* version 0.4–27[20]. Process mining for events visualization was performed using package *bupaR*, version 0.5.4[10].

Results

To achieve the objective of grouping individuals according to their pattern of infection, reinfection and vaccination sequences, 318 participants were included in the study based on the previously defined inclusion criteria (see Supplementary Figure S3 online). Table 1 reports characteristics of individuals in the study cohort and Table 2 shows characteristics of their COVID episodes. Among them, 230 were women (72.3%), and 141 were healthcare workers (44.3%). The average age was 43 years (SD 13.8), and a total of 31 participants were reinfected twice (9.3%), including 6 who were reinfected three times (1.9%). The average dose of vaccine at primary infection was 1.08 (SD 1.30), and the average doses of vaccine at reinfection were 2.36 (SD 0.876), 2.29 (SD 0.864), and 2.50 (SD 0.548).

Figure 1 presents the global sequence of events identified for the cohort. Boxes represent events while the edges represent the temporal sequence between the events. All boxes and edges are completed with the relative frequency of each event or transition, as well as the median time between each transition. We used the sum of the medians of the various transitions as an approximation of sequence duration for illustrative purposes. While the medians are not addable due to their statistical properties and do not represent the actual median of the complete sequence, they can support the identification of overall trends. To avoid confusion, this approximation method will hereafter be referred to as *summed medians*. Mapping revealed that 56.3% of individuals in the cohort began their sequence with the primary infection, while 43.7% started with a first dose of vaccine. Among all those who received the first dose, regardless of the previous event, 83.3% were subsequently vaccinated a second time, within a median of 87 days. Among those who received a second dose, regardless of the previous trajectory, 50.3% then received a third vaccine within a median of 191 days, while 16% contracted their first infection within a median of 183 days.

Sociodemographic, participant's condition, habits and environment characteristics

To group participants based on their infection (I), vaccination (V), and reinfection (R) patterns, we used event sequences as time series in a cluster analysis. Among the different numbers of clusters tested (2 to 6), both the five- and six-cluster solutions yielded the highest average silhouette coefficient (0.65), compared to 0.53, 0.58, and 0.62 for the two-, three-, and four-clusters solutions, respectively. However, the five-cluster solution was retained as it provided a more balanced partition, avoiding the creation of clusters with very small sample sizes. The first cluster included 138 participants (43.4%), while others respectively included 42 (13.2%), 11 (3.5%), 51 (16%) and 76 individuals (23.9%). Although small in size, cluster 3 was retained in the analysis due to its distinct temporal pattern. There was no significant difference between clusters in terms of age ($p = 0.137$) or body mass index (BMI) ($p = 0.545$), and the proportion of males and females in each cluster was similar to the cohort proportion. However, in the first three clusters, the ratio of health workers was lower (between 21.4% and 36.4%) than the cohort proportion (44.3%), whereas in the remaining clusters, the ratio was reversed, with 51% and 76.3%. The detailed characteristics of each group are shown in Tables 1 and 2.

Temporal description

The following section describes each group in terms of temporal sequence and presents illustrations of these sequences (Fig. 2A to Fig. 2E).

Cluster 1. Participants in the first cluster were mostly infected in the first three waves of the pandemic (Table 2). As shown in Fig. 2A, which presents the event mapping for this group, individuals were first infected before receiving two doses of vaccine within a *summed medians* of 225.5 days (median I-V1 139.5; median V1-V2 86). The sequence then split, with 44.2% of the group who were reinfected, while the remaining received their third dose of vaccine, both events within similar median timescales (185 and 181 days, respectively). Patients who

Characteristics	Cohort (n = 318)	Clusters				
		1 (n = 138)	2 (n = 42)	3 (n = 11)	4 (n = 51)	5 n = 76)
Age years: mean (SD)	43.0 (13.8)	43.2 (14.0)	40.5 (12.3)	41.8 (16.5)	40.2 (11.8)	46.3 (14.6)
18 to 34 *n*(%)	94 (29.6%)	42 (30,4%)	16 (38.1%)	4 (36.4%)	16 (31.4%)	16 (21.1%)
35 to 44 *n*(%)	84 (26.4%)	38 (27.5%)	7 (16.7%)	1 (9.1%)	20 (39.2%)	18 (23.7%)
45 à 64 *n*(%)	125 (39.3%)	51 (37.0%)	19 (45.2%)	5 (45.5%)	14 (27.5%)	36 (47.4%)
+ 65 years *n*(%)	15 (4.7%)	7 (5.1%)	0 (0%)	1 (9.1%)	1 (2.0%)	6 (7.9%)
BMI: mean (SD)	26.9 (5.3)	28.9 (8.6)	26.5 (5.3)	27.0 (5.9)	28.0 (5.1)	26.9 (5.3)
Sex at birth						
Female *n*(%)	230 (72.3%)	97 (70.3%)	26 (61.9%)	7 (63.6%)	41 (80.4%)	59 (77.6%)
Male *n*(%)	88 (27.7%)	41 (29.7%)	16 (38.1%)	4 (36,4%)	10 (19.6%)	17 (22.4%)
Smoking status						
Nonsmoker *n*(%)	230 (72.3%)	101 (73.2%)	26 (61.9%)	5 (45.5%)	40 (78.4%)	58 (76.3%)
Smoker *n*(%)	22 (6.9%)	9 (6.5%)	8 (19.0%)	0 (0%)	2 (3.9%)	3 (3.9%)
Former smoker *n*(%)	55 (17.3%)	20 (14.5%)	6 (14.3%)	5 (45.5%)	9 (17.6%)	15 (19.7%)
Passive smoker *n*(%)	1 (0.3%)	1 (0.7%)	- (0%)	- (0%)	- (0%)	- (0%)
Missing *n*(%)	10 (3.1%)	7 (5.1%)	2 (4.8%)	1 (9.1%)	- (0%)	- (0%)
Use electronic cigarettes						
Yes *n*(%)	13 (4.1%)	3 (2.2%)	4 (9.5%)	1 (9.1%)	3 (5.9%)	2 (2.6%)
No *n*(%)	299 (94.0%)	131 (94.9%)	37 (88.1%)	9 (81.8%)	48 (94.1%)	74 (97.4%)
Missing *n*(%)	6 (1.9%)	4 (2.9%)	1 (2.4%)	1 (9.1%)	- (0%)	- (0%)
Use Drugs						
Yes *n*(%)	30 (9.4%)	10 (7.2%)	10 (23.8%)	- (0%)	5 (9.8%)	5 (6.6%)
No *n*(%)	282 (88.7%)	124 (89.9%)	31 (73.8%)	10 (90.9%)	46 (90.2%)	71 (93.4%)
Missing *n*(%)	6 (1.9%)	4 (2.9%)	1 (2.4%)	1 (9.1%)	- (0%)	- (0%)
Healthcare worker						
Yes *n*(%)	141 (44.3%)	44 (31.9%)	9 (21.4%)	4 (36.4%)	26 (51.0%)	58 (76.3%)
No *n*(%)	171 (53.8%)	91 (65.9%)	31 (73.8%)	7 (63.6%)	24 (47.1%)	18 (23.7%)
Missing *n*(%)	6 (1.9%)	3 (2.2%)	2 (4.8%)	- (0%)	1 (2.0%)	- (0%)
Live where						
Home	315 (99.1%)	136 (98.6%)	41 (97.6%)	11 (100%)	51 (100%)	76 (100%)
Residence for eldery (RPA)	2 (0.6%)	2 (1.4%)	- (0%)	- (0%)	- (0%)	- (0%)
Nursing home (CHSLD)	- (0%)	- (0%)	- (0%)	- (0%)	- (0%)	- (0%)
Intermediate and family-type resources (RI-RTF)	- (0%)	- (0%)	- (0%)	- (0%)	- (0%)	- (0%)
In rooming house	- (0%)	- (0%)	- (0%)	- (0%)	- (0%)	- (0%)
Missing n(%)	1 (0.3%)	- (0%)	1 (2.4%)	- (0%)	- (0%)	- (0%)
Live with						
Family member(s)	279 (87.7%)	123 (89.1%)	38 (90.5%)	11 (100%)	45 (88.2%)	62 (81.6%)
Caretaker	- (0%)	- (0%)	- (0%)	- (0%)	- (0%)	- (0%)
Alone	31 (9.7%)	1- (7.2%)	4 (9.5%)	- (0%)	5 (9.8%)	12 (15.8%)
Roommate(s)	4 (1.3%)	2 (1.4%)	- (0%)	- (0%)	1 (2.0%)	1 (1.3%)
Inconnu n(%)	4 (1.3%)	3 (2.2%)	- (0%)	- (0%)	- (0%)	1 (1.3%)

Table 1. Sociodemographic characteristics of the study population between 2020/03 and 2023/08 and clusters.

received this last dose of vaccine took a median 118 days before being reinfected. Reinfection occurred in waves five, six and seven of the pandemic.

Cluster 2. Similarly to cluster 1, most individuals in this group were infected in the first waves (66.7%), although a few individuals contracted their primary infection later in the pandemic (Table 2). Almost the entire group (97.6%) was infected before a first dose of vaccine (Fig. 2B). The sequence of the entire group converged toward a reinfection, either following first dose (47.6%, *summed medians* 380.5 days (172; 208.8) from primary infection), or following the initial infection (52.4%, median 270 days). Some individuals had contracted the disease twice without vaccine in their sequences, which means that this group has the particularity of containing patients with natural immunity instead of hybrid immunity (N = 21). Reinfection occurred in waves five, six and seven but mainly in waves five and six (61.9%).

Cluster 3. In this cluster, individuals were mainly infected in the second and third waves of the pandemic (81.9%), with none having been infected in the first wave (Table 2). Thus, on average, they were infected slightly later than the previous two groups. This is interesting given that, as Fig. 2C shows, the entire group started their event sequence with the first vaccine. However, the delay between this vaccine and primary infection is

Characteristics	Cohort (n = 318)	Clusters				
		1 (n = 138)	2 (n = 42)	3 (n = 11)	4 (n = 51)	5 n = 76)
COVID severity						
Mild	299 (94.0%)	123 (89.1%)	4- (95.2%)	11 (100%)	51 (100%)	74 (97.4%)
Moderate	11 (3.5%)	11 (8.0%)	- (0%)	- (0%)	- (0%)	- (0%)
Severe	8 (2.5%)	4 (2.9%)	2 (4.8%)	- (0%)	- (0%)	2 (2.6%)
Death	- (0%)	- (0%)	- (0%)	- (0%)	- (0%)	- (0%)
Infection's wave						
1	22 (6.9%)	16 (11.6%)	6 (14.3%)	- (0%)	- (0%)	- (0%)
2	13- (40.9%)	11- (79.7%)	15 (35.7%)	5 (45.5%)	- (0%)	- (0%)
3	22 (6.9%)	11 (8.0%)	7 (16.7%)	4 (36.4%)	- (0%)	- (0%)
4	26 (8.2%)	1 (0.7%)	7 (16.7%)	1 (9.1%)	16 (31.4%)	1 (1.3%)
5	56 (17.6%)	- (0%)	3 (7.1%)	1 (9.1%)	3- (58.8%)	22 (28.9%)
6	16 (5.0%)	- (0%)	1 (2.4%)	- (0%)	2 (3.9%)	13 (17.1%)
7	46 (14.5%)	- (0%)	3 (7.1%)	- (0%)	3 (5.9%)	4- (52.6%)
Reinfection's wave						
1	- (0%)	- (0%)	- (0%)	- (0%)	- (0%)	- (0%)
2	- (0%)	- (0%)	- (0%)	- (0%)	- (0%)	- (0%)
3	- (0%)	- (0%)	- (0%)	- (0%)	- (0%)	- (0%)
4	1 (0.3%)	1 (0.7%)	- (0%)	- (0%)	- (0%)	- (0%)
5	72 (22.6%)	45 (32.6%)	19 (45.2%)	5 (45.5%)	3 (5.9%)	- (0%)
6	43 (13.5%)	32 (23.2%)	7 (16.7%)	- (0%)	4 (7.8%)	- (0%)
7	202 (63.5%)	6- (43.5%)	16 (38.1%)	6 (54.5%)	44 (86.3%)	76 (100%)
Delay between events (days): mean (SD)						
Vaccine1-Vaccine2	103 (63.6)	110 (70.5)	387 (43.8)	129 (84.5)	78.5 (24.3)	90.3 (23.9)
Vaccine1-Vaccine3	298 (67.8)	290 (76.4)	–	331 (66.6)	331 (77.4)	293 (49.0)
Vaccine1-Infection	70.7 (267)	-151 (91.9)	-173 (118)	49.5 (70.8)	255 (81.0)	416 (83.5)
Vaccine1-Reinfection	463 (172)	379 (114)	247 (141)	405 (108)	518 (123)	648 (112)
Vaccine1-Reinfection2	523 (151)	494 (90.9)	413 (114)	–	–	747 (177)
Vaccine1-Reinfection3	583 (113)	553 (133)	–	–	–	643 (21.9)
Vaccine2-Vaccine3	202 (60.1)	189 (67.5)	–	199 (65.1)	260 (71.6)	202 (32.0)
Vaccine2-Infection	-18.6 (285)	-261 (103)	-666 (163)	-79.1 (34.2)	177 (72.5)	326 (72.3)
Vaccine2-Reinfection	373 (182)	268 (121)	-213 (125)	277 (124)	439 (121)	557 (102)
Vaccine2-Reinfection2	407 (190)	371 (107)	-100 (–)	–	–	656 (150)
Vaccine2-Reinfection3	467 (101)	410 (62.7)	–	–	–	581 (18.4)
Vaccine3-Infection	-169 (269)	-427 (103)	–	-278 (74.4)	-96.0 (45.8)	125 (62.7)
Vaccine3-Reinfection	213 (166)	111 (134)	–	88.7 (129)	179 (129)	357 (93.6)
Vaccine3-Reinfection2	245 (169)	171 (98.0)	–	–	–	486 (111)
Vaccine3-Reinfection3	315 (92.1)	275 (84.9)	–	–	–	396 (NA)
Infection-Reinfection	391 (177)	529 (118)	392 (183)	356(126)	263 (103)	232 (84.2)
Infection-Reinfection2	564 (154)	614 (107)	593 (178)	–	–	330 (48.0)
Infection-Reinfection3	583 (170)	686 (75.2)	–	–	–	376 (7.07)
Reinfection-Reinfection2	175 (86.7)	162 (88.5)	250 (59.3)	–	–	139 (61.4)
Reinfection-Reinfection3	210 (69.6)	222 (78.4)	–	–	–	186 (62.9)
Reinfection2-Reinfection3	96.5 (69.6)	102 (88.4)	–	–	–	85.5 (19.1)
Doses of vaccine at primary infection: mean (SD)	1.08 (1.30)	- (0)	0.024 (0.154)	1 (0)	2.02 (0.140)	2.99 (0.115)
Doses of vaccine at first reinfection: mean (SD)	2.36 (0.876)	2.55 (0.514)	0.50 (0.506)	2.64(0.505)	2.39 (0.493)	2.99 (0.115)
Doses of vaccine at second reinfection: mean (SD)	2.29 (0.864)	2.60 (0.503)	0.83 (0.408)	–	–	2.80 (0.447)
Doses of vaccine at third reinfection: mean (SD)	2.50 (0.548)	2.50 (0.577)	–	–	–	2.50 (0.707)
Number of reinfections: mean (SD)	1.12 (0.375)	1.17 (0.451)	1.14 (0.354)	1.00 (0)	1.00 (0)	1.09 (0.372)

Table 2. COVID characteristics of the study population between 2020/03 and 2023/08 and clusters.

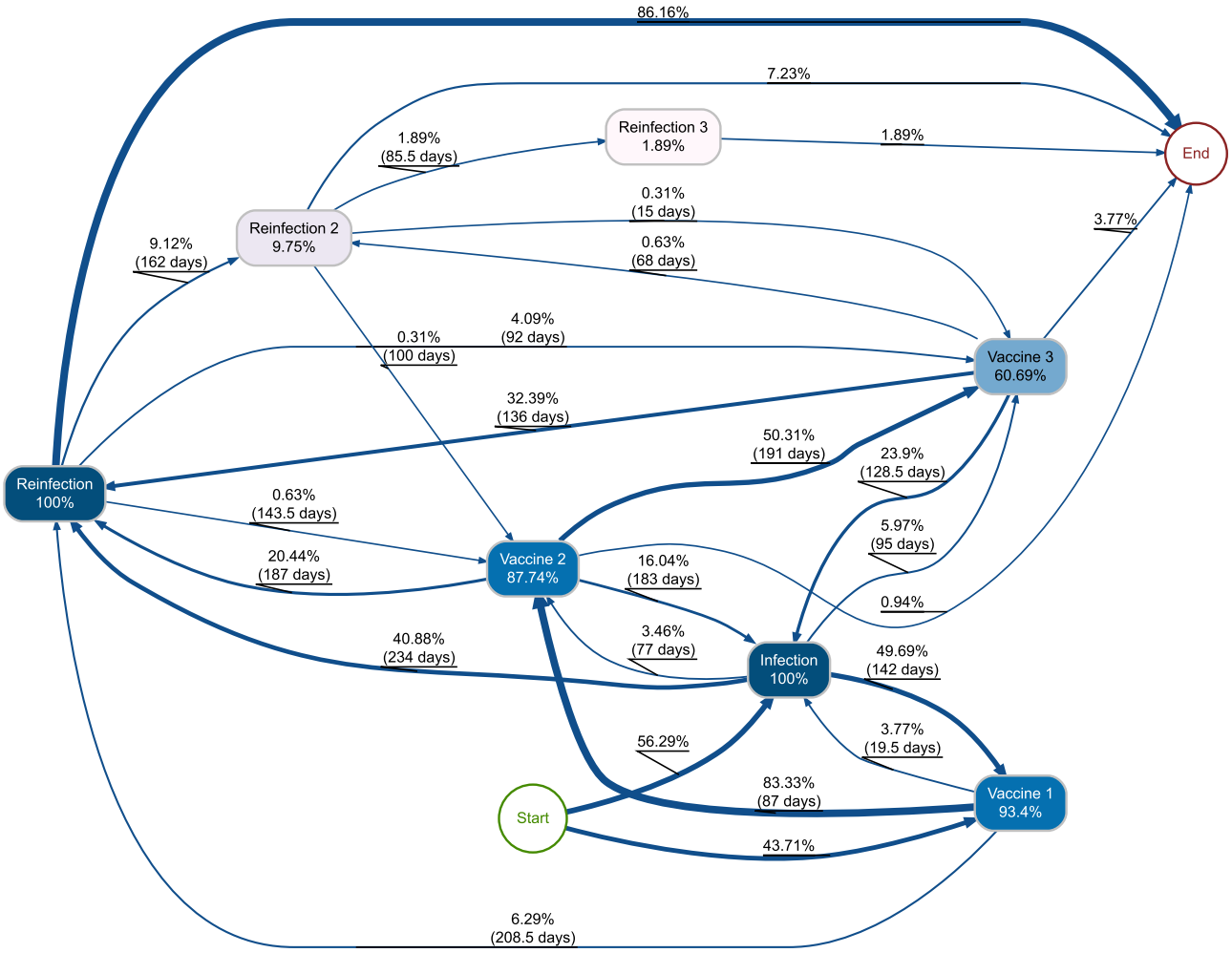


Fig. 1. Sequencing map of infection, reinfection and vaccination events in the cohort. The map was generated using process mining techniques. Boxes indicate main events and their relative frequency through the whole cohort. Edges represent the consecutive sequence of events and display the median time between each event as well as the relative frequency of participants following a specific sequence of events.

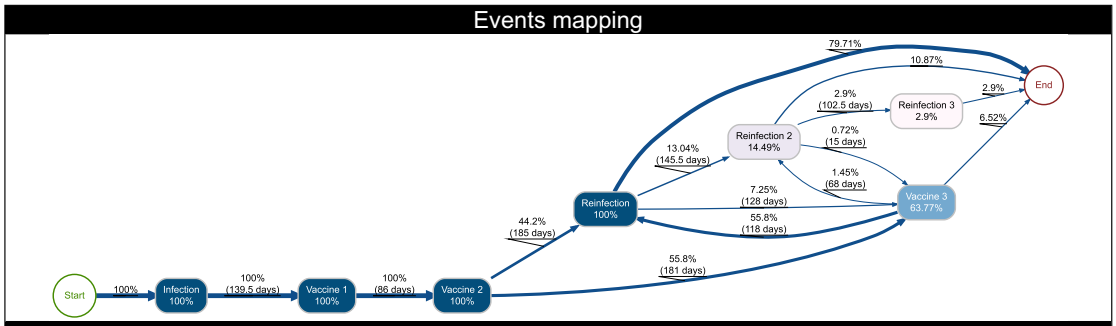
relatively short, with a mean and median delays of 49.5 and 16 days. Postinfection, individuals in the group all received a second vaccine before they split into two subsequences: those who were reinfected (36.4%) within a median of 215.5 days and those who received a third dose of vaccine (63.6%) within a median of 186 days. Those individuals took a median delay of 162 additional days before their reinfection. Within this group, 45.5% of individuals were reinfected in the fifth wave, and the remaining 54.5% were reinfected after the sixth wave, indicating that no reinfection occurred in wave 6 for this group.

Cluster 4. The fourth cluster was mainly infected in the fourth (31.4%) and fifth waves (58.8%), positioning the group, in terms of timeline of infection, between cluster 3 and cluster 5 (Table 2). For the entire group, the sequence of events began with two vaccines, as shown in Fig. 2D. For a majority (98%), the subsequent event is the primary infection. Once this event is reached, the group separates into two distinct trajectories: toward the third vaccine dose (37.3%, median delay 95 days) or the reinfection (62.8%, median delay 226 days). Individuals who received the 3rd dose after their primary infection were reinfected within a median of 198 days. In terms of the median delay from first vaccination, individuals with the V1-V2-I-R sequence had a *summed medians* delay of 479 days (71; 182; 226 days), compared with patients with the V1-V2-I-V3-R sequence, who had a *summed medians* delay of 546 days (71; 182; 95;198 days). Reinfections in this group occurred mainly from the seventh wave onward. It may be noted that one individual in this group presented the main temporal sequence of group 5 (V1-V2-V3-I-R1). Analysis revealed that this individual, even if it has the same sequence, received a third vaccine only one day before the first infection, making it more likely to belong to group 4 (V1-V2-I-R1).

Cluster 5. The last cluster contained the individuals who were most vaccinated prior to their primary infection. Indeed, 98.7% had received their third dose at the time of initial infection. Reinfection occurred latest among the other four groups, i.e., during waves five and six and mainly from wave seven onward (52.6%). This represents a median delay from infection of 235.5 days.

In addition, we noted interesting differences between the groups. As presented previously, at the time of their primary infection, individuals in the first cluster did not receive a vaccine, individuals in the third had

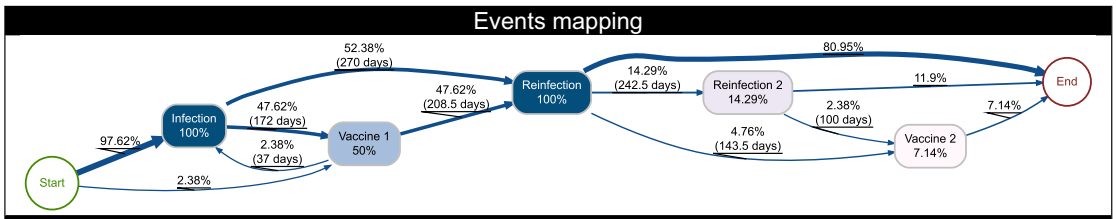
A



Events sequences

1	2	3	4	5	6	7	n	Cumulative frequency (%)
I	V1	V2	V3	R1			67	48,55
I	V1	V2	R1				43	79,71
I	V1	V2	R1	V3			8	85,51
I	V1	V2	V3	R1	R2		8	91,31
I	V1	V2	R1	R2			5	94,93
I	V1	V2	R1	R2	R3		2	96,38
I	V1	V2	V3	R1	R2	R3	2	97,83
I	V1	V2	R1	V3	R2		2	99,28
I	V1	V2	R1	R2	V3		1	100,00

B



Events sequences

1	2	3	4	5	6	7	n	Cumulative frequency (%)
I	R1						20	47,62
I	V1	R1					13	78,57
I	V1	R1	R2				4	88,10
I	V1	R1	V2				2	92,86
I	V1	R1	R2	V2			1	95,24
V1	I	R1					1	97,62
I	R1	R2					1	100,00

C

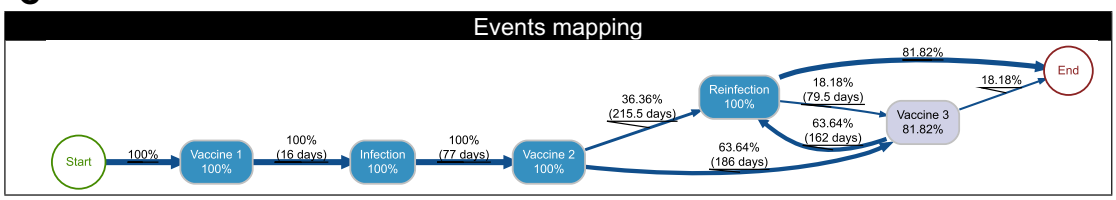
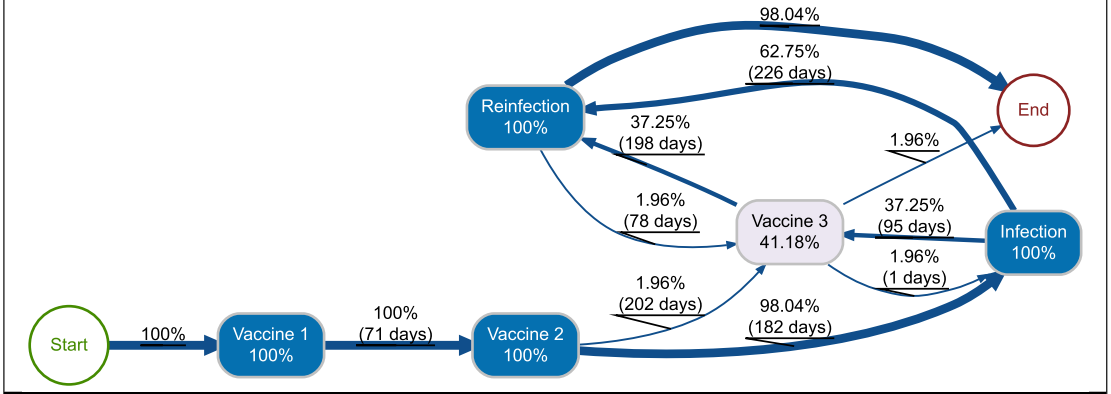


Fig. 2. Sequences and mapping of interest events, grouped by cluster. where **A**) is cluster 1, **B**) cluster 2, **C**) cluster 3, **D**) cluster 4 and **E**) cluster 5. Events mappings present the relative frequency of the trajectory (%) and the median in days between each event (I = Infection, V1 = 1st dose of vaccine, V2 = 2nd dose of vaccine, V3 = 3rd dose of vaccine, R1 = 1st reinfection, R2 = 2nd reinfection, R3 = 3rd reinfection).

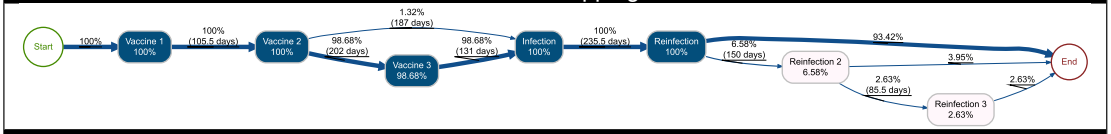
Events sequences								
1	2	3	4	5	6	7	n	Cumulative frequency (%)
V1	I	V2	V3	R1			7	63,64
V1	I	V2	R1				2	81,82
V1	I	V2	R1	V3			2	100,00

D



Events sequences								
1	2	3	4	5	6	7	n	Cumulative frequency (%)
V1	V2	I	R1				30	58,82
V1	V2	I	V3	R1			19	96,08
V1	V2	I	R1	V3			1	98,04
V1	V2	V3	I	R1			1	100,00

E



Events sequences								
1	2	3	4	5	6	7	n	Cumulative frequency (%)
V1	V2	V3	I	R1			71	93,42
V1	V2	V3	I	R1	R2		3	97,37
V1	V2	V3	I	R1	R2	R3	1	98,68
V1	V2	I	R1	R2	R3		1	100,00

Fig. 2. (continued)

mostly received one dose, individuals in the fourth, two doses (average 2.02; 95% confidence interval CI: 1.98–2.06), and those in the fifth, three doses (2.99; CI: 2.96–3.02). This situation presented a statistical difference between groups ($p < 0.001$), except for cluster 1, with cluster 2 ($p = 0.876$), and cluster 3, with cluster 4 ($p = 0.232$). Considering the doses at first reinfection, there was evidence of statistical difference in the average doses of the vaccine, except for cluster 1 (2.55; CI: 2,46–2,64), with cluster 3 (2.64; CI: 2.30–2.98) and cluster 4 (2.39; CI: 2.25–2.53), and cluster 3, with cluster 4 and cluster 5 (2.99; CI: 2.964–3.016). In clusters containing individuals who were reinfected twice (clusters 1, 2 and 5), there was a statistical difference between doses at second reinfection for clusters 1 and cluster 2 ($p < 0.001$) and between clusters 2 and cluster 5 ($p < 0.001$). However, for individuals who were reinfected three times (clusters 1 and 5), there was no evidence of statistical difference between groups.

Follow-up duration

In terms of follow-up duration, an analysis of the distributions revealed certain disparities. Group 1 showed a moderate distribution around the median (729 days) but was distinguished by the presence of outliers, suggesting that some participants in this group had below- and above-average follow-up times. This group was significantly different from groups 2–4-5 ($p < 0.001$). Group 3 stood out for its good homogeneity, low variability and absence of outliers, suggesting uniform follow-up (interquartile range (IQR) = 190, median 694 days),

and had no statistical difference with any of the other groups. Group 2, on the other hand, shows the greatest heterogeneity, with a much wider interquartile range (479 days), suggesting significant differences in follow-up duration between participants in this group. Despite showing similar visual distributions with close medians and moderate variability (median of 562 and 676 days; IQRs of 191 and 151 days), there was a significant difference in follow-up duration between groups 4 and 5 ($p < 0.001$).

In summary, the cohort participants were grouped into five clusters, and their characterization revealed that the clusters followed a temporal progression according to the infection timing and its positioning across the pandemic waves. Reinfections, on the other hand, occurred from the fifth wave onward. The most highly vaccinated groups appear to have been infected and consequently reinfected later in the pandemic. Some groups featured a greater proportion of healthcare workers, while for others, it was the trajectory and their timeframes that were of interest. There were some disparities in follow-up duration, which need to be considered when drawing conclusions from the results.

Discussion

This project aimed to study hybrid immunity by identifying and characterizing SARS-CoV-2 reinfection profiles. Using machine learning techniques, we grouped individuals from BQC19 according to characteristics leading to similar patterns of vaccination, infection, and reinfection in a five-cluster classification.

The study showed no significant differences between the groups in terms of sociodemographic variables, except for the proportion of healthcare workers. For this variable, groups 4 (51%) and 5 (76.3%) had a higher proportion than the cohort (44.3%). These same groups had a more sustained initial vaccination sequence than the other groups did (2 doses and 3 doses, respectively, before primary infection). This seemed consistent with the vaccination policies in place during the pandemic for this at-risk population in close contact with the virus. The results therefore suggest that these policies had a positive impact, given that, for this group, primary infection occurred later during the pandemic. However, this finding, implying that healthcare workers in the cohort were infected late (59.6% of them), differs from the results of Carazo et al. (2023) in their study about healthcare workers' protection against Omicron BA.2 reinfection conferred depending on the primary infection variant, where, for around the same period, approximately 20.7% of healthcare workers were infected[21]. The difference may be explained by the inclusion criteria for the documented dates in our study, which reduced the size of our cohort. This contrasts with their study, which exploited data sources from the Ministry of Health and Social Services that were potentially more exhaustive at this level.

Similarly, it was possible to observe that group 3, which received a first vaccine before being infected, was spared in wave 1. Thus, compared with groups 1 and 2, which received no vaccine prior to infection, group 3's primary infection occurred later in waves 2 and 3, allowing us to hypothesize that although one dose was missing to achieve so-called complete vaccine immunization, the first vaccine dose may have generated a positive impact by delaying the initial infection. However, this hypothesis must be interpreted with caution, given the small number of individuals in the group.

The first group also presented interesting features in terms of vaccination efficacy. Indeed, following the second vaccine, the trajectory of the individuals split in two, some toward reinfection (median of 185 days after), while the others toward the third vaccine (median of 181 days after). This separation occurred within an almost identical median time, which might suggest that the policy of administering the third dose was relatively synchronized with a weakening of immunity. This timeframe is in line with the results of Asamoah-Boaheng et al. (2023) showing that antibody levels decrease with a half-life of 94 days and plateauing at 294 days[22]. Although these results relate to mRNA vaccines and vaccine type was not a variable in the present study, the results remain consistent. Also, the 3rd vaccine appeared to have delayed reinfection by 118 days (median) compared with patients who received only 2 doses. This suggests that an earlier 3rd dose could potentially have delayed more reinfections. Group 3 had a similar separation between the reinfection event and the 3rd vaccine. The latter, whose sequence prior to separation was V1-I-V2, compared with the first group's I-V1-V2, had a greater *summed medians* time to reinfection (215 vs. 185 days). Similarly, when comparing the median time from second vaccine to reinfection via third vaccine dose, group 3 had a longer *summed medians* time (348 vs. 299 days). This might suggest that, in terms of hybrid immunity, being infected between two vaccine doses could offer slightly longer-lasting immunity, but the size of the 3rd group makes it difficult to draw such a definitive conclusion.

The study also revealed a group of patients who had not been vaccinated and, consequently, had not achieved hybrid immunity. This same group also contained individuals who had received only one vaccine, implying that they had not fully achieved hybrid immunity, given that full vaccine immunity required 2 doses. Thus, based exclusively on the sequence of events and their temporality, individuals with partial or nonexistent hybrid immunity were grouped together by the clustering algorithm. Despite its interest, this finding needs to be nuanced according to the variance in follow-up duration within the group. In fact, some individuals may simply not have had the follow-up time required to be fully vaccinated. Despite this limitation, it is worth mentioning that the algorithm's data-driven grouping of these participants supports the finding of Sanchez-de Prada et al. (2024) that there is no significant difference between individuals vaccinated once within five months of infection and those who were not vaccinated at all.

The results showed that vaccination has a positive effect in delaying infection or reinfection. They also showed that the temporality of events greatly influenced the formation of groups by the algorithm, in the sense that primary infections and reinfections are distributed according to a temporal progression, from group 1 (the earliest infections) to group 5 (the latest).

This study has several strengths. First, we used data collected as early as the beginning of the pandemic, which allowed us to use valuable data for this study. The data management process is also a great strength of this study, as substantial work has been performed to consolidate the data, allowing us to increase our sample size. Finally, the use of machine learning made it possible to identify more complex patterns by taking events

and their temporality into account, using a data-driven approach. In fact, the method enabled us to consider not only the delay between events but also their chronology, thereby accounting for pandemic waves. Thus, by first forming the groups and then characterizing them using variables that were not used in the clustering process, we were able to highlight elements that were more difficult to identify using conventional methods. Unsupervised techniques reveal interesting avenues for future investigations. To do so, we intend to use existing genomic data in BQC19 to characterize the groups, including the use of random forest to determine the most relevant variables for this purpose. As genomic data were sequenced only for a rather small sample size, it was not possible to include it in this study, as our sample size was already reduced due to inclusion criterion about documented reinfection with a date.

While there are strengths, there are also some limitations to this work. First, this study is based on longitudinal data obtained from individuals who agreed to participate in the BQC19 and its follow-up, which may introduce a selection bias. Also, even though significant attention has been given to data management, the sample size has remained small. This is mostly due to our inclusion criteria, where people need to be reinfected to be included. Second, while the infection and vaccination dates were properly collected within the datasets, we had to establish a strategy to correct the reinfection date, as two variables were present within the same dataset. However, the dates were the same for the vast majority of participants, and the same treatment was applied otherwise, limiting potential biases. In addition, the delay used to define the reinfection in this study differed from the delay found in the literature and could be considered as a limitation. Considering the lack of official consensus in the scientific community regarding this timeframe, the choice of timeframe (14 days) has been made using the threshold used among the BQC19 research community to enable meaningful comparisons. However, even with this short delay, the number of patients concerned was relatively small and, according to the distribution of delays between reinfections, increasing the threshold of consideration would have only a minimal impact on the number of individuals. Also, results showing differences between groups need to be interpreted with caution due to the variance in follow-up time within groups, given that some individuals may simply not have had the follow-up time required to have been fully vaccinated. Further analysis would be relevant to assess the impact of the difference in follow-up time for participants. Similarly, some events may also have been missed if they occurred outside the scope of participating hospitals to BQC19. Additionally, even if it assures consistency with other BQC19 works, using pandemic waves as a proxy for dominant circulating variants in the absence of variant-level data has limitations, given that multiple variants may have co-circulated during the same wave. Due to the role of variants in immune escape risk, the reinfections profiles obtained should be interpreted with caution, as they may thus reflect unmeasured heterogeneity in terms of variants. Finally, the use of *summed medians* can induce a distortion in the estimation of overall times. However, it was only used to provide an overall idea of the temporality of the sequences and the global trajectory, as we know that it is not the actual median of the whole sequence.

Conclusion

To our knowledge, this is the first study using data from the *Biobanque québécoise de la COVID-19* to investigate reinfection patterns and hybrid immunity using a data-driven approach. In addition to highlighting the effectiveness of vaccination policies, it identified, by leveraging machine learning techniques on complex multidimensional time series, distinct groups and COVID-19 patterns of infection, reinfection and vaccination, thus providing interesting insights for further investigation. It also highlights that beyond the sequence of events, the temporal delays between events seem to play an important role in the acquisition of primary and secondary infections. In terms of hybrid immunity, the results of this study suggest that an infection between two vaccines could offer greater immunity. This finding should be treated with caution, however, given the size of the group from which it is drawn and the disparity in follow-up times. In any case, this is an interesting perspective to pursue. The delay between events played a determining role in the formation of the study groups. Consequently, their consideration in the development and adaptation of health policies, particularly regarding vaccine administration and boosters, is necessary. In addition, the study shows that machine learning algorithms represent, for public health practices, an innovative and complementary approach to analyze health data and discover hidden information that can have an impact on public health decisions. These two approaches, which combine delay analysis and machine learning, offer promising perspectives for future work, particularly in preparation for possible pandemics.

Data availability

The data that support the findings of this study are available from the BQC19 following the standard access procedures required to obtain participant-derived data at https://www.quebeccovidbiobank.ca/acceder-au-materiel-de-la-bqc19 .

Received: 15 May 2025; Accepted: 3 September 2025
Published online: 07 October 2025

References

1. Livieratos, A., Gogos, C. & Akinosoglou, K. Impact of Prior COVID-19 Immunization and/or Prior Infection on Immune Responses and Clinical Outcomes. *Viruses* https://doi.org/10.3390/v16050685 (2024).
2. World Health Organization: WHO COVID-19 dashboard - Data reported on 28 april 2024. https://data.who.int/dashboards/covid19/. Accessed 17 juillet 2024. (2024)
3. Lapuente, D., Winkler, T. H. & Tenbusch, M. B-cell and antibody responses to SARS-CoV-2: infection, vaccination, and hybrid immunity. *Cell Mol. Immunol.* **21**(2), 144–158. https://doi.org/10.1038/s41423-023-01095-w (2024).

4. Hu, B., Guo, H., Zhou, P. & Shi, Z. L. Characteristics of SARS-CoV-2 and COVID-19. *Nat. Rev. Microbiol.* **19**(3), 141–154. https://doi.org/10.1038/s41579-020-00459-7 (2021).
5. Zhu, N. et al. A Novel Coronavirus from Patients with Pneumonia in China, 2019. *N. Engl. J. Med.* **382**(8), 727–733. https://doi.org/10.1056/NEJMoa2001017 (2020).
6. World Health Organization, Kryuchkov I: Post COVID-19 condition (Long COVID). https://www.who.int/europe/news-room/fact-sheets/item/post-covid-19-condition. Accessed 9 septembre 2024. (2022)
7. Misra, A. & Theel, E. S. Immunity to SARS-CoV-2: What Do We Know and Should We Be Testing for It?. *J. Clin. Microbiol.* **60**(6), e0048221. https://doi.org/10.1128/jcm.00482-21 (2022).
8. Rodriguez Velásquez, S. et al. Long-term levels of protection of different types of immunity against the Omicron variant: a rapid literature review. *Swiss Med. Wkly.* **154**, 3732. https://doi.org/10.57187/s.3732 (2024).
9. Tremblay, K. et al. The Biobanque québécoise de la COVID-19 (BQC19)—A cohort to prospectively study the clinical and biological determinants of COVID-19 clinical trajectories. *PLoS ONE* **16**(5), e0245031. https://doi.org/10.1371/journal.pone.0245031 (2021).
10. Janssenswillen, G., Depaire, B., Swennen, M., Jans, M. & Vanhoof, K. bupaR: Enabling reproducible business process analysis. *Knowl.-Based Syst.* **163**, 927–930. https://doi.org/10.1016/j.knosys.2018.10.018 (2019).
11. Sakoe, H. & Chiba, S. Dynamic programming algorithm optimization for spoken word recognition. *IEEE Trans. Acoust. Speech Signal Process.* **26**(1), 43–49. https://doi.org/10.1109/TASSP.1978.1163055 (1978).
12. Rousseeuw, P. J. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *J. Comput. Appl. Math.* **20**, 53–65. https://doi.org/10.1016/0377-0427(87)90125-7 (1987).
13. Giorgino, T. Computing and Visualizing Dynamic Time Warping Alignments in R: The dtw Package. *J. Stat. Softw.* **31**(7), 1–24. https://doi.org/10.18637/jss.v031.i07 (2009).
14. GISAID via CoVariants.org: – with major processing by Our World in Data. "Alpha" . GISAID, via CoVariants.org, COVID-19, sequencing [original data]. Accessed 2024. (2025).
15. Mathieu E, Ritchie H, Rodés-Guirao L, Appel C, Gavrilov D, Giattino C, et al.: COVID-19 Pandemic. https://ourworldindata.org/coronavirus. Accessed 2024 (2020).
16. R Core Team. R : A Language and Environment for Statistical Computing. Vienna, Austria: R Foundation for Statistical Computing (2023).
17. Posit team. RStudio: Integrated Development Environment for R. PBC, Boston, MA: Posit Software (2024).
18. Wickham, H. et al. Welcome to the Tidyverse. *J. Open Sour. Softw.* https://doi.org/10.21105/joss.01686 (2019).
19. Sarda-Espinosa A. dtwclust: Time Series Clustering Along with Optimizations for the Dynamic Time Warping Distance. R package 5.5.12 ed2023.
20. Meyer D, Buchta C. proxy: Distance and Similarity Measures. R Package. 0.4–27 ed2022.
21. Carazo, S. et al. Protection against omicron (B.1.1.529) BA.2 reinfection conferred by primary omicron BA.1 or pre-omicron SARS-CoV-2 infection among health-care workers with and without mRNA vaccination: a test-negative case-control study. *Lancet Infect. Dis.* **23**(1), 45–55. https://doi.org/10.1016/S1473-3099(22)00578-3 (2023).
22. Asamoah-Boaheng, M. et al. Eleven-month SARS-CoV-2 binding antibody decay, and associated factors, among mRNA vaccinees: implications for booster vaccination. *Access Microbiol.* https://doi.org/10.1099/acmi.0.000678.v3 (2023).

Acknowledgements

This work was made possible through open sharing of data and samples from the Biobanque québécoise de la COVID-19, funded by the Fonds de recherche du Québec—Santé, Génome Québec, the Public Health Agency of Canada and, as of March 2022, the ministère de la Santé et des Services sociaux. We thank all participants to BQC19 for their contribution. https://www.quebeccovidbiobank.ca

Author contributions

All authors contributed to the study design. J.F.B. cleaned, analyzed and interpreted the data, and drafted the manuscript while D.L., S.R. and D.B.R. revised it extensively. S.R. and D.B.R. contributed equally. All authors read and approved the final manuscript.

Funding

This project was funded through an award from the Fonds de recherche du Québec—Santé supported networks: Réseau de recherche en santé respiratoire du Québec (RSRQ) et le Réseau de recherche en santé des populations du Québec (RRSPQ).

Declarations

Competing interests

The authors declare no competing interests.

Ethics approval and consent to participate

Ethics approval has been obtained from Centre universitaire de santé McGill's ethics committee within the framework of the project *Determining the impact of hybrid immunity on the evolving landscape of host responses to SARS-CoV-2 in the Biobanque Québécoise de la COVID-19 (BQC19)* (ref. 2023–9261). Each BQC19 enrolling site has established a consent process that reflects the BQC19's standard operating procedures. See https://doi.org/10.1371/journal.pone.0245031 for more information.

Additional information

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1038/s41598-025-18706-3.

Correspondence and requests for materials should be addressed to D.B.-R.