Title: Influence of Prior SARS-CoV-2 Infection on COVID-19 Severity: Evidence from the National COVID Cohort Collaborative

Authors: Nathaniel Hendrix, PharmD, PhD; Hythem Sidky, PhD; David K. Sahner, MD, The N3C Consortium

Corresponding author:

Nathaniel Hendrix
Center for Professionalism and Value in Health Care, American Board of Family Medicine
1016 16th St NW Ste 800, Washington, DC 20036
nhendrix@theabfm.org

Full name, department, institution, city, and country of authors:

Hythem Sidky, PhD
National Center for Advancing Translational Sciences, National Institutes of Health, Bethesda, MD, USA

David K. Sahner, MD
National Center for Advancing Translational Sciences, National Institutes of Health, Bethesda, MD, USA

Consortium contributions from: Christopher G. Chute

Keywords: COVID-19; SARS-CoV-2; COVID-19 Reinfection; SARS-CoV-2 Reinfection

Funding: This project was sponsored by an award from the National COVID Cohort Collaborative's Public Health Answers to Speed Tractable Results (PHASTR) program and the National Center for Advancing Translational Sciences (Award #NCATS-P00438-E-2).

Word count: 2,907

Acknowledgment: The analyses described in this publication were conducted with data or tools accessed through the NCATS N3C Data Enclave https://covid.cd2h.org and N3C Attribution & Publication Policy v 1.2-2020-08-25b supported by NCATS U24 TR002306, Axle Informatics Subcontract: NCATS-P00438-B. This research was possible because of the patients whose information is included within the data and the organizations (https://ncats.nih.gov/n3c/resources/data-contribution/data-transfer-agreement-signatories) and scientists who have contributed to the on-going development of this community resource [https://doi.org/10.1093/jamia/ocaa196].

Disclaimer: The N3C Publication committee confirmed that this publication <msid:1491.197> is in accordance with N3C data use and attribution policies; however, this content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health or the N3C program.

Research ethics: The N3C data transfer to NCATS is performed under a Johns Hopkins University Reliance Protocol # IRB00249128 or individual site agreements with NIH. The N3C Data Enclave is managed under the authority of the NIH; information can be found at

https://ncats.nih.gov/n3c/resources. The study design was exempted from human subjects research review by the American Academy of Family Medicine institutional review board.

Abstract

Background: As SARS-CoV-2 has transitioned from a pandemic to endemic disease, the majority of new infections have been among previously infected individuals. To manage the risks and benefits of ongoing COVID-19 policies, it is important to understand whether prior infection modifies the severity of subsequent infections.

Methods: We used data from first and second COVID-19 episodes in the National COVID Cohort Collaborative (N3C), a collection of health systems who provide de-identified electronic health records for research purposes. Our analysis was a sequential series of nested trial emulations. In the first of two analytic stages, we created a month-specific model of the probability of prior infection for each individual. In the second stage, we used an ordinal logistic regression with inverse probability weights calculated in the first stage to simulate a series of monthly trials comparing severity between the cohorts of first and second infections. In addition to cohort-wide effect estimates, we also conducted analyses among race/ethnicity, sex, and age subgroups.

Results: From an initial cohort of 7,446,481 combined first and second infections, we identified a cohort of 2,227,484 infections, among which 7.6% were second infections. Ninety-four percent of patients with two recorded infections experienced mild disease for both. The overall odds ratio (OR) for more severe disease with prior infection was 1.06 (95% confidence interval [CI]: 1.03 – 1.10). Monthly point estimates of the OR ranged from 0.56 (95% CI: 0.37 – 0.84) in October 2020 to 1.64 (95% CI: 1.33 – 2.00) in February 2023. In most subgroups, the effect of prior infection was significant. In 8 out of 10 subgroups, the maximum monthly OR occurred after the minimum monthly OR, suggesting that protection has waned throughout the pandemic.

Conclusion: Overall, prior infection was associated with a significant slightly elevated risk of severe disease. This effect varied month to month. As the pandemic proceeded, the effect of prior infection tended to evolve from generally protective during the pre-Omicron era to unprotective during the Omicron era. This points to the need for continued strategies to avert and minimize the harms of COVID-19, rather than relying upon immunity acquired through previous infection.

Key Points

Question: Does prior infection with SARS-CoV-2 affect the severity of subsequent COVID-19 episodes?

Findings: We observed a mild protective effect of prior infection during the early and mid-stages of the pandemic that waned after the rise of the Omicron variants, ultimately resulting in loss of protection or a tendency toward more severe second infections.

Meaning: Prior infection alone is likely not enough to avert the worst public health harms of endemic SARS-CoV-2. Interventions to avoid infection and reduce the severity of COVID-19 will still be important in the post-pandemic era.

Introduction

The SARS-CoV-2 pandemic's shift into an endemic disease has necessarily been accompanied by substantial adjustments to public health policy. Policies around recommendations for vaccine boosters, public investment in novel treatments for COVID-19, and reimbursement for telemedicine have already changed or begun to be debated.(1) As of late 2022, an estimated 94% of Americans had experienced at least one episode of COVID-19.(2) Since the majority of new cases have been among people with prior exposure to the SARS-CoV-2 virus, an understanding of how prior infection changes disease severity is vital to making policy decisions informed by reasonable risk-benefit calculations.

One of the hallmarks of an infectious disease's transition from epidemic to endemic status is a reduction in disease severity due, in part, to an increasing number of the population with some immune protection against the disease.(3) In many cases, this takes place through prior infections. It is therefore reasonable to hypothesize that prior infection with SARS-CoV-2 also confers some protective effect, although some evidence has suggested that the protective effect of natural antibodies against SARS-CoV-2 may decay relatively quickly compared to antibodies against other viruses.(4)

Published efforts to understand the severity of repeat COVID-19 episodes have been surprisingly rare. This may have been due in part to the notable challenges of studying repeat infection.(5) Among other challenges, these have included the emergence of several new strains since the start of the pandemic, the introduction of vaccines, and the likely increase in the rate of self-management for COVID-19. Rigorously studying reinfection severity requires both large datasets of comprehensive health information and sophisticated analytic methods.

In this study, we used the National COVID Cohort Collaborative (N3C) dataset to conduct an analysis of reinfection severity. The N3C collected deidentified data from 76 health systems covering 18.9 million individuals and 7.5 million COVID-19 cases to provide comprehensive information on a large, diverse swath of patients throughout the United States.(6) We used causal methods with these data to assess the relevance of reinfection to COVID-19 severity both at the population level and across subgroups defined by gender, race/ethnicity, and age.

Methods

Methods Summary

To answer the question of how prior infection affects disease severity, we used a nested trial design to conduct this target trial emulation in which each month was treated as a separate trial and the effect size was estimated within that month. This design allowed us to cope with the challenges created by time-varying risks of (re)infection, vaccine penetration, severe disease related to variant pathogenicity, and evolving availability of therapies.

Data Source

The National Institutes of Health (NIH) sponsored the creation of the N3C dataset, which is the largest collection of information on COVID-19 infections in the United States. The N3C uses electronic health record data volunteered from healthcare organizations on patients in their systems going as far back as January 1, 2018 and extending to the present. The data are harmonized into a single data standard using the Observational Medical Outcomes Partnership

(OMOP) ontology and hosted in a secure, online portal for researchers to access. Episodes of COVID were defined by the N3C COVID-19 Phenotype, v4.0.(7) This phenotype required either a positive lab test (PCR or antigen) or diagnosis code specific to COVID-19. For cases prior to May 1, 2020, at which point diagnostic testing was not widely available, the phenotype allowed for identification of COVID-19 via two weak positive diagnostic codes documented on the same day. These weak positive codes included exposure to virus and symptoms.

Cohort Identification

We included only individuals who had at least one recorded COVID-19 infection in our analysis. All included patients were required to have at least one recorded visit in the year prior to their first recorded infection. This lookback period increased our confidence in the availability of baseline medical history, including 32 comorbidities with possible associations to COVID-19 outcomes (Supplemental Material I).(8) We also only included data from eight health systems who had been established to have vaccination rates that generally matched with their locale's overall vaccination rate.(9)

We defined reinfection as the first infection to take place at least 60 days after the earliest recorded infection. To further ensure that participant privacy was protected, we only included reinfections taking place in months with a total of at least twenty infections.

Outcomes

We used the five-level classification of COVID-19 severity defined by N3C. The classifications were:

- Mild disease: Seen in an outpatient, non-emergent setting without subsequent hospitalization or death
- Mild disease in emergency department (ED): Seen in emergency department and released without subsequent hospitalization or death
- Moderate disease: Hospitalized without extracorporeal membrane oxygenation (ECMO), invasive mechanical ventilation (IMV), or death
- Severe disease: Hospitalized with ECMO and/or IMV, but without death
- Death: Discharge status to "death" or "hospice" after a hospitalization with COVID-19

Mortality data in the N3C Enclave was enriched through privacy-preserving record linkage to ancillary sources of death data, including supplemental entries (SSA), private obituary entries, and obit.com entries. Our outcome of interest was the counterfactual distribution of COVID-19 severity among the cohort of those with a second infection at the time of infection had it instead been their first.

Analytical Approach

We analyzed data on disease outcomes as if it were produced by a sequential series of nested trials, each starting within a given month. Thus, within each month with at least 20 reinfections, we compared the outcomes of those with and without prior COVID-19. This approach allowed us to manage the challenges associated with the multitude of time-varying influences on severity throughout the pandemic.(10)

We analyzed the observational data using a two-stage approach. In the first stage, we used a logistic regression within each month to predict the probability that the individual had a prior

recorded episode of COVID-19. The regression included the following independent variables: age, age squared, race/ethnicity, sex, health system data source, whether the patient's postal code was recorded, count of vaccinations having taken place at least 2 weeks prior to the infection date (categorical from 0 to more than 4), and body mass index (categorical as <18.5, 18.5 - 24.9, 25 - 29.9, 30 - 34.9, 35 – 40, >40, or missing). We also included in the regression a binary indicator for each of the comorbidities listed above. We converted the predicted probability of prior exposure into stabilized inverse probability of treatment weights. The balance of potentially outcome-relevant covariates was confirmed by plots of standardized mean differences using a threshold of 0.1 (Supplemental Material II).

In the second stage, we used an ordinal logistic regression to identify the effect of prior infection on disease severity. We estimated the pandemic-wide effect on the 5-dimensional severity outcome with a regression wherein the only independent variable was prior infection. We also calculated the monthly effect of exposure by including an interaction term between the binary variable for prior infection and a categorical variable for the month. In both regressions, each patient was weighted by the inverse probability of treatment weights calculated in the first stage of the analysis, thereby also accounting for covariates that affect disease severity. We also accounted for temporal correlation in the error term by clustering standard errors at the month level. This method recognizes that observations within the same month may be more similar to each other than observations in different months, thus potentially violating the assumption of independent errors.

Subgroup Analyses

Other studies of reinfection severity were conducted among relatively homogenous groups. Our use of the N3C dataset allowed us to examine the effect of prior infection for several subgroups:

- Black / African American, non-Hispanic
- Hispanic of any race
- White, non-Hispanic
- Female
- Male
- Age less than 18 years
- Age 18-44 years
- Age 45-59 years
- Age 60-69 years
- Age 70 or more years

We recalculated the inverse probability weight in each subgroup analysis because of the differing reference groups in the subgroup versus population analyses.

Sensitivity Analyses

Our analysis hinged on several important assumptions. We tested the relevance of these assumptions to our findings through several sensitivity analyses. Namely, we first evaluated the impact of prior infection on a cohort without the one-year lookback period required in the main analysis. In another analysis, we changed the definition of a second infection from an infection taking place at least 60 days after initial infection to one taking place at least 90 days later. Finally, we used a simpler analytic strategy to examine the effects of the more stringent

assumptions required for ordinal versus binary logistic regression. We looked at the association of prior infection with two outcomes: death and/or hospitalization.

Results

Cohort

From an initial cohort of 7,446,481 combined first and second infections, we excluded observations due to lack of visits during the look-back period, origination from a health system without vaccination rates comparable to locally reported rates (i.e., presumably invalid vaccine data related to incomplete capture in the EHR), or infection in a month with fewer than 20 reinfections until we arrived at a combined cohort of 2,227,484 infections from 2,058,274 distinct individuals (Figure 1).

Individuals with a recorded second infection were relatively similar to the cohort of individuals during their first infections (Table 1). The cohort with a recorded second infection was more likely to be female and somewhat more likely to be vaccinated, but overall had similar rates of comorbidities (Supplemental Material III).

Longitudinal Trends among Previously Reinfected Patients

At least two outcome-relevant characteristics of the cohort of previously infected individuals changed dramatically over the course of the pandemic (Figure 2). During 2020, the average time between first and second infection among this cohort was under 5 months, and vaccination against SARS-CoV-2 was minimal. During the first three months of 2023, however, first and second infections were 16 months apart on average, and approximately two-thirds of the reinfected cohort had at least one vaccination.

Reinfection Outcomes

Among patients with a recorded reinfection, 94% experienced mild disease during both infections. More severe outcomes were much rarer. Approximately 0.5% of reinfections resulted in death. Among individuals who had a severe initial infection, 5.9% died as a result of their second infection compared to much lower rates for individuals who had a less severe initial infection.

Cohort-Level Results

The pandemic-wide effect of prior infection on the complete cohort was slightly but significantly predictive of greater severity (Figure 3). The estimated OR of more severe disease given prior infection was 1.06 (95% confidence interval [CI]: 1.03 – 1.10).

There was substantial heterogeneity in the impact of prior infection on COVID-19 outcomes across the pandemic, however. There was a total of 7 months where reinfection offered a protective effect, all prior to March 2022. After this point, there were 6 months in which prior infection was associated with increased odds of more severe disease. Many monthly confidence intervals were large and, in 20 of the 33 months analyzed, crossed unity.

Subgroup Analyses

The pandemic-wide analysis revealed that, for most of the subgroups we analyzed, prior infection had a significant association with disease severity, though there were differences in the magnitude and directionality of this effect (Supplemental Material IV). The greatest protective effect was observed among patients aged 70 or greater (OR: 0.89 [0.85 to 0.94]), although prior infection among patients in all groups aged 45 and greater was associated with a significant protective effect (OR for 45- to 59-year-old patients: 0.92 [0.87 to 0.97], OR for 60- to 69-year-old patients: 0.92 [0.86 to 0.99]). In the pandemic-wide analysis, prior infection among male patients (OR: 1.11 [1.06 to 1.16]), patients aged less than 18 years (OR: 1.34 [1.22 – 1.49]), and black / African American patients (OR: 1.08 [1.01 – 1.15]) was significantly associated with more severe disease. No subgroup had a consistently significant difference of effects compared to other groups in the monthly analysis.

Sensitivity Analyses

The sensitivity analyses assessing the impact of our inclusion criteria and definition of second infection showed largely the same results as our main analysis, albeit with somewhat more attenuated effects (Supplemental Material VI, VII). Our binomial logistic regressions also validated the findings of our ordinal logistic regression, while also showing an increased risk of hospitalization and mortality in the periods soon after the shift from pre-Alpha to Alpha variants and from Delta to Omicron (Figure 4).

Discussion

Across the pandemic, we found that prior infection slightly but significantly increased the odds of more severe COVID-19 illness. This analysis is perhaps less relevant than the monthly effects of disease, which shifted substantially over the approximately 3 years of the analysis. In 8 of the 10 subgroups and in the cohort at large, the maximum observed OR for prior infection was after the minimum observed OR, indicating the evaporation of protective effect and the possibility of a harmful one.

The changes in the impact of prior infection that we observed must be interpreted in the context of monthly cohorts with very different characteristics. The introduction and broad uptake of vaccination, the shifting dominance of new SARS-CoV-2 variants, and the increasing time between first and second infection all likely played a role in the patterns we identified in this study. At the same time, we did not design this study to test those effects. Therefore, our interpretation of the roles played by these time-varying influences is necessarily very speculative.

Our findings may be consistent with a scenario in which infection has both protective and harmful effects of differing durations: the protective effects predominate in the short term, while harmful effects are longer lasting. In this explanation, the benefits of infection to immune readiness subside while damage to the lungs, vasculature, and other systems have not been fully healed.(11) Prior studies have found that COVID-19 confers protection against subsequent infection for approximately one year.(12) The point in the pandemic when average time between first and second infection began to exceed one year was the same point at which the first harmful effect of prior infection was observed in the cohort-wide analysis. This lends some credibility to our hypothesis, as does the appearance of potential harm from previous infection at points when SARS-CoV-2 had mutated most dramatically.(13)

Our findings have relevance to important questions of public health policy: specifically, how much can we rely on immunity from prior infection to reduce the public health consequences of

endemic COVID-19 and how much will we need to continue to invest in vaccination? The lack of a persistent protective effect suggests that relying exclusively on prior infection will be deleterious to public health.

At least three previous studies of reinfection severity have been published with contradictory results. The studies by Abu-Raddad, et al. and Mensah, et al. used nationwide surveillance from Qatar and the United Kingdom, respectively.(14,15) Both found significantly diminished severity among individuals with prior infection. On the other hand, Bowe, et al. found a substantially increased risk of death and multiple post-acute sequelae among patients in the United States Veterans Affairs healthcare system with more than one SARS-CoV-2 infection, although their study design was compromised by immortal time bias and issues with the timing of outcome ascertainment among singly infected individuals.(16)

Our study had several strengths. Among these was the use of inverse probability weights to simulate randomization to reinfection. This allowed us to simulate an analysis in which the previously and never-before infected groups had the same distribution of outcome-relevant covariates. Our dataset was also a major strength. The N3C dataset allowed us to access high-quality information on comorbidities, demographics, and vaccination from diverse health systems, which supported our ability to perform subgroup analyses. The sample size available within N3C also meant that we could explore how the impact of prior infection has varied across the duration of the pandemic. This is a novel contribution of the study that increases our study's policy relevance, despite the large confidence intervals observed in many months.

At the same time, our study had multiple limitations. The presumed unreliability of vaccination data at some sites meant that we had to exclude data partners for whom we could not validate the observed prevalence of vaccination against their local vaccination rates. We also had no way of validating individual-level vaccination data. Next, we had no way of validating that first and second recorded infections were, in fact, the first or second infections that patients had experienced. In practice, the cohort of infections we deemed "first" was likely a mix of first and subsequent infections with ratios that varied across the pandemic. Although we attempted to control for this possibility by using time in the ordinal logistic regression, we may underestimate any protective effect of prior infection. Similarly, our IPW approach could not include all key variables related to risk of prior infection, such as masking and distancing behavior, given the limitations of the data set. In addition, the widespread use of home testing late in the pandemic may have led to missed mild Omicron infections, inflating Omicron severity. Finally, we cannot precisely attribute death to COVID since it was based on discharge status instead of cause of death.

Conclusion

Our analysis of a large, diverse collection of electronic health records for COVID-19 patients revealed that prior infection was slightly but significantly predictive of more severe disease in the entire cohort. More importantly, though, we observed an apparent evaporation of protective effect over the pandemic, as well as suggestions that prior infection may be associated with more severe disease in the Omicron era. Our results imply that acquired immunity through infection cannot be relied on to provide comprehensive protection against severe disease. In turn, public health should benefit from continued investments in vaccination and treatment.

References

1. Shen AK, Schwartz JL. COVID-19 From Pandemic To Endemic: Strengthening Routine Adult Vaccination Programs. Health Affairs Forefront [Internet]. [cited 2023 Apr 21]; Available from: https://www.healthaffairs.org/do/10.1377/forefront.20220906.804775/full/

2. Klaassen F, Chitwood MH, Cohen T, Pitzer VE, Russi M, Swartwood NA, et al. Changes in population immunity against infection and severe disease from SARS-CoV-2 Omicron variants in the United States between December 2021 and November 2022 [Internet]. medRxiv; 2022 [cited 2023 Apr 21]. p. 2022.11.19.22282525. Available from: https://www.medrxiv.org/content/10.1101/2022.11.19.22282525v3

3. Antia R, Halloran ME. Transition to endemicity: Understanding COVID-19. Immunity. 2021 Oct 12;54(10):2172–6.

4. Ibarrondo FJ, Fulcher JA, Goodman-Meza D, Elliott J, Hofmann C, Hausner MA, et al. Rapid Decay of Anti–SARS-CoV-2 Antibodies in Persons with Mild Covid-19. New England Journal of Medicine. 2020 Sep 10;383(11):1085–7.

5. Babiker A, Marvil CE, Waggoner JJ, Collins MH, Piantadosi A. The Importance and Challenges of Identifying SARS-CoV-2 Reinfections. Journal of Clinical Microbiology. 2021 Mar 19;59(4):e02769-20.

6. Haendel MA, Chute CG, Bennett TD, Eichmann DA, Guinney J, Kibbe WA, et al. The National COVID Cohort Collaborative (N3C): Rationale, design, infrastructure, and deployment. Journal of the American Medical Informatics Association. 2021 Mar 1;28(3):427–43.

7. Latest Phenotype · National-COVID-Cohort-Collaborative/Phenotype_Data_Acquisition Wiki [Internet]. [cited 2023 Apr 25]. Available from: https://github.com/National-COVID-Cohort-Collaborative/Phenotype_Data_Acquisition/wiki/Latest-Phenotype

8. Sidky H, Young JC, Girvin AT, Lee E, Shao YR, Hotaling N, et al. Data quality considerations for evaluating COVID-19 treatments using real world data: learnings from the National COVID Cohort Collaborative (N3C). BMC Medical Research Methodology. 2023 Feb 17;23(1):46.

9. Brannock MD, Chew RF, Preiss AJ, Hadley EC, McMurry JA, Leese PJ, et al. Long COVID Risk and Pre-COVID Vaccination: An EHR-Based Cohort Study from the RECOVER Program. medRxiv. 2022 Oct 7;2022.10.06.22280795.

10. Hernán MA, Robins JM. Using Big Data to Emulate a Target Trial When a Randomized Trial Is Not Available. Am J Epidemiol. 2016 Apr 15;183(8):758–64.

11. Han X, Chen L, Fan Y, Alwalid O, Jia X, Zheng Y, et al. Longitudinal Assessment of Chest CT Findings and Pulmonary Function after COVID-19 Infection. Radiology. 2023 Apr;307(2):e222888.

12. Hall V, Foulkes S, Insalata F, Kirwan P, Saei A, Atti A, et al. Protection against SARS-CoV-2 after Covid-19 Vaccination and Previous Infection. New England Journal of Medicine. 2022 Mar 31;386(13):1207–20.

13. Markov PV, Ghafari M, Beer M, Lythgoe K, Simmonds P, Stilianakis NI, et al. The evolution of SARS-CoV-2. Nat Rev Microbiol. 2023 Jun;21(6):361–79.

14. Abu-Raddad LJ, Chemaitelly H, Bertollini R. Severity of SARS-CoV-2 Reinfections as Compared with Primary Infections. New England Journal of Medicine. 2021 Dec 23;385(26):2487–9.

15. Mensah AA, Lacy J, Stowe J, Seghezzo G, Sachdeva R, Simmons R, et al. Disease severity during SARS-COV-2 reinfection: a nationwide study. Journal of Infection. 2022 Apr 1;84(4):542–50.

16. Bowe B, Xie Y, Al-Aly Z. Acute and postacute sequelae associated with SARS-CoV-2 reinfection. Nat Med. 2022 Nov;28(11):2398–405.

Figure 1: Data flow for the cohort included in the analysis.

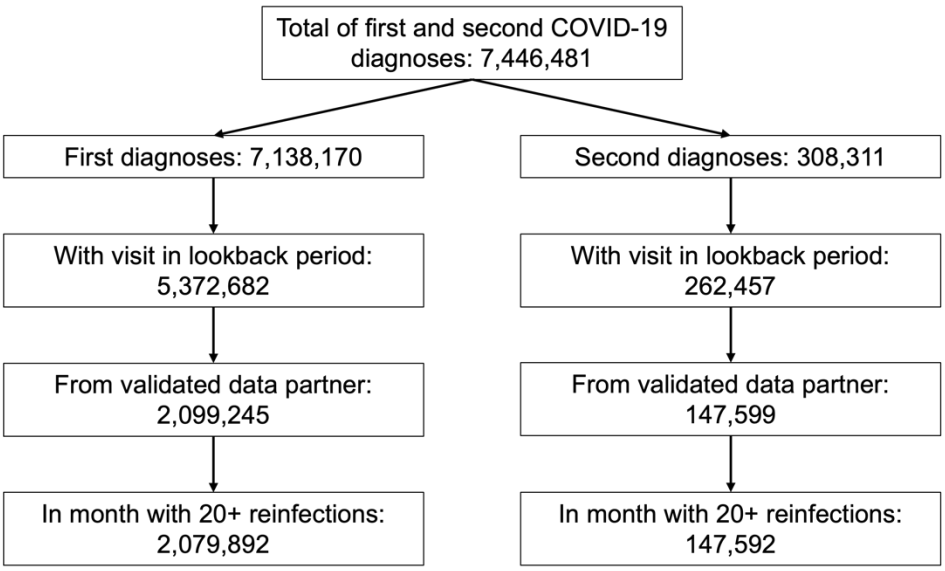

Figure 2: Longitudinal trends among the previously infected cohort in number of vaccines received and time between first and second infection. The shaded area represents the interquartile range of time between first and second infection.

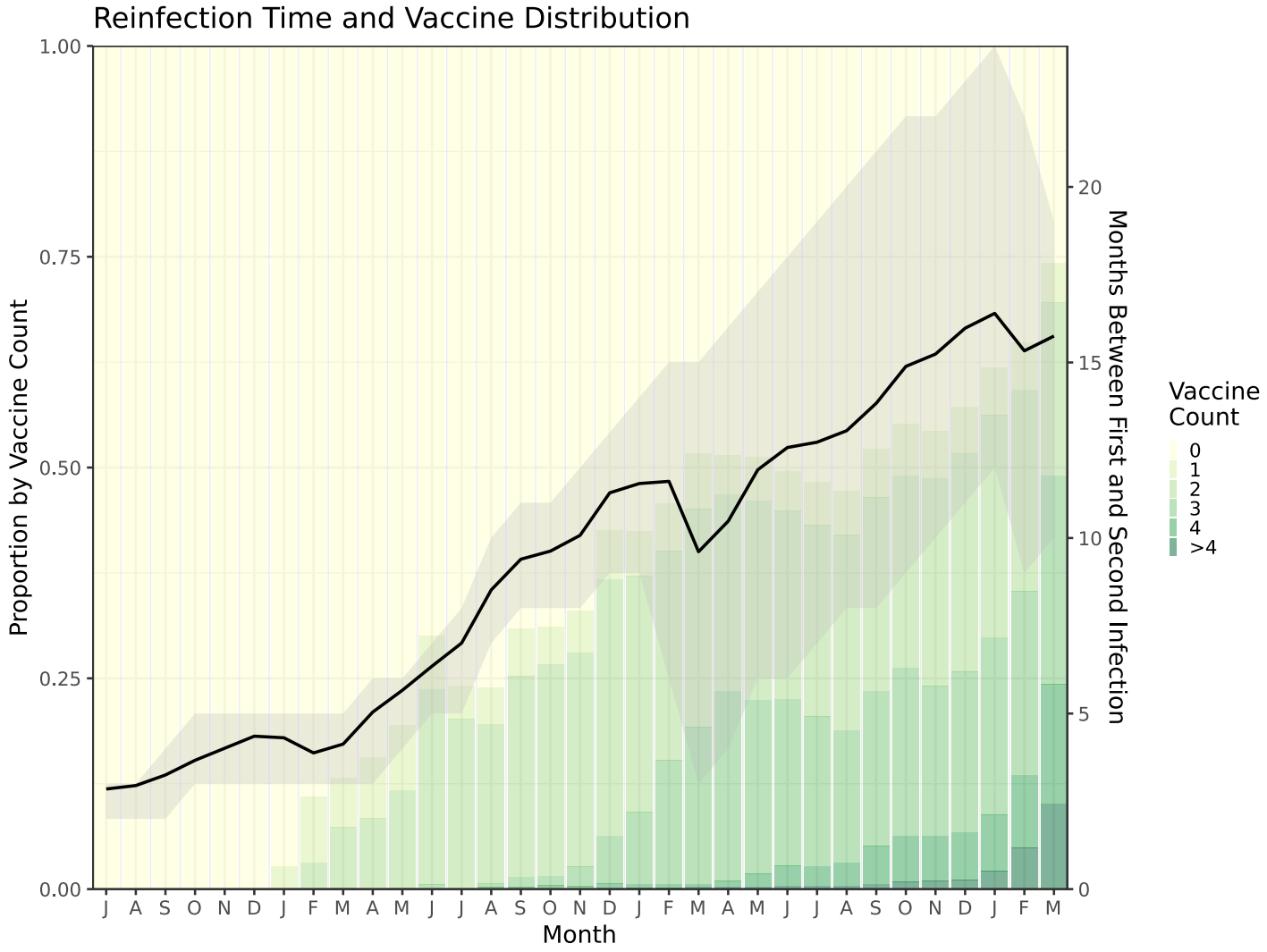

Figure 3: Odds ratio for more severe disease among the full cohort. The solid red line represents the point estimate of the pandemic-wide effect of prior infection, while the dashed red lines represent the 95% confidence interval around that estimate.

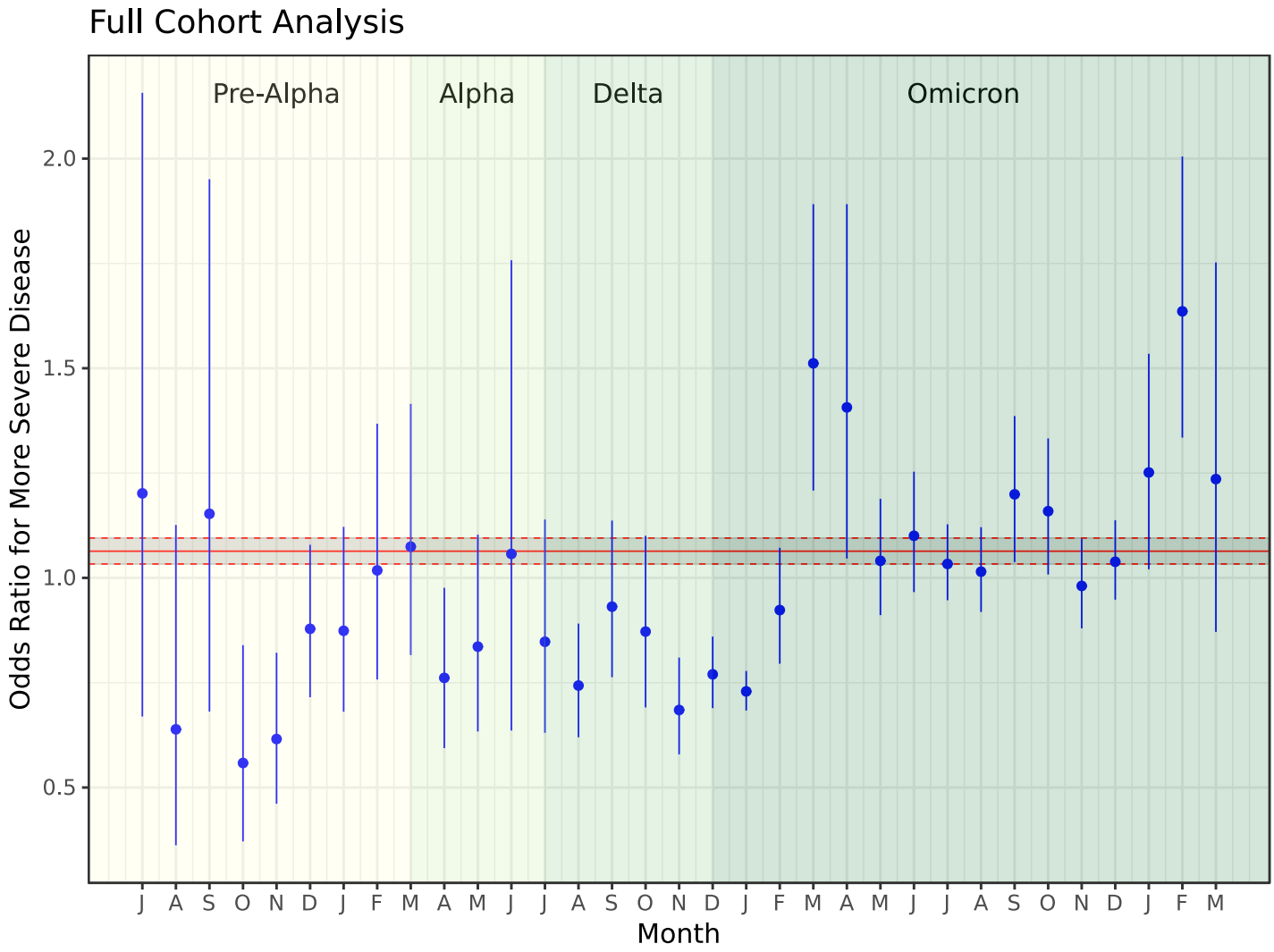

Figure 4: Odds ratios for death (top) and hospitalization (bottom) as binary outcomes using a logistic regression in a sensitivity analysis

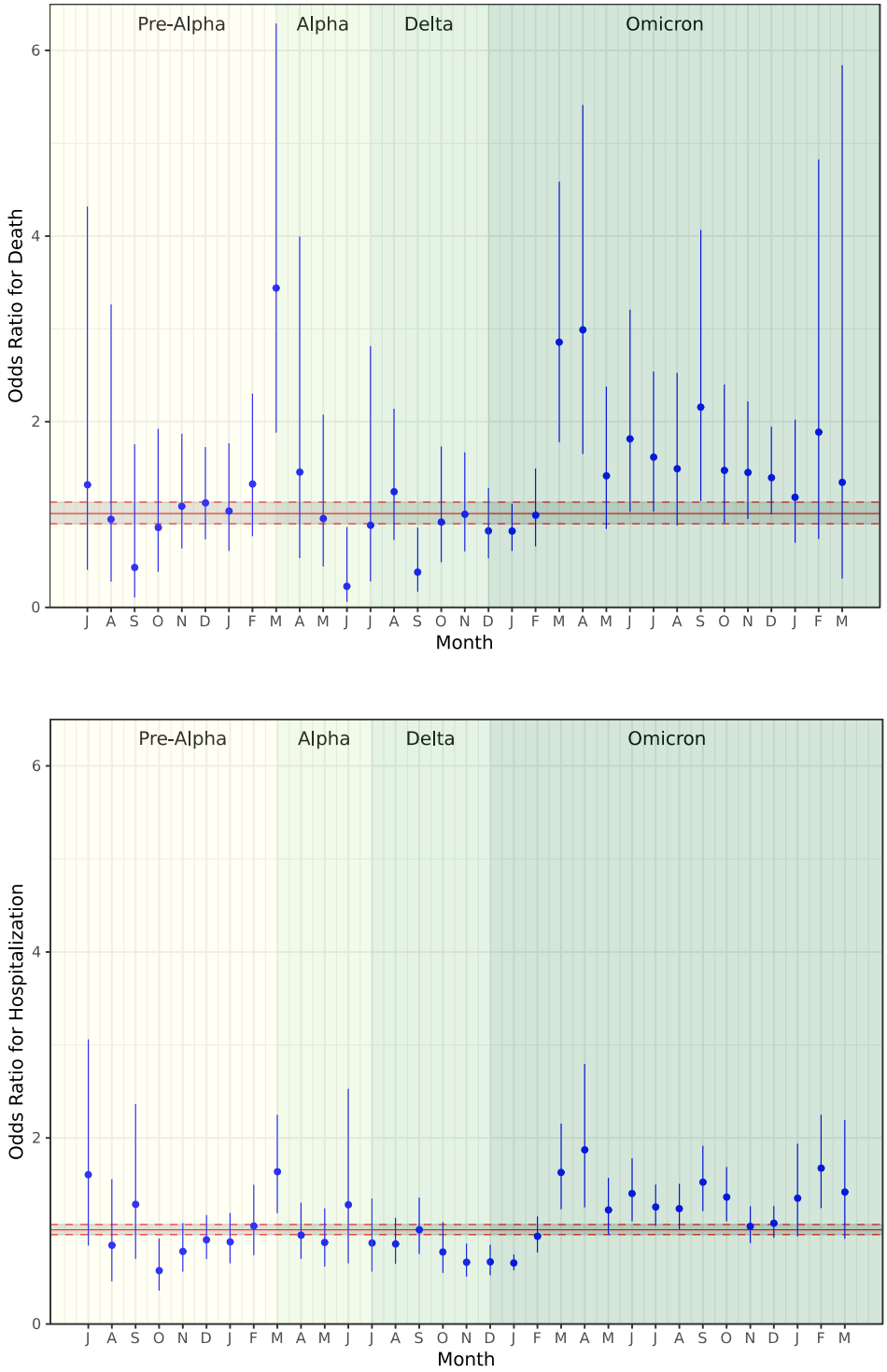

Table 1: Demographics at the time of first and second infection

	First infections (n = 2,079,892)	Second infections (n = 147,592)
Age at infection, mean (SD)	42.1 (22.1)	41.6 (20.2)
Sex: Female, n (%)	1,195,238 (57.5%)	95,961 (65.0%)
Sex: Male, n (%)	882,972 (42.5%)	51,825 (35.0%)
Race / ethnicity: Black, non-Hispanic, n (%)	154,800 (7.4%)	11,626 (7.9%)
Race / ethnicity: Hispanic, any race, n (%)	103,374 (5.1%)	4,493 (3.1%)
Race / ethnicity: White, non-Hispanic, n (%)	1,502,224 (72.2%)	109,979 (74.5%)
Race / ethnicity: Other, non-Hispanic, n (%)	67,483 (3.2%)	2,720 (1.8%)