Exploratory analysis of immunization records highlights decreased SARS-CoV-2 rates in individuals with recent non-COVID-19 vaccinations

Colin Pawlowski1,3, Arjun Puranik1,3, Hari Bandi3, A. J. Venkatakrishnan1, Vineet Agarwal1, Richard Kennedy2, John C. O’Horo2, Gregory J. Gores2, Amy W. Williams2, John Halamka2, Andrew D. Badley2 & Venky Soundararajan1*

Clinical studies are ongoing to assess whether existing vaccines may afford protection against SARS-CoV-2 infection through trained immunity. In this exploratory study, we analyze immunization records from 137,037 individuals who received SARS-CoV-2 PCR tests. We find that polio, Haemophilus influenzae type-B (HIB), measles-mumps-rubella (MMR), Varicella, pneumococcal conjugate (PCV13), Geriatric Flu, and hepatitis A/hepatitis B (HepA–HepB) vaccines administered in the past 1, 2, and 5 years are associated with decreased SARS-CoV-2 infection rates, even after adjusting for geographic SARS-CoV-2 incidence and testing rates, demographics, comorbidities, and number of other vaccinations. Furthermore, age, race/ethnicity, and blood group stratified analyses reveal significantly lower SARS-CoV-2 rate among black individuals who have taken the PCV13 vaccine, with relative risk of 0.45 at the 5 year time horizon (n: 653, 95% CI (0.32, 0.64), p-value: 6.9e−05). Overall, this study identifies existing approved vaccines which can be promising candidates for pre-clinical research and Randomized Clinical Trials towards combating COVID-19.

Since the genome for SARS-CoV-2 was released on January 11, 2020, scientists around the world have been racing to develop a vaccine1. However, vaccine development is a long and expensive process, which takes on average over 10 years under ordinary circumstances2. Even for the previous epidemics of the past decade, including SARS, Zika, and Ebola, vaccines were not available before the virus spread was largely contained3.

Conventionally vaccinations are intended to train the adaptive immune system by generating an antigen-specific immune response. However, studies are also suggesting that certain vaccines lead to protection against other infections through trained immunity for up to 1 year and in the case of live vaccines for up to 5 years4. For instance, vaccination against smallpox showed protection against measles and whooping cough5. Live vaccinia virus was successfully used against smallpox. Due to the urgent need to reduce the spread of COVID-19, scientists are turning to alternate methods to reduce the spread, such as reprogramming existing vaccines. There are some hypotheses that the Bacillus Calmette–Guérin (BCG) and live poliovirus vaccines may provide some protective effect against SARS-CoV-2 infection6–8. There are several ongoing/recruiting clinical trials testing the protective effects of existing vaccines against SARS-CoV-2 infection, including: Polio9, Measles-Mumps-Rubella vaccine10, Influenza vaccine11, and BCG vaccine12–15.

In this work, we conduct a systematic analysis to determine whether or not a set of existing non-COVID-19 vaccines in the United States are associated with decreased rates of SARS-CoV-2 infection. In Fig. 1, we provide an overview of the study design and statistical analyses. We consider data from 137,037 individuals from the Mayo Clinic electronic health record (EHR) database who received PCR tests for SARS-CoV-2 between February 15, 2020 and July 14, 2020 and have at least one ICD diagnostic code recorded in the past five years (see “Methods” section). In Table 1, we show the clinical characteristics of the study population. In particular, 92,673 (67%) individuals have at least 1 vaccine in the past 5 years relative to the PCR testing date. In Fig. 2, we present the SARS-CoV-2 infection rates for subsets of the study population with particular clinical covariates. We note that

1Nference, Inc., One Main Street, Suite 400, East Arcade, Cambridge, MA 02142, USA. 2Mayo Clinic, Rochester, MN, USA. 3These authors contributed equally: Colin Pawlowski and Arjun Puranik. *email: venky@nference.net
the rates of SARS-CoV-2 infection are higher in Black, Asian, and Hispanic racial and ethnic subgroups compared to the overall study population. This is likely due to higher rates of COVID-19 spread and/or decreased access to PCR testing. In addition, the rates of SARS-CoV-2 infection are lower in individuals with pre-existing conditions (e.g., hypertension, diabetes, obesity) possibly due to greater caution in avoiding exposure and/or higher PCR testing rates. Given this study population, we assess the rates of SARS-CoV-2 infection among individuals who did and did not receive one of 18 vaccines in the past 1, 2, and 5 years relative to the date of PCR testing. In Table 2, we present the full names, common formulations, and counts for the 18 vaccines that we consider.

Figure 1. Overview of study design and statistical analyses. (a) Study design, datasets, and inclusion criteria used for the study; (b) comparisons of SARS-CoV-2 rates between propensity-matched vaccinated and unvaccinated cohorts in the overall study population; (c) comparisons of SARS-CoV-2 rates between propensity-matched vaccinated and unvaccinated cohorts in subgroups of the population stratified by age, race/ethnicity, and blood type.
| Clinical characteristics | Count (proportion) |
|--------------------------|--------------------|
| Total number of individuals | 137,037 |
| **County-level COVID-19** | |
| Incidence rate (± 1 week from PCR date) | 0.0014 |
| Test positive rate (± 1 week from PCR date) | 5.1% |
| **Age** | |
| 0–18 | 10,855 (7.9%) |
| 19–49 | 52,179 (38%) |
| 50–64 | 33,297 (24%) |
| 65+ | 40,706 (30%) |
| **Gender** | |
| Male | 60,712 (44%) |
| Female | 76,308 (56%) |
| **Race** | |
| White | 119,979 (88%) |
| Black | 5473 (4%) |
| Asian | 3267 (2.4%) |
| Other | 8318 (6.1%) |
| **Ethnicity** | |
| Hispanic | 7720 (5.6%) |
| Not hispanic | 124,877 (91%) |
| Unknown | 4440 (3.2%) |
| **Body Mass Index** | |
| Underweight (<18.5) | 2340 (1.7%) |
| Normal (18.5 to 25.0) | 35,183 (26%) |
| Overweight (25.0 to 30.0) | 36,655 (27%) |
| Obese (>30.0) | 45,860 (33%) |
| Unknown | 16,999 (12%) |
| **Number of individuals with at least 1 recorded vaccine** | |
| Over past 1 year | 61,209 (45%) |
| Over past 2 years | 74,923 (55%) |
| Over past 5 years | 92,278 (67%) |
| Lifetime | 106,420 (78%) |
| **Elixhauser comorbidities in the past 5 years** | |
| Hypertension | 47,767 (35%) |
| Arrhythmias | 39,423 (29%) |
| Depression | 34,687 (25%) |
| Obesity | 33,334 (24%) |
| Pulmonary disease | 30,135 (22%) |
| Fluid and electrolyte disorders | 23,778 (17%) |
| Hypothyroidism | 20,284 (15%) |
| Peripheral vascular disorders | 18,848 (14%) |
| Valvular disease | 16,985 (12%) |
| Renal | 14,876 (11%) |
| Tumor (solid, without metastasis) | 13,533 (9.9%) |
| Liver disease | 12,756 (9.3%) |
| Congestive heart failure | 11,841 (8.6%) |
| Neurodegenerative disorders | 11,804 (8.6%) |
| Diabetes (complicated) | 11,801 (8.6%) |
| Anemia | 11,491 (8.4%) |
| Rheumatic diseases | 11,086 (8.1%) |
| Weight loss | 9487 (6.9%) |
| Coagulopathy | 9134 (6.7%) |
| Alcohol | 7401 (5.4%) |
| Metastatic cancer | 7048 (5.1%) |
| Drug abuse | 6969 (5.1%) |

Continued
In Table 3, we provide a breakdown of patient counts by vaccine type (inactivated, live attenuated, recombinant, unknown) for each of these 18 vaccines.

Given this dataset, first we assess the overall association of vaccination status with the risk of SARS-CoV-2 infection (see "Methods" section). We use propensity score matching to construct unvaccinated control groups for each of the vaccinated populations at the 1 year, 2 year, and 5-year time horizons. The unvaccinated control groups are balanced in covariates including demographics, county-level incidence and testing rates for SARS-CoV-2, comorbidities, and number of other vaccines taken in the past 5 years. Then, we compare the SARS-CoV-2 rates between each of the vaccinated cohorts and corresponding matched, unvaccinated control groups which have similar clinical characteristics. Second, we repeat the analysis on a set of age, race, ethnicity, Elixhauser comorbidities, and number of unique vaccines taken.

In Table 3, we provide a breakdown of patient counts by vaccine type (inactivated, live attenuated, recombinant, unknown) for each of these 18 vaccines.

Table 1. General characteristics of study population. Descriptive statistics for the study population, including: County-level COVID-19 incidence and testing rates, demographics (age, gender, race, ethnicity), vaccine counts, and Elixhauser comorbidities.

| Clinical characteristics | Count (proportion) |
|--------------------------|--------------------|
| Pulmonary hypertension   | 6607 (4.8%)        |
| Diabetes (uncomplicated) | 6529 (4.8%)        |
| Peptic ulcer disease     | 3442 (2.5%)        |
| Lymphoma                 | 3128 (2.3%)        |
| Blood loss               | 2413 (1.8%)        |
| Paralysis                | 1719 (1.3%)        |
| Psychosis                | 1394 (1%)          |
| HIV/AIDS                 | 192 (0.14%)        |

Figure 2. SARS-CoV-2 infection risk ratios by clinical covariate. SARS-CoV-2 rates among individuals with particular clinical covariates along with 95% confidence intervals. A dotted line indicates the study population SARS-CoV-2 rate of 4.4%. The clinical covariates include: county-level COVID-19 incidence and testing rates, age brackets (< 18, 18–49, 50–64, 65+ years), gender, race, ethnicity, Elixhauser comorbidities, and number of unique vaccines taken.
Results
Polio, HIB, MMR, Varicella, PCV13, Geriatric Flu, and HepA–HepB vaccines consistently show associations with lower rates of SARS-CoV-2 infection across 1, 2, and 5-year time horizons. The results of the propensity score matching for the 1 year, 2 year, and 5-year time horizons are presented in Tables 4, 5 and 6, respectively. We observe that across all time horizons, Polio, Hemophilus Influenzae type B (HIB), Pneumococcal conjugate (PCV13), Geriatric Flu, Hepatitis A/Hepatitis B (HepA–HepB), and Hepatitis A/Hepatitis B (HepA–HepB) vaccines consistently show associations with lower rates of SARS-CoV-2 infection across 1, 2, and 5-year time horizons.

Table 2. Summary of vaccines and common formulations. The 18 vaccines taken by at least 1000 individuals within 5 years prior to their PCR test date, along with the most common formulations and patient counts. Note that some common formulations are combinations of multiple vaccines (e.g. Pentacel is a combination of DPT, polio, and HIB vaccines).

| Vaccine name                          | Common formulations                        | Number of individuals taking in the past 5 years |
|---------------------------------------|--------------------------------------------|-------------------------------------------------|
| Diphtheria–Pertussis–Tetanus (DPT)   | TDAP; TD preservative free                  | 47,949                                          |
| Geriatric Flu                         | High dose geriatric (65 + years)           | 22,259                                          |
| Haemophilus Influenzae type B (HIB)   | DTAP-IPV/HIB (PENTACEL)                    | 4534                                            |
| Human Papillomavirus (HPV)            | 9VHPV, 4VHPV                               | 6179                                            |
| Hepatitis A/Hepatitis B (HepA–HepB)  | HepA adult; HepB adult                     | 15,392                                          |
| Influenza (general)                   | Flublok/Fluarix/Fluzone                    | 77,890                                          |
| Influenza (live)                      | Influenza LAIV (Nasal)                     | 2296                                            |
| Measles–mumps–rubella (MMR)          | MMR, MMRV                                  | 6480                                            |
| Meningococcal                         | MCV4, MENB                                 | 7008                                            |
| Polio                                 | DTAP-IPV/HIB (PENTACEL), IPV, DTAP-IPV     | 5793                                            |
| Pediatric Flu                         | H1V4                                       | 11,676                                          |
| Pneumococcal conjugate (PCV13)        | PCV13                                      | 25,634                                          |
| Pneumococcal polysaccharide (PPSV23)  | PPSV23                                     | 16,846                                          |
| Rotavirus                             | RV5 (Rotateq)                              | 3235                                            |
| RZV Zoster (Zostavax, Shingrix)       | Zostavax, Shingrix                         | 16,897                                          |
| Tetanus                               | Td                                         | 2801                                            |
| Typhoid                               | TyVi, Ty21a                                | 2384                                            |
| Varicella                             | VAR, MMRV                                  | 5497                                            |

Table 3. Patient counts by vaccination type. Breakdown of patient counts by vaccine type for each of the 18 vaccines taken by at least 1000 individuals within 5 years prior to their PCR test date. In particular, counts and percentages are shown for the following vaccine types: (1) Inactivated, (2) Live attenuated, (3) Recombinant, and (4) Unknown.

| Vaccine name                          | Total | Inactivated | Live attenuated | Recombinant | Unknown |
|---------------------------------------|-------|-------------|-----------------|-------------|---------|
| Diphtheria–Pertussis–Tetanus (DPT)   | 47,949| 37,256 (78%)| 0 (0%)          | 0 (0%)      | 10,693 (22%) |
| Geriatric Flu                         | 22,259| 21,979 (99%)| 0 (0%)          | 0 (0%)      | 280 (1.3%)  |
| Hepatitis A/Hepatitis B (HepA–HepB)  | 15,392| 10,138 (66%)| 0 (0%)          | 0 (0%)      | 5254 (34%)  |
| Haemophilus Influenzae type B (HIB)   | 4534  | 4495 (99%)  | 0 (0%)          | 0 (0%)      | 39 (0.86%)  |
| Human Papillomavirus (HPV)            | 6179  | 0 (0%)      | 0 (0%)          | 6143 (99%)  | 36 (0.58%)  |
| Influenza (general)                   | 77,890| 57,620 (74%)| 2296 (2.9%)     | 0 (0%)      | 17,974 (23%)|
| Influenza (live)                      | 2296  | 0 (0%)      | 2296 (100%)     | 0 (0%)      | 0 (0%)      |
| Meningococcal                         | 7008  | 6289 (90%)  | 0 (0%)          | 0 (0%)      | 719 (10%)   |
| Measles–mumps–rubella (MMR)          | 6480  | 0 (0%)      | 6282 (97%)      | 0 (0%)      | 198 (3.1%)  |
| Pediatric Flu                         | 11,676| 11,621 (100%)| 0 (0%)          | 0 (0%)      | 55 (0.47%)  |
| Pneumococcal (PPSV23)                | 16,846| 16,378 (97%)| 0 (0%)          | 0 (0%)      | 468 (2.8%)  |
| Pneumococcal conjugate (PCV13)       | 25,634| 24,407 (95%)| 0 (0%)          | 0 (0%)      | 1227 (4.8%) |
| Polio                                 | 5793  | 5559 (96%)  | 3 (0.03%)       | 0 (0%)      | 231 (4%)    |
| Rotavirus                             | 3235  | 0 (0%)      | 3213 (99%)      | 0 (0%)      | 22 (0.68%)  |
| RZV Zoster (Zostavax, Shingrix)       | 16,897| 0 (0%)      | 3017 (18%)      | 10,551 (62%)| 3329 (20%)  |
| Tetanus                               | 2801  | 2600 (93%)  | 0 (0%)          | 0 (0%)      | 201 (7.2%)  |
| Typhoid                               | 2384  | 1686 (71%)  | 185 (7.8%)      | 0 (0%)      | 513 (22%)   |
| Varicella                             | 5497  | 0 (0%)      | 5346 (97%)      | 0 (0%)      | 151 (2.7%)  |
### Table 4. Summary of SARS-CoV-2 rates for vaccinated and unvaccinated propensity score matched cohorts (1 year time horizon).

Table of SARS-CoV-2 infection rates for vaccinated and unvaccinated (matched) cohorts for vaccines administered within 1 year prior to PCR testing. Rows in which the SARS-CoV-2 rate is lower (adjusted p-value < 0.05) in the vaccinated cohort are highlighted in bold. The columns are (1) Vaccine: Name of the vaccine, (2) Total matched pairs: Number of pairs from the propensity matching procedure, which is the sample size of both vaccinated and unvaccinated cohorts after matching, (3) Vaccinated (matched) COVIDpos: Number of COVIDpos cases among the vaccinated (matched) cohort, along with the percentage in parentheses, (4) Unvaccinated (matched) COVIDpos: Number of COVIDpos cases among the unvaccinated (matched) cohort, along with the percentage in parentheses, (5) Relative risk (95% CI): Relative risk of COVIDpos in the vaccinated (matched) cohort compared to the unvaccinated (matched) cohort, along with 95% confidence interval in parentheses, (6) BH-adjusted p-value: Benjamini–Hochberg-adjusted Fisher exact test p-value.

| Vaccine                          | Total matched pairs | Vaccinated (matched) COVIDpos | Unvaccinated (matched) COVIDpos | Relative risk (95% CI) | BH-adjusted p-value |
|----------------------------------|---------------------|-------------------------------|---------------------------------|------------------------|---------------------|
| POLIO                            | 2402                | 64 (2.66%)                    | 113 (4.7%)                      | 0.57 (0.42, 0.77)      | 3.1E−03             |
| HIB                              | 2061                | 43 (2.09%)                    | 81 (3.93%)                      | 0.53 (0.37, 0.77)      | 3.2E−03             |
| MMR                              | 1700                | 53 (3.12%)                    | 94 (5.53%)                      | 0.56 (0.41, 0.79)      | 3.2E−03             |
| Geriatric Flu vaccine (65+ years)| 12,085              | 190 (1.57%)                   | 257 (2.13%)                     | 0.74 (0.61, 0.89)      | 5.6E−03             |
| Influenza (any)                  | 12,791              | 442 (3.46%)                   | 521 (4.07%)                     | 0.85 (0.75, 0.96)      | 0.03                |
| Pneumococcal conjugate (PCV13)   | 4693                | 102 (2.17%)                   | 142 (3.03%)                     | 0.72 (0.56, 0.92)      | 0.03                |
| VARICELLA                        | 1416                | 39 (2.75%)                    | 63 (4.45%)                      | 0.62 (0.42, 0.92)      | 0.04                |
| HepA–HepB                        | 5858                | 189 (3.23%)                   | 235 (4.01%)                     | 0.80 (0.67, 0.97)      | 0.05                |
| Meningococcal                    | 1456                | 96 (6.59%)                    | 73 (5.01%)                      | 1.32 (0.98, 1.76)      | 0.12                |
| Diphtheria (with P/T)            | 12,020              | 423 (3.52%)                   | 474 (3.94%)                     | 0.89 (0.78, 1.01)      | 0.12                |
| RZV Zoster (ZOSTAVAX, SHINGRIX)  | 9381                | 209 (2.23%)                   | 230 (2.45%)                     | 0.91 (0.76, 1.09)      | 0.43                |
| HPV                              | 1467                | 91 (6.2%)                     | 79 (5.39%)                      | 1.15 (0.86, 1.54)      | 0.45                |
| Pneumococcal (PPSV23)            | 4636                | 112 (2.42%)                   | 106 (2.29%)                     | 1.06 (0.81, 1.37)      | 0.79                |

### Table 5. Summary of SARS-CoV-2 rates for vaccinated and unvaccinated propensity score matched cohorts (2 year time horizon).

Table of SARS-CoV-2 infection rates for vaccinated and unvaccinated (matched) cohorts for vaccines administered within 2 years prior to PCR testing. Rows in which the SARS-CoV-2 rate is lower (adjusted p-value < 0.05) in the vaccinated cohort are highlighted in bold, and rows in which the SARS-CoV-2 rate is lower in the unvaccinated cohort are highlighted in italic. The columns are (1) Vaccine: Name of the vaccine, (2) Total matched pairs: Number of pairs from the propensity matching procedure, which is the sample size of both vaccinated and unvaccinated cohorts after matching, (3) Vaccinated (matched) COVIDpos: Number of COVIDpos cases among the vaccinated (matched) cohort, along with the percentage in parentheses, (4) Unvaccinated (matched) COVIDpos: Number of COVIDpos cases among the unvaccinated (matched) cohort, along with the percentage in parentheses, (5) Relative risk (95% CI): Relative risk of COVIDpos in the vaccinated (matched) cohort compared to the unvaccinated (matched) cohort, along with 95% confidence interval in parentheses, (6) BH-adjusted p-value: Benjamini–Hochberg-adjusted Fisher exact test p-value.

| Vaccine                          | Total matched pairs | Vaccinated (matched) COVIDpos | Unvaccinated (matched) COVIDpos | Relative risk (95% CI) | BH-adjusted p-value |
|----------------------------------|---------------------|-------------------------------|---------------------------------|------------------------|---------------------|
| POLIO                            | 2821                | 88 (3.12%)                    | 173 (6.13%)                     | 0.51 (0.40, 0.66)      | 1.2E−06             |
| HepA–HepB                        | 8443                | 265 (3.14%)                   | 379 (4.49%)                     | 0.70 (0.60, 0.82)      | 4.0E−05             |
| Pneumococcal conjugate (PCV13)   | 8285                | 185 (2.23%)                   | 275 (3.32%)                     | 0.67 (0.56, 0.81)      | 9.5E−05             |
| HIB                              | 2711                | 56 (2.07%)                    | 110 (4.06%)                     | 0.51 (0.37, 0.70)      | 9.5E−05             |
| Meningococcal                    | 3009                | 223 (7.41%)                   | 155 (5.15%)                     | 1.14 (1.18, 1.75)      | 1.1E−03             |
| Typhoid                          | 1051                | 72 (6.85%)                    | 36 (3.43%)                      | 2.00 (1.35, 2.93)      | 1.2E−03             |
| Varicella                        | 2544                | 85 (3.34%)                    | 135 (5.31%)                     | 0.63 (0.48, 0.82)      | 1.5E−03             |
| MMR                              | 3055                | 111 (3.63%)                   | 161 (5.27%)                     | 0.69 (0.55, 0.87)      | 4.3E−03             |
| RZV Zoster (ZOSTAVAX, SHINGRIX)  | 14,000              | 290 (2.07%)                   | 360 (2.57%)                     | 0.81 (0.69, 0.94)      | 0.01                |
| Pneumococcal (PPSV23)            | 8751                | 210 (2.4%)                    | 265 (3.03%)                     | 0.79 (0.66, 0.95)      | 0.02                |
| Geriatric Flu vaccine (65+ years)| 12,360              | 217 (1.76%)                   | 269 (2.18%)                     | 0.81 (0.68, 0.96)      | 0.03                |
| Diphtheria (with P/T)            | 21,705              | 805 (3.71%)                   | 879 (4.05%)                     | 0.92 (0.83, 1.01)      | 0.09                |
| Influenza (any)                  | 17,652              | 665 (3.77%)                   | 723 (4.1%)                      | 0.92 (0.83, 1.02)      | 0.14                |
| HPV                              | 2654                | 177 (6.72%)                   | 168 (6.38%)                     | 1.05 (0.86, 1.29)      | 0.70                |
| ROTAVIRUS                        | 612                 | 10 (1.63%)                    | 8 (1.31%)                       | 1.25 (0.50, 3.03)      | 0.81                |
### Table 6. Summary of SARS-CoV-2 rates for vaccinated and unvaccinated propensity score matched cohorts (5 year time horizon).

| Vaccine                          | Total matched pairs | Vaccinated (matched) COVID<sub>pos</sub> | Unvaccinated (matched) COVID<sub>pos</sub> | Relative risk (95% CI) | BH-adjusted p-value |
|----------------------------------|---------------------|-----------------------------------------|-------------------------------------------|------------------------|---------------------|
| Pneumococcal conjugate (PCV13)  | 23,194              | 503 (2.17%)                             | 745 (3.21%)                               | 0.68 (0.60, 0.76)      | 7.2E−11             |
| Meningococcal                   | 7008                | 552 (7.88%)                             | 366 (5.22%)                               | 1.51 (1.33, 1.71)      | 2.1E−09             |
| POLIO                            | 3072                | 131 (4.26%)                             | 213 (6.93%)                               | 0.62 (0.50, 0.76)      | 3.8E−05             |
| HPV                              | 6179                | 494 (7.99%)                             | 376 (6.09%)                               | 1.31 (1.13, 1.48)      | 1.7E−04             |
| Geriatric Flu vaccine (65+ years)| 13,860              | 246 (1.77%)                             | 330 (2.38%)                               | 0.75 (0.63, 0.88)      | 1.7E−03             |
| Hib                              | 2913                | 73 (2.51%)                              | 120 (4.12%)                               | 0.61 (0.46, 0.81)      | 2.2E−03             |
| MMR                              | 4965                | 226 (4.55%)                             | 299 (6.02%)                               | 0.76 (0.64, 0.89)      | 3.1E−03             |
| RZV Zoster (ZOSTAVAX, SHINGRIX)  | 16,889              | 355 (2.1%)                              | 440 (2.61%)                               | 0.81 (0.70, 0.93)      | 5.7E−03             |
| Typhoid                          | 2383                | 149 (6.25%)                             | 103 (4.32%)                               | 1.45 (1.13, 1.84)      | 7.0E−03             |
| HepA–HepB                        | 13,377              | 541 (4.04%)                             | 628 (4.69%)                               | 0.86 (0.77, 0.96)      | 0.02                |
| Varicella                        | 3623                | 175 (4.83%)                             | 218 (6.02%)                               | 0.80 (0.66, 0.97)      | 0.05                |
| Pediatric Flu vaccine            | 11,676              | 426 (3.65%)                             | 375 (3.21%)                               | 1.14 (0.99, 1.30)      | 0.11                |
| TETANUS                          | 2800                | 88 (3.14%)                              | 67 (2.39%)                                | 1.31 (0.96, 1.79)      | 0.14                |
| Influenza (Live)                 | 2296                | 110 (4.79%)                             | 87 (3.79%)                                | 1.26 (0.94, 1.66)      | 0.14                |
| Diphtheria (with P/T)            | 40,334              | 1590 (3.94%)                            | 1678 (4.16%)                              | 0.95 (0.89, 1.01)      | 0.14                |
| Pneumococcal (PPSV23)            | 16,836              | 414 (2.46%)                             | 446 (2.65%)                               | 0.93 (0.81, 1.06)      | 0.32                |
| Influenza (any)                  | 22,057              | 985 (4.47%)                             | 955 (4.33%)                               | 1.03 (0.95, 1.13)      | 0.53                |
| ROTAVIRUS                        | 694                 | 18 (2.59%)                              | 23 (3.31%)                                | 0.78 (0.43, 1.43)      | 0.53                |

### Pairwise correlation analysis reveals strong associations between administration of Hib, Polio, Rotavirus, Varicella, and MMR vaccines.

In order to identify vaccines which may be confounding factors for other vaccines that are linked to reduced rates of SARS-CoV-2 infection, we conduct a pairwise correlation analysis. For example, it is possible that the lower rates of SARS-CoV-2 infection that we observe for one vaccine are in fact caused by another vaccine which is highly correlated with the former. To measure the correlations we use Cohen's kappa, which is a measure of correlation for categorical variables that ranges from −1 to +1. In particular, Cohen’s kappa = +1 indicates that the pair of vaccines are always administered together, Cohen’s kappa = 0 indicates that the pair of vaccines are independent of each other, and Cohen’s kappa = −1 indicates that the pair of vaccines are never administered together.
In Fig. 5, we present a heatmap of the pairwise correlations for each of the 18 vaccines administered in the 5 years prior to the PCR test date. Sorted by Cohen's kappa value, the top vaccine pairs with kappa ≥ 0.60 are: HIB and Rotavirus (0.83), HIB and Polio (0.80), MMR and Varicella (0.74), Polio and Varicella (0.72), Polio and Rotavirus (0.71), MMR and Polio (0.68). From this, we see that there is a cluster of vaccines which are commonly administered together, which includes: HIB, Polio, Rotavirus, Varicella, and MMR vaccines. The majority of individuals who receive this cluster of vaccines are children < 18 years old (see Fig. 4). We note that in this cluster, the vaccines HIB, Polio, Varicella, and MMR are all consistently associated with lower SARS-CoV-2 rates. This suggests that some of the lower rates of SARS-CoV-2 observed in these vaccinated cohorts may be confounded by the other vaccines in this group.
Stratification by race reveals that Polio, HIB, and PCV13 vaccines are associated with lower SARS-CoV-2 rates in particular racial subgroups across 1, 2, and 5-year time periods. In Tables 7, 8 and 9, we present the results of propensity score matching at the 1, 2, and 5-year time horizon, respectively, on study cohorts stratified by race. We observe that PCV13 vaccination is linked with significantly decreased SARS-CoV-2 rates in the Black subpopulation. In particular, the relative risk of SARS-CoV-2 infection for black individuals who have been administered PCV13 is 0.24 at the 1 year time horizon (n: 197, 95% CI (0.09, 0.71), p-value: 0.03); 0.33 at the 2 year time horizon (n: 239, 95% CI (0.16, 0.74), p-value: 0.03); and 0.45 at the 5 year time horizon (n: 653, 95% CI (0.32, 0.64), p-value: 6.9e−5). Furthermore, at the 5-year time horizon, the relative risk for the PCV13 vaccinated cohort of black individuals is significantly lower than the relative risk for the PCV13 vaccinated cohort overall (p-value: 0.03).

Figure 4. Age distribution plots for vaccinated cohorts. Distributions of age in cohorts of individuals who received vaccines associated with lower SARS-CoV-2 rates. For each vaccine, percentages of vaccinated individuals within age brackets (<18, 18–49, 50–64, 65+ years) are shown along with 95% confidence intervals. Includes age distributions for the following vaccines for the past 1, 2, and 5 year time horizons: (A–C) Geriatric Flu vaccine, (D–F) Hepatitis A/Hepatitis B (HepA-HepB), (G–I) Haemophilus Influenzae type B (HIB), (J–L) Measles-Mumps-Rubella (MMR), (M–O) Pneumococcal Conjugate (PCV13), (P–R) Polio, (S–U) Varicella.
In addition, we observe that Polio, HIB, and PCV13 vaccines are linked with decreased SARS-CoV-2 rates in the White subpopulation. However, since 119,979 (88%) of individuals in the study population are white, the relative risks for these vaccinated cohorts are close to the relative risks for the overall population (see Tables 4, 5, 6). Matching within subgroups was done by age group (0–18, 19–49, 50–64, 65 +) and blood group (A, B, AB, O) as well, but no significant within-subgroup associations between any vaccine and SARS-CoV-2 rates were found. This suggests that associations between vaccines and SARS-CoV-2 infection rates may not be strongly specific to particular age ranges/blood groups.

Hospitalization and ICU rates of COVID‑19 patients are similar among vaccinated and non‑vaccinated cohorts. In Tables 10 and 11, we show the COVID-19 hospitalization and ICU rates among the vaccinated and non-vaccinated (matched) cohorts for the 1-year time horizon, respectively. We observe that these rates are relatively similar between the two cohorts, and there are no statistically significant differences. This lack of a statistically significant association may be due to the relatively low rates of hospitalization/ICU admission among COVID-19 patients in this dataset. Considering these findings along with the results from the previous analysis, these results suggest that vaccination status is associated with differential rates of SARS-CoV-2 infection, but there is not enough evidence to determine if vaccination status is associated with COVID-19 disease severity.

Sensitivity analysis. Tipping point analysis shows that associations between reduced SARS-CoV-2 rates and Polio vaccine (1, 2 year time horizons), PCV13 (3 year time horizon) are most robust to unobserved confounders. In this retrospective study, we evaluate the correlations between vaccination and SARS-CoV-2 infection, taking into account a number of possible confounding variables, such as demographic variables and geographic COVID-19 incidence rate (see “Methods” section). However, it is possible that the results from this study have been influenced by unobserved confounders. For example, we do not explicitly control for travel history, which was a significant risk factor for SARS-CoV-2 infection early on in the pandemic.

In Fig. 6, we present the results from the tipping point analysis on the statistically significant associations between vaccination and reduced rates of SARS-CoV-2 infection in the overall study population. For each time horizon, we show the relative prevalence and effect size that would be required for an unobserved confounder to overturn the conclusion for a given (vaccine, time horizon) pair. For reference, we show the effect size of the
Table 7. Summary of SARS-CoV-2 rates for race/ethnicity-stratified vaccinated and unvaccinated propensity score matched cohorts (1 year time horizon). Table of SARS-CoV-2 infection rates for vaccinated and unvaccinated (matched) race/ethnicity subgroup cohorts for vaccines administered within 1 year prior to PCR testing. Only rows with adjusted p-values ≤ 0.1 are included. Rows in which the SARS-CoV-2 rate is lower (adjusted p-value < 0.05) in the vaccinated cohort are highlighted in bold. The columns are (1) Vaccine: Name of the vaccine, (2) Race/ethnicity: Race/ethnicity subgroup, (3) Total matched pairs: Number of pairs from the propensity matching procedure, which is the sample size of both vaccinated and unvaccinated cohorts after matching, (4) Vaccinated (matched) COVIDpos: Number of COVIDpos cases among the vaccinated (matched) cohort, along with the percentage in parentheses, (5) Unvaccinated (matched) COVIDpos: Number of COVIDpos cases among the unvaccinated (matched) cohort, along with the percentage in parentheses, (6) Relative risk (95% CI): Relative risk of COVIDpos in the vaccinated (matched) cohort compared to the unvaccinated (matched) cohort, along with 95% confidence interval in parentheses, (7) BH-adjusted p-value: Benjamini–Hochberg-adjusted Fisher exact test p-value.

| Vaccine                                              | Race/ethnicity | Total matched pairs | Vaccinated (matched) COVIDpos | Unvaccinated (matched) COVIDpos | Relative risk (95% CI) | BH-adjusted p-value |
|------------------------------------------------------|----------------|---------------------|------------------------------|-------------------------------|------------------------|----------------------|
| Geriatric Flu vaccine (65+ years)                    | White          | 13,021              | 178 (1.37%)                  | 256 (1.97%)                  | 0.70 (0.58, 0.84)      | 4.8E−03              |
| POLIO                                                | Black          | 117                 | 5 (4.27%)                    | 24 (20.5%)                   | 0.21 (0.09, 0.55)      | 4.8E−03              |
| HIB                                                  | White          | 1744                | 20 (1.15%)                   | 49 (2.81%)                   | 0.41 (0.25, 0.69)      | 7.7E−03              |
| RZV Zoster (ZOSTAVAX, SHINGRIX)                      | Asian          | 179                 | 5 (2.79%)                    | 21 (11.7%)                   | 0.24 (0.10, 0.64)      | 0.02                 |
| POLIO                                                | White          | 2033                | 28 (1.38%)                   | 57 (2.8%)                    | 0.49 (0.32, 0.77)      | 0.02                 |
| Pneumococcal conjugate (PCV13)                       | White          | 4116                | 68 (1.65%)                   | 105 (2.55%)                  | 0.65 (0.48, 0.88)      | 0.03                 |
| Pneumococcal conjugate (PCV13)                       | Black          | 197                 | 4 (2.03%)                    | 17 (8.63%)                   | 0.24 (0.09, 0.71)      | 0.03                 |
| RZV Zoster (ZOSTAVAX, SHINGRIX)                      | Black          | 221                 | 10 (4.52%)                   | 24 (10.9%)                   | 0.42 (0.21, 0.86)      | 0.09                 |
| Influenza (any)                                      | White          | 11,731              | 298 (2.54%)                  | 357 (3.04%)                  | 0.83 (0.72, 0.97)      | 0.09                 |
| Pneumococcal conjugate (PCV13)                       | Hispanic       | 301                 | 19 (6.31%)                   | 36 (12%)                     | 0.53 (0.32, 0.90)      | 0.09                 |

At the 1 year and 2 year time horizons, the associations of the Polio vaccine to lower rates of SARS-CoV-2 infection are most robust to the impact of a potential unobserved confounder. In particular, an unobserved confounder with a large effect size of 2.78 would need to have an absolute difference in prevalence between vaccinated and unvaccinated cohorts of 17.8% (30.9%) in order to overturn the results for the 1 year (2 year) time horizon. On the other hand, at the 5 year time horizon, the association of PCV13 and lower rates of SARS-CoV-2 infection is most robust to the influence by unobserved confounders. An unobserved confounder with a large effect size of 2.78 would need to have an absolute difference in prevalence between vaccinated and unvaccinated cohorts of 19.1% in order to render the findings insignificant.

**Discussion**

Ongoing clinical studies offer preliminary evidence that existing vaccines may reduce risk of SARS-CoV-2 infection. For example, interim results from the ACTIVATE trial indicate that the BCG vaccine reduces SARS-CoV-2 infection rates up to 53%. While specific vaccines such as BCG are being tested for cross-protective effects against SARS-CoV-2 infection based upon their prior potential for protection against other diseases, to our knowledge, a systematic hypothesis-free analysis to identify potential vaccines that can have beneficial effects against SARS-CoV-2 infection is lacking. Our retrospective study has analyzed 18 different vaccines and identified key vaccines that are associated with lower rates of SARS-CoV-2 infection after controlling for confounding factors (see "Results" section). In particular, we find that individuals who have been recently vaccinated with one of Polio, HIB, MMR, Varicella, PCV13, Geriatric Flu, or HepA–HepB vaccines have lower rates of SARS-CoV-2 infection. These vaccines are promising candidates for follow-up pre-clinical animal studies and clinical trials in the COVID-19. For the rest of the 18 vaccines that we considered, the correlations with SARS-CoV-2 infection were either insignificant or varied across the time horizons of interest. In some cases, these vaccines may serve as negative controls in clinical trials testing the safety and efficacy of novel COVID-19 vaccines. For example, a clinical trial evaluating the COVID-19 vaccine candidate ChAdOx1 uses Meningococcal vaccine as a comparator arm. In this case, Meningococcal vaccine was used as a control instead of the typical saline solution in order to reduce the risk of unblinding, because viral vector vaccinations are known to be associated with certain typical adverse reactions. Preliminary results from this trial indicate that as expected, Meningococcal vaccine does not induce antibody responses against SARS-CoV-2 spike protein. It may be interesting to evaluate the antibody responses for some of the vaccines that we have found to be significantly correlated with lower rates of SARS-CoV-2 infection, to explore if there is any underlying immunologic mechanism for the associations that we observe.
Table 8. Summary of SARS-CoV-2 rates for race/ethnicity-stratified vaccinated and unvaccinated propensity score matched cohorts (2 year time horizon). Table of SARS-CoV-2 infection rates for vaccinated and unvaccinated (matched) race/ethnicity subgroup cohorts for vaccines administered within 2 years prior to PCR testing. Only rows with adjusted p-values ≤ 0.1 are included. Rows in which the SARS-CoV-2 rate is lower (adjusted p-value < 0.05) in the vaccinated cohort are highlighted in bold. The columns are (1) Vaccine: Name of the vaccine, (2) Race/ethnicity: Race/ethnicity subgroup, (3) Total matched pairs: Number of pairs from the propensity matching procedure, which is the sample size of both vaccinated and unvaccinated cohorts after matching, (4) Vaccinated (matched) COVIDpos: Number of COVIDpos cases among the vaccinated (matched) cohort, along with the percentage in parentheses, (5) Unvaccinated (matched) COVIDpos: Number of COVIDpos cases among the unvaccinated (matched) cohort, along with the percentage in parentheses, (6) Relative risk (95% CI): Relative risk of COVIDpos in the vaccinated (matched) cohort compared to the unvaccinated (matched) cohort, along with 95% confidence interval in parentheses, (7) BH-adjusted p-value: Benjamini-Hochberg-adjusted Fisher exact test p-value.

Because the BCG vaccine is rarely administered in the US, this vaccine did not meet the sample size threshold for inclusion in our analysis. From the limited data available, there were 51 individuals in the study population who had taken BCG vaccine in the past 5 years, and among these 0 individuals tested positive for SARS-CoV-2 infection (95% CI (0.0%, 7.0%)). Among the 198 individuals who had taken BCG vaccine at least once in their lifetime, there were 6 (3.0%) individuals who tested positive for SARS-CoV-2 infection (95% CI (1.4%, 6.5%)). We note that no individuals in the study population received BCG over the 1-year time horizon, and only 1 over the 2-year time horizon. As a result, more data from additional medical centers would be required for us to assess the associations between BCG vaccine and SARS-CoV-2 infection.

There are prior studies highlighting mechanisms of activation of broad immune signalling pathways by vaccines, which might also be providing protection against SARS-CoV-2. This nonspecific innate response conferring protection to other infections is termed as ‘trained immunity’17. For example, in the case of tuberculosis vaccine—Bacillus Calmette-Guerin (BCG) induces immune response against micro-organisms beyond Mycobacterium tuberculosis, such as Candida albicans and Staphylococcus aureus18,19. There is also evidence of epigenetic histone modifications observed in the monocytes/macrophages promoting the expression of pattern-recognition molecules upon stimulation through BCG20,21. Recently there have been a number of studies systematically exploring the effect of BCG vaccine in treating COVID-19 patients22,23,24. In the case of Haemophilus influenzae type-B, the activation of complement system by Haemophilus influenzae type-B is well studied25 and recently there has been a report on decreased complement C3 levels being associated with poor prognosis in patients with COVID-1926. The complement C3 inhibitor AMY-101 is currently in phase-2 clinical trial for treatment of COVID-1927. Here, the cross-protection provided through Haemophilus influenzae type-B vaccine could potentially be mediated through regulation of the immune complement system. In the case of MMR, engineered live measles vaccine has previously been suggested to confer protection from SARS-CoV-2 in animal models28. At a molecular level IFNAR2 deficiency is reported to cause hemophagocytic lymphohistiocytosis (HLH) following measles-mumps-rubella vaccination because of excessive IFN signalling. Although there are reports of SARS-CoV-2 inhibiting the production of IFNβ, externally administered interferons are observed to block the replication of viruses. Thus, interferon signalling indirectly mediated through MMR vaccine could potentially contribute to cross-protection towards SARS-CoV-2. Overall, there are interesting hypotheses around trained immunity from pre-existing vaccines having a potential effect against SARS-CoV-2 and further studies to investigate these are warranted.

Due to the observational nature of this study, there are potential biases which may have impacted the findings, including confounding, selection bias, and measurement bias. The motivation for using propensity score matching was to account for confounding. Although we take into account some potential confounders through propensity score matching, there may still be residual confounding from unobserved factors (e.g. socioeconomic conditions).
| Vaccine                          | Race/ Ethnicity | Total matched pairs | Vaccinated (matched) COVID<sub=pos</sub> | Unvaccinated (matched) COVID<sub=pos</sub> | Relative risk (95% CI) | BH-adjusted p-value |
|---------------------------------|----------------|--------------------|-----------------------------------------|------------------------------------------|------------------------|---------------------|
| Meningococcal                   | White          | 5772               | 335 (5.8%)                              | 202 (3.5%)                              | 1.66 (1.40, 1.96)     | 2.9E−07             |
| Pneumococcal conjugate (PCV13)  | White          | 23,706             | 413 (1.74%)                             | 581 (2.45%)                             | 0.71 (0.63, 0.81)     | 2.0E−06             |
| POLIO                           | White          | 3321               | 66 (1.99%)                              | 142 (4.28%)                             | 0.46 (0.35, 0.62)     | 2.0E−06             |
| Meningococcal                   | Black          | 460                | 123 (26.7%)                             | 61 (13.3%)                              | 2.02 (1.52, 2.65)     | 6.4E−06             |
| Pneumococcal conjugate (PCV13)  | Black          | 653                | 41 (6.28%)                              | 91 (13.9%)                              | 0.45 (0.32, 0.64)     | 6.9E−05             |
| MMR                             | White          | 5285               | 130 (2.46%)                             | 201 (3.8%)                              | 0.65 (0.52, 0.80)     | 9.2E−04             |
| RZV Zoster (ZOSTAVAX, SHINGRIX) | Black          | 359                | 15 (4.18%)                              | 44 (12.3%)                              | 0.34 (0.20, 0.61)     | 9.7E−04             |
| HIB                             | Black          | 170                | 10 (5.88%)                              | 31 (18.2%)                              | 0.32 (0.17, 0.65)     | 5.5E−03             |
| Pediatric Flu vaccine           | Black          | 517                | 100 (19.3%)                             | 62 (12%)                                | 1.61 (1.20, 2.15)     | 0.01                |
| Typhoid                         | Black          | 268                | 71 (26.5%)                              | 42 (15.7%)                              | 1.69 (1.20, 2.36)     | 0.02                |
| Diptheria (with P/T)            | White          | 38,816             | 1153 (2.97%)                            | 1298 (3.34%)                            | 0.89 (0.82, 0.96)     | 0.02                |
| VARICELLA                       | White          | 4456               | 117 (2.63%)                             | 163 (3.66%)                             | 0.72 (0.57, 0.91)     | 0.03                |
| HIB                             | White          | 3731               | 62 (1.66%)                              | 97 (2.6%)                               | 0.64 (0.47, 0.88)     | 0.03                |
| HepA–HepB                       | White          | 12,999             | 356 (2.74%)                             | 432 (3.32%)                             | 0.82 (0.72, 0.95)     | 0.03                |
| Geriatric Flu vaccine (65+ years)| Black         | 312                | 19 (6.09%)                              | 39 (12.5%)                              | 0.49 (0.29, 0.83)     | 0.03                |
| HPV                             | Black          | 354                | 84 (23.7%)                              | 56 (15.8%)                              | 1.50 (1.10, 2.02)     | 0.04                |
| Pediatric Flu vaccine           | Asian          | 463                | 35 (7.56%)                              | 17 (3.67%)                              | 2.06 (1.16, 3.54)     | 0.05                |
| Geriatric Flu vaccine (65+ years)| White        | 14,410             | 226 (1.57%)                             | 281 (1.95%)                             | 0.80 (0.68, 0.96)     | 0.05                |
| Influenza (any)                 | Black          | 1181               | 171 (14.5%)                             | 132 (11.2%)                             | 1.30 (1.05, 1.60)     | 0.06                |
| Pneumococcal conjugate (PCV13)  | Hispanic       | 897                | 66 (7.36%)                              | 93 (10.4%)                              | 0.71 (0.53, 0.96)     | 0.10                |

Table 9. Summary of SARS-CoV-2 rates for race/ethnicity-stratified vaccinated and unvaccinated propensity score matched cohorts (5 year time horizon). Table of SARS-CoV-2 infection rates for vaccinated and unvaccinated (matched) race/ethnicity subgroup cohorts for vaccines administered within 5 years prior to PCR testing. Only rows with adjusted p-values ≤ 0.1 are included. Rows in which the SARS-CoV-2 rate is lower (adjusted p-value < 0.05) in the vaccinated cohort are highlighted in bold, and rows in which the SARS-CoV-2 rate is lower in the unvaccinated cohort are highlighted in italics. The columns are (1) Vaccine: Name of the vaccine, (2) Race/ethnicity: Race/ethnicity subgroup, (3) Total matched pairs: Number of pairs from the propensity matching procedure, which is the sample size of both vaccinated and unvaccinated cohorts after matching, (4) Vaccinated (matched) COVID<sub=pos</sub>: Number of COVID<sub=pos</sub> cases among the vaccinated (matched) cohort, along with the percentage in parentheses, (5) Unvaccinated (matched) COVID<sub=pos</sub>: Number of COVID<sub=pos</sub> cases among the unvaccinated (matched) cohort, along with the percentage in parentheses, (6) Relative risk (95% CI): Relative risk of COVID<sub=pos</sub> in the vaccinated (matched) cohort compared to the unvaccinated (matched) cohort, along with 95% confidence interval in parentheses, (7) BH-adjusted p-value: Benjamini-Hochberg-adjusted Fisher exact test p-value.

status, adherence to social distancing measures, use of personal protective equipment etc.) which may be different for each vaccine. For example, travel history is a risk factor for exposure to SARS-CoV-2 infection that we do not explicitly account for in this study. Our motivation for the tipping point sensitivity analysis is to estimate the effect size and prevalence of an unobserved confounder which would be required to overturn the statistically significant findings (see Fig. 6). Even among the variables that we consider, there is potential for bias if the cohorts are poorly matched on those covariates. In Tables S1–S7, we present the propensity score matching results for a number of vaccines at the 1 year time horizon, in order to show the matching quality for each of these statistical comparisons. Furthermore, we present plots showing the distribution of the age covariate in particular in Fig. S1. We note that for some vaccines, differences in age between the vaccinated and unvaccinated (matched) cohorts may have influenced the results.

In addition, it is possible that restricting the study population to SARS-CoV-2 PCR tested individuals may have introduced selection bias. For example, vaccinated individuals may engage in more health-seeking behaviors to reduce their potential COVID-19 risk, and also have a higher likelihood of seeking out a PCR test. This type of bias is known as the “healthy user effect”, which is suspected to have influenced the findings of recent COVID-19 observational studies. We performed sensitivity analyses using breast cancer and colon cancer screening as negative controls which suggest that the propensity score matching analysis is in part effective in filtering out healthy user effect for the associations between vaccination status and SARS-CoV-2 risk. Finally, measurement bias is a concern as vaccination records may be incomplete for some individuals in our cohort since they may have received the vaccines outside of the Mayo Clinic system. We plan to perform additional sensitivity analyses to further explore these potential sources of bias.
Table 10. Summary of COVID-19 hospitalization rates for vaccinated and unvaccinated propensity score matched cohorts (1 year time horizon). Table of COVID-19 hospitalization rates for vaccinated and unvaccinated (matched) cohorts for vaccines administered within 1 years prior to PCR testing. The columns are (1) Vaccine: Name of the vaccine, (2) Vaccinated (matched) COVIDpos Hospitalized: Number of patients who were COVID-positive among the vaccinated (matched) cohort and had hospitalization status data, (3) Unvaccinated (matched) COVIDpos Hospitalized: Number of patients hospitalized for COVID-19 among the vaccinated (matched) cohort, along with the percentage (taken over the vaccinated COVIDpos patients) in parentheses, (4) Unvaccinated (matched) COVIDpos: Number of patients who were COVID-positive in the unvaccinated (matched) cohort and had hospitalization data, (5) Unvaccinated (matched) COVIDpos Hospitalized: Number of patients hospitalized for COVID-19 among the unvaccinated (matched) cohort, along with the percentage (taken over unvaccinated COVIDpos patients) in parentheses, (6) Relative risk (95% CI): Relative risk of COVID-19-related hospitalization in the vaccinated (matched) cohort compared to the unvaccinated (matched) cohort, along with 95% confidence interval in parentheses, (7) Fisher exact p-value.

| Vaccine                          | Vaccinated (matched) COVIDpos | Vaccinated (matched) COVIDpos hospitalized | Unvaccinated (matched) COVIDpos | Unvaccinated (matched) COVIDpos hospitalized | Relative risk (95% CI) | Fisher exact p-value |
|---------------------------------|-------------------------------|------------------------------------------|-------------------------------|------------------------------------------|------------------------|---------------------|
| Pneumococcal (PPSV23)          | 111                           | 35 (32%)                                 | 106                           | 25 (24%)                                 | 1.3 (0.86, 2.1)        | 0.23                |
| RZV Zoster (ZOSTA-VAX, SHINGRIX) | 209                           | 51 (24%)                                 | 230                           | 68 (30%)                                 | 0.83 (0.61, 1.1)       | 0.24                |
| Hib                             | 43                            | 3 (7%)                                   | 81                            | 11 (14%)                                 | 0.51 (0.18, 1.8)       | 0.38                |
| HepA-HepB                       | 189                           | 31 (16%)                                 | 235                           | 46 (20%)                                 | 0.84 (0.56, 1.3)       | 0.45                |
| Influenza (any)                 | 441                           | 74 (17%)                                 | 518                           | 78 (15%)                                 | 1.1 (0.83, 1.5)        | 0.48                |
| Pneumococcal conjugate (PCV13) | 102                           | 19 (19%)                                 | 140                           | 21 (15%)                                 | 1.2 (0.71, 2.2)        | 0.49                |
| Varicella                       | 39                            | 2 (5.1%)                                 | 62                            | 1 (1.6%)                                 | 3.2 (0.36, 19)         | 0.56                |
| Meningococcal                   | 95                            | 8 (8.4%)                                 | 73                            | 8 (11%)                                  | 0.77 (0.31, 1.9)       | 0.61                |
| Diphtheria (with P/T)           | 421                           | 65 (15%)                                 | 471                           | 67 (14%)                                 | 1.1 (0.79, 1.5)        | 0.64                |
| POLIO                           | 64                            | 3 (4.7%)                                 | 113                           | 8 (7.1%)                                 | 0.66 (0.22, 2.4)       | 0.75                |
| Geriatric Flu vaccine (65+ years) | 189                        | 82 (43%)                                 | 257                           | 110 (43%)                                | 1 (0.82, 1.3)          | 0.92                |
| HPV                             | 91                            | 3 (3.3%)                                 | 78                            | 3 (3.8%)                                 | 0.86 (0.2, 3.7)        | 1.00                |
| MMR                             | 53                            | 3 (5.7%)                                 | 94                            | 5 (5.3%)                                 | 1.1 (0.31, 4.1)        | 1.00                |

As an initial exploratory analysis linking historical vaccination records to SARS-CoV-2 PCR testing results, more research is warranted in order to confirm the findings. We plan to update this analysis in coming months as more PCR testing data becomes available. Also, we note that this study is based on data from one academic medical center in the United States, which restricts the analysis to vaccines administered in this geographic region. Notably, we do not have sufficient immunization record data on the BCG vaccine, which has shown promise in early clinical trials. As a result, the findings from this study would be well complemented by similar studies from hospitals across the world.

Methods

Study design. This is an observational study in a cohort of individuals who underwent polymerase chain reaction (PCR) testing for suspected SARS-CoV-2 infection at the Mayo Clinic and hospitals affiliated to the Mayo health system. The full dataset includes 152,548 individuals who received PCR tests between February 15, 2020 and July 14, 2020. We restricted the study population to 137,037 individuals from this dataset who have at least one ICD code recorded in the past 5 years. This exclusion criteria is applied in order to restrict the analysis to individuals with medical history data. Within this PCR tested cohort, we define COVIDpos to be persons with at least one positive PCR test result for SARS-CoV-2 infection, which includes 5679 individuals. Similarly, we define COVIDneg to be persons with all negative PCR test results, which includes 131,358 individuals.

For the study population of 137,037 individuals, we obtain a number of clinical covariates from the Mayo Clinic electronic health record (EHR) database, including: demographics (age, gender, race, ethnicity, county of residence), ICD diagnostic billing codes from the past 5 years, and immunization records from the past 5 years (68 unique vaccines; we focus on the 18 taken by at least 1000 individuals over the past 5 years). We use the Elixhauser Comorbidity Index to map the ICD codes from each individual from the past 5 years to a set of 30 medically relevant comorbidities.[19] In addition to the Mayo Clinic EHR database, we use the Corona Data Scraper online database to obtain incidence rates of COVID-19 at the county-level in the United States.[19,20] By linking the county of residence data from the EHR with the incidence rates of COVID-19 from Corona Data Scraper, we are able to obtain county-level incidence rates of COVID-19 for 136,313 individuals in the study population.

We also obtain county-level testing data for 100,433 individuals in the study population from (i) Minnesota state government records and (ii) public county-level testing data scraped from other state/county websites. In Table 1, we present the average values for each of the clinical covariates in the study population.

Given these clinical covariates, we conduct a series of statistical analyses to assess whether or not each of the 19 vaccines has an association with lower rates of SARS-CoV-2 infection at the 1 year, 2 years, and 5 year time...
For each vaccine and time horizon, the vaccinated cohort is defined as the set of individuals in the study population who received the vaccine within the past time horizon. For example, the "2-year polio vaccinated cohort" is the set of individuals who received the polio vaccine within the past 2 years. Similarly, for each vaccine and time horizon, the unvaccinated cohort is defined as the set of individuals in the study population who did not receive the vaccine within the past time horizon. For example, the "5-year influenza unvaccinated cohort" is the set of individuals who did not receive the influenza vaccine within the past 5 years.

In the following sections, we describe the statistical methods that we use to compare the rates of COVID-19 between the vaccinated and unvaccinated cohorts for each of the (vaccine, time horizon) pairs. First, we describe the propensity score matching analysis to construct unvaccinated control groups that have similar clinical characteristics. Second, we describe the statistical tests that we use to determine which of the (vaccine, time horizon) pairs have the most significant association with lower rates of SARS-CoV-2 infection for the 1 year, 2 year, and 5 year time horizons, both overall and for particular demographic subgroups. Third, we describe the covariate-level stratification analysis to identify vaccines which have the largest association with lower rates of SARS-CoV-2 infection for particular demographic subgroups. Finally, we describe the sensitivity analyses that we use to evaluate the robustness of the statistical methods to potential biases from unobserved confounders or other factors that could impact the overall results from this observational study.

Propensity score matching to construct unvaccinated control groups. Before running the propensity score matching step, first we filtered to vaccinated cohorts with at least 1000 persons. For the overall statistical analysis, there were 13, 15, and 18 vaccines which met this threshold for the 1 year, 2 year, and 5 year time horizons, respectively.

For each vaccine and time horizon, the vaccinated cohort is defined as the set of individuals in the study population who received the vaccine within the past time horizon. For example, the "2-year polio vaccinated cohort" is the set of individuals who received the polio vaccine within the past 2 years. Similarly, for each vaccine and time horizon, the unvaccinated cohort is defined as the set of individuals in the study population who did not receive the vaccine within the past time horizon. For example, the "5-year influenza unvaccinated cohort" is the set of individuals who did not receive the influenza vaccine within the past 5 years.

In the following sections, we describe the statistical methods that we use to compare the rates of COVID-19 between the vaccinated and unvaccinated cohorts for each of the (vaccine, time horizon) pairs. First, we describe the propensity score matching analysis to construct unvaccinated control groups that have similar clinical characteristics. Second, we describe the statistical tests that we use to determine which of the (vaccine, time horizon) pairs have the most significant association with lower rates of SARS-CoV-2 infection for the 1 year, 2 year, and 5 year time horizons, both overall and for particular demographic subgroups. Third, we describe the covariate-level stratification analysis to identify vaccines which have the largest association with lower rates of SARS-CoV-2 infection for particular demographic subgroups. Finally, we describe the sensitivity analyses that we use to evaluate the robustness of the statistical methods to potential biases from unobserved confounders or other factors that could impact the overall results from this observational study.

| Vaccine                          | Vaccinated (matched) COVIDpos | Vaccinated (matched) COVIDpos ICU | Unvaccinated (matched) COVIDpos | Unvaccinated (matched) COVIDpos ICU | Relative risk (95% CI) | Fisher exact p-value |
|----------------------------------|-------------------------------|----------------------------------|--------------------------------|-------------------------------------|------------------------|---------------------|
| Meningococcal                   | 95                            | 1 (1.1%)                         | 73                             | 4 (5.5%)                            | 0.19 (0.042, 1.6)     | 0.17                |
| Diphtheria (with P/T)           | 421                           | 7 (1.7%)                         | 471                            | 15 (3.2%)                           | 0.52 (0.23, 1.3)      | 0.19                |
| Hib                              | 43                            | 0 (0%)                           | 81                             | 2 (2.5%)                            | 0 (0.018, 7.6)        | 0.54                |
| Polio                            | 64                            | 0 (0%)                           | 113                            | 3 (2.7%)                            | 0 (0.013, 4.8)        | 0.55                |
| Pneumococcal conjugate (PCV13)  | 102                           | 6 (5.9%)                         | 140                            | 6 (4.3%)                            | 1.4 (0.48, 3.9)       | 0.57                |
| RZV Zoster (ZOSTAVAX, SHINGRIX)  | 209                           | 13 (6.2%)                        | 230                            | 12 (5.2%)                           | 1.2 (0.56, 2.5)       | 0.68                |
| HepA-HepB                        | 189                           | 7 (3.7%)                         | 235                            | 10 (4.3%)                           | 0.87 (0.35, 2.2)      | 0.81                |
| Influenza (any)                  | 441                           | 15 (3.4%)                        | 518                            | 16 (3.1%)                           | 1.1 (0.56, 2.2)       | 0.86                |
| Geriatric Flu vaccine (65+ years)| 189                           | 20 (11%)                         | 257                            | 28 (11%)                            | 0.97 (0.57, 1.7)      | 1.00                |
| Pneumococcal (PPSV23)            | 111                           | 10 (9%)                          | 106                            | 9 (8.5%)                            | 1.1 (0.46, 2.4)       | 1.00                |
| HPV                              | 91                            | 0 (0%)                           | 78                             | 0 (0%)                              | 1.00                  |                     |
| MMR                              | 53                            | 0 (0%)                           | 94                             | 0 (0%)                              | 1.00                  |                     |
| Varicella                        | 39                            | 0 (0%)                           | 62                             | 0 (0%)                              | 1.00                  |                     |

Table 11. Summary of COVID-19 ICU rates for vaccinated and unvaccinated propensity score matched cohorts (1 year time horizon). Table of COVID-19 ICU rates for vaccinated and unvaccinated (matched) cohorts for vaccines administered within 1 years prior to PCR testing. The columns are (1) Vaccine: Name of the vaccine, (2) Vaccinated (matched) COVIDpos: Number of patients who were COVID-positive among the vaccinated (matched) cohort and had ICU data, (3) Vaccinated (matched) COVIDpos ICU: Number of patients admitted to the ICU for COVID-19 among the vaccinated (matched) cohort, along with the percentage (taken over the vaccinated COVIDpos patients) in parentheses, (4) Unvaccinated (matched) COVIDpos: Number of patients who were COVID-positive among the unvaccinated (matched) cohort and had ICU data, (5) Unvaccinated (matched) COVIDpos ICU: Number of patients admitted to the ICU for COVID-19 among the unvaccinated (matched) cohort, along with the percentage (taken over unvaccinated COVIDpos patients) in parentheses, (6) Relative risk (95% CI): Relative risk of COVID-19-related ICU admission in the vaccinated (matched) cohort compared to the unvaccinated (matched) cohort, along with 95% confidence interval in parentheses, (7) Fisher exact p-value.
Figure 6. Sensitivity of associations between vaccines and SARS-CoV-2 rates to unobserved confounders. Tipping point analysis for associations of vaccines and lower rates of SARS-CoV-2 infection for (A) 1-year, (B) 2-year, and (C) 5-year time horizons. For each vaccine that is associated with lower SARS-CoV-2 rates in a particular time horizon, we plot the (prevalence, effect size) combinations of an unobserved confounder that would be required to overturn the results. The x-axis indicates the absolute difference in prevalence of the confounder between vaccinated and unvaccinated (matched) cohorts. For example, if the unobserved confounder is present in 25% of the vaccinated cohort and 5% of the unvaccinated cohort, then the absolute difference in prevalence would be 20%. The y-axis indicates the relative COVID pos risk (effect size) of the unobserved confounder. For reference, we show the relative risk of (county-level COVID-19 incidence rate ≥ median value) as a horizontal dotted line, which is equal to 2.78. Each plot is annotated with the top 3 vaccines that are most robust to unobserved confounders, along with the intersection point between the vaccine curve and the reference line. For example, for the polio vaccine at the 1 year time horizon, an unobserved confounder with a relative risk of 2.78 which is prevalent in 17.8% of the vaccinated cohort and 0% of the unvaccinated cohort could explain the differences in SARS-CoV-2 infection rates that we observe in the data.
County-level COVID-19 test positive rate (Number of positive SARS-CoV-2 PCR tests in county)/(Number of PCR tests in county) within ± 1 week of PCR testing date.

Elixhauser comorbidities Medical history derived from ICD diagnostic billing codes in the past 5 years relative to the PCR testing date. Includes indicators for the following conditions: (1) congestive heart failure, (2) cardiac arrhythmias, (3) valvular disease, (4) pulmonary circulation disorders, (5) peripheral vascular disorders, (6) hypertension, (7) paralysis, (8) neurodegenerative disorders, (9) chronic pulmonary disease, (10) diabetes, (11) diabetes with complications, (12) hypothyroidism, (13) renal failure, (14) liver disease, (15) peptic ulcer disease (excluding bleeding), (16) AIDS/HIV, (17) lymphoma, (18) metastatic cancer, (19) solid tumor without metastasis, (20) rheumatoid arthritis/collagen vascular diseases, (21) coagulopathy, (22) obesity, (23) weight loss, (24) fluid and electrolyte disorders, (25) blood loss anemia, (26) deficiency anemia, (27) alcohol abuse, (28) drug abuse, (29) psychoses, (30) depression.

Pregnancy Whether or not the individual had a pregnancy-related ICD code recorded in the past 90 days relative to the PCR testing date.

Number of other vaccines Count of the total number of unique vaccines (excluding the vaccine which is the treatment variable) taken by the individual in the past 5 years relative to the PCR testing date.

For each of the vaccinated cohorts, we fit a logistic regression model to predict whether or not the individual was vaccinated, using these covariates as predictors. We trained the logistic regression model using the scikit-learn package in Python. Then, we used the model-predicted probability of an individual receiving the vaccine as the propensity score for the individual. Matching was done without replacement using greedy nearest-neighbor matching within calipers. Some subjects were dropped from the positive cohort in this procedure. The matching was performed with caliper width 0.2 × (pooled standard deviation of scores), as suggested in the literature.

Statistical assessment of associations between vaccines and SARS-CoV-2 infection rates for the overall study population. After the propensity score matching step, we compare the COVIDpos rates for the vaccinated and unvaccinated (matched) cohorts. First, we compute the relative risk, which is equal to the COVIDpos rate for the vaccinated (matched) cohort divided by the COVIDpos rate for the unvaccinated (matched) cohort. We use a Fisher exact test to compute the p-value for this association. We then apply the Benjamini-Hochberg (BH) adjustment on the p-values over all vaccines for each time horizon to control the False Discovery Rate (at 0.05). We also compute and report 95% confidence intervals for the relative risks.

Statistical assessment of associations between vaccines and SARS-CoV-2 infection rates for age, race/ethnicity, and blood type stratified subgroups. We repeat the statistical analysis on subsets of the study population stratified by age, race/ethnicity, and blood type. For age, we consider the subgroups: 0 to 18 years, 19 to 49 years old, 50 to 64 years old, and ≥ 65 years old. For race/ethnicity, we consider the subgroups: White, Black, Asian, and Hispanic. For blood type, we consider the subgroups: O, A, B, and AB. We note that age and race/ethnicity were recorded in the dataset for all subjects, but blood type information was only available for 41,828 subjects.

For each vaccine, at the 1, 2, and 5 year time horizons, we use propensity score matching to construct unvaccinated control groups for each age bracket, race/ethnicity, and blood type subgroup. Matching was done on the same covariates as in the overall analysis (apart from the Race/Ethnicity covariates for the race/ethnicity subgroups). We then compared the COVIDpos rates between the vaccinated and unvaccinated (matched) cohorts, and reported the relative risk, 95% confidence interval, and BH-corrected p-values.

Sensitivity analyses. We performed two sets of sensitivity analyses, as described below.

Cancer screens as negative controls for propensity score matching procedure. To assess the effectiveness of the propensity score matching procedure, we ran the statistical analysis using cancer screens as the exposure variable instead of vaccinations (i.e. negative control exposure). This set of experiments serves as a negative control because it is highly unlikely that cancer screenings are causally linked to risk of SARS-CoV-2 infection. In particular, we considered the following two cancer screens as negative controls:

- Colon cancer screen Whether or not the individual received a screening for colon cancer (within a specified time horizon relative to PCR testing date).
- Mammogram Whether or not the individual received a mammogram screening for breast cancer (within a specified time horizon relative to PCR testing date).

In Table 12, we present the results from the negative control experiments. In the unmatched cohorts, we observe that persons who have had a mammogram in the past 1, 2, or 5 years have significantly lower rates of SARS-CoV-2 infection compared to persons who have not had mammograms during the same time period. For example, the SARS-CoV-2 infection rate is 2.5% among persons with mammograms in the past 5 years and 4.5% among persons without mammograms in the past 5 years (p-value: 1.9e−47). This significant difference in SARS-CoV-2 infection rate can be explained by confounding variables, because the unmatched cohorts have different underlying clinical characteristics. However, after propensity score matching, the SARS-CoV-2 infection rate is 2.8% among persons with mammograms in the past 5 years and 2.8% among persons without mammograms in the past 5 years (p-value: 1).
We observe similar results for the colon cancer screening covariate. For example, the SARS-CoV-2 infection rate is 2.5% among persons with colon cancer screens in the past 5 years and 4.4% among persons without colon cancer screens in the past 5 years (p-value: 9.3e−44). After propensity score matching, the SARS-CoV-2 infection rate is 2.5% with and 2.4% without colon cancer screens in the past 5 years (p-value: 1). In total, 6 comparisons (2 controls, 3 time horizons each) were done. After applying Fisher’s method to combine p-values, we get a combined p-value of 0.22 (X² = 15, df = 12) against the combined hypothesis that none of the controls have an association with SARS-CoV-2 after propensity score matching.

We expect that the individuals who have recently taken cancer screens may have lower rates of SARS-CoV-2 infection due to the “healthy user effect.” In particular, persons who have recently had mammograms or colonoscopies may engage in general health-seeking behaviors which decrease their risk of SARS-CoV-2 infection or generally decrease their risk of a positive PCR test result. The results from the negative control experiment demonstrates that the propensity score matching is able to correct for confounding variables which may contribute to spurious findings such as those caused by the healthy user effect.

Tipping point analysis. In order to evaluate how robust the associations between vaccinations and SARS-CoV-2 infection found in this study are to the effects of potential confounders, we conduct a “tipping point” analysis. The purpose of this analysis is to find the point at which an unobserved confounder would “tip” the conclusion on each vaccine, making the results no longer statistically significant. Here, there are two dimensions to consider: (1) the effect size (i.e. relative risk of SARS-CoV-2 infection) of the confounder, and (2) the relative prevalence of the confounder in the vaccinated vs. unvaccinated (matched) cohorts. For each vaccine, we compute the relative prevalence and effect size that would be required for an unobserved confounder to overturn the conclusion for a given (vaccine, time horizon) pair. We present the results from the tipping point analysis in Fig. 6.

Institutional Review Board (IRB) for study at the Mayo Clinic. This research was conducted under IRB 20-003278 at the Mayo Clinic, “Study of COVID-19 patient characteristics with augmented curation of Electronic Health Records (EHR) to inform strategic and operational decisions”. The Mayo Clinic granted IRB/ethical approval for this study and waived the need for informed consent (https://www.mayo.edu/research/institutional-review-board/overview). All methods were performed in accordance with the relevant guidelines and regulations supplied by the Mayo Clinic and HIPAA regulations regarding patient privacy protection. Subjects without research authorization on file were excluded.

Data availability
The primary data underlying this study was accessed via the Mayo Clinic upon approval of the IRB 20-003278 entitled “Study of COVID-19 patient characteristics with augmented curation of Electronic Health Records (EHR) to inform strategic and operational decisions”. On a case by case basis, requests for accessing the deidentified data sets will be considered by the Mayo Clinic, in keeping with HIPAA guidelines for patient privacy protection and the specific contents of the data requests. Address data requests to the corresponding authors of this manuscript.
Author contributions

The authors thank Murali Aravamudan, Patrick Lenehan, Jake Martin, Walter Kremers, and Hilal Maradit-Kremers, for their careful review and helpful feedback on this manuscript.

Acknowledgements

The authors thank Murali Aravamudan, Patrick Lenehan, Jake Martin, Walter Kremers, and Hilal Maradit-Kremers, for their careful review and helpful feedback on this manuscript.

Author contributions

C.P. and A.P. developed the methods and analytical techniques, interpreted the results, and wrote the manuscript. H.B. and V.A. supported the statistical analysis conducted. A.I. and V.S. led the study design and wrote the manuscript. C.P. and V.S. conceptualized the study and reviewed the manuscript. R.K., J.C.O.H., G.J.G., A.W.W., J.H. and A.D.B. interpreted the results and reviewed the manuscript. All authors reviewed the findings and revised the manuscript based on critical feedback from reviewers and colleagues.

Funding

The funding was provide by Division of Intramural Research, National Institute of Allergy and Infectious Diseases (AI110173), amfAR, The Foundation for AIDS Research (109593-62-RGRL) and also by National Institute of Allergy and Infectious Diseases (AI120698).

Competing interests

One or more of the investigators associated with this project and Mayo Clinic have a Financial Conflict of Interest in technology used in the research and that the investigator(s) and Mayo Clinic may stand to gain financially from the successful outcome of the research. This research has been reviewed by the Mayo Clinic Conflict of Interest Review Board and is being conducted in compliance with Mayo Clinic Conflict of Interest policies. ADB is a consultant for Abbvie, is on scientific advisory boards for Nference and Zentalis, and is founder and President of Splissen therapeutics. The authors from nference are employees of Nference and have financial interests in the...
success of inference. A provisional patent application filed covers some of the findings from this study with CP, AP and VS are named as inventors, with inference as the assignee for this patent application.

Additional information

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1038/s41598-021-83641-y.

Correspondence and requests for materials should be addressed to V.S.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher’s note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article’s Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article’s Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/.

© The Author(s) 2021