BMC Medical Research Methodology

RESEARCH

Open Access

Design and analysis of outcomes following SARS-CoV-2 infection in veterans

Valerie A. Smith[1,2,3], Theodore S. Z. Berkowitz[1], Paul Hebert[4,5], Edwin S. Wong[4,5], Meike Niederhausen[6,7,8], John A. Pura[1], Kristin Berry[4,9], Pamela Green[4], Anna Korpak[9], Alexandra Fox[9], Aaron Baraff[9], Alex Hickok[6], Troy A Shahoumian[10], Amy S.B. Bohnert[11,12], Denise M. Hynes[6,13], Edward J. Boyko[9], George N. Ioannou[4,14], Theodore J. Iwashyna[11,15,16], C. Barrett Bowling[1,17,18], Ann M. O'Hare[4,19] and Matthew L. Maciejewski[1,2,3*]

Abstract

Background Understanding how SARS-CoV-2 infection impacts long-term patient outcomes requires identification of comparable persons with and without infection. We report the design and implementation of a matching strategy employed by the Department of Veterans Affairs' (VA) COVID-19 Observational Research Collaboratory (CORC) to develop comparable cohorts of SARS-CoV-2 infected and uninfected persons for the purpose of inferring potential causative long-term adverse effects of SARS-CoV-2 infection in the Veteran population.

Methods In a retrospective cohort study, we identified VA health care system patients who were and were not infected with SARS-CoV-2 on a rolling monthly basis. We generated matched cohorts within each month utilizing a combination of exact and time-varying propensity score matching based on electronic health record (EHR)-derived covariates that can be confounders or risk factors across a range of outcomes.

Results From an initial pool of 126,689,864 person-months of observation, we generated final matched cohorts of 208,536 Veterans infected between March 2020-April 2021 and 3,014,091 uninfected Veterans. Matched cohorts were well-balanced on all 37 covariates used in matching after excluding patients for: no VA health care utilization; implausible age, weight, or height; living outside of the 50 states or Washington, D.C.; prior SARS-CoV-2 diagnosis per Medicare claims; or lack of a suitable match. Most Veterans in the matched cohort were male (88.3%), non-Hispanic (87.1%), white (67.2%), and living in urban areas (71.5%), with a mean age of 60.6, BMI of 31.3, Gagne comorbidity score of 1.4 and a mean of 2.3 CDC high-risk conditions. The most common diagnoses were hypertension (61.4%), diabetes (34.3%), major depression (32.2%), coronary heart disease (28.5%), PTSD (25.5%), anxiety (22.5%), and chronic kidney disease (22.5%).

Conclusion This successful creation of matched SARS-CoV-2 infected and uninfected patient cohorts from the largest integrated health system in the United States will support cohort studies of outcomes derived from EHRs and sample selection for qualitative interviews and patient surveys. These studies will increase our understanding of the long-term outcomes of Veterans who were infected with SARS-CoV-2.

Keywords COVID-19, Veterans, Matching, Study design, Retrospective, Causal inference, Longitudinal

*Correspondence:
Matthew L. Maciejewski
mlm34@duke.edu

Full list of author information is available at the end of the article

Background

The U.S. faces continued infections with Severe Acute Respiratory Syndrome Coronavirus 2 (SARS-CoV-2) following earlier waves in 2020 and 2021. Numerous studies have examined short-term symptoms, hospitalization, and death [1, 2] among patients infected with SARS-CoV-2 both before and after vaccine availability. The coronavirus disease 2019 (COVID-19) pandemic caused unprecedented disruptions across a wide range of domains relevant to health and health care including to health systems and care processes, informal family support, long-term care, safety net programs, social organizations, and economic stability, making it challenging to isolate the direct impact of infection with SARS-CoV-2 on individual health. Thus, more work is needed to characterize the direct long-term health effects of SARS-CoV-2 infection distinct from these systemic disruptions.

To research the long-term health consequences of COVID-19, the U.S. Department of Veterans Affairs (VA) provided resources in May 2021 to create the COVID-19 Outcomes Research Collaboratory (CORC). The main purposes of this program are investigating long-term health outcomes associated with SARS-CoV-2 infection using national electronic health record (EHR) data and survey research to assess factors not well captured by the EHR. The focus of this paper is on the study design utilized to facilitate EHR and survey-based research through cohort construction of Veterans with SARS-CoV-2 infection and well-matched controls, with matching for a wide range of covariates potentially associated with the exposure (SARS-CoV-2 infection) and outcomes. The creation of these populations will enable study of outcomes associated with this infection and potential mediating factors using observational research methods applied to both retrospective EHR and prospective survey-based data.

To best identify potential causal associations, we selected the target trial emulation approach to evaluate the causal effect of SARS-CoV-2 infection on long-term health-related outcomes. This approach is intended to minimize sources of bias, including observed and unobserved confounding and immortal time bias in estimating the effect of SARS-CoV-2 infection [3, 4]. To address unobserved confounding and selection bias, we matched Veterans infected with SARS-CoV-2 to similar contemporaneous uninfected Veterans before comparing a broad array of outcomes available in the VA's comprehensive longitudinal EHR.

Analyses based on the EHR will be supplemented with a prospective longitudinal survey in which a subset of matched cohorts of infected and uninfected persons will be invited to participate. Longitudinal survey responses will provide more detailed information on patient-reported outcomes that will complement outcomes ascertained from the EHR. Selection and matching of survey samples will be conducted in such a way to ensure covariate balance is maintained in the survey subsample. In addition, source data from the EHR and survey will be used to guide purposive sampling for a prospective qualitative study to understand the diversity of experiences of Veterans infected with SARS-CoV-2.

Designing a matched cohort study to address a wide array of EHR-based outcomes and embedded survey subsets requires a more inclusive consideration of confounding than when estimating the effect of SARS-CoV-2 infection on a single outcome. This protocol paper describes the design and methodological approach to identify a matched cohort of comparable patients infected and not infected with SARS-CoV-2. This matched cohort will be used in future research to analyze the effect of SARS-CoV-2 infection on clinical, functional, and economic outcomes among Veterans. This work could also inform and support efforts by other groups interested in creating matched cohorts to address a wide range of unanswered questions related to SARS-CoV-2 infection.

Methods

Study design and data

We designed a retrospective cohort study of EHR-based outcomes with a non-equivalent comparator of uninfected Veterans. To facilitate measurement of patient-reported outcomes, this retrospective cohort is paired with an embedded smaller post-only survey-based prospective cohort study. In both components, comparator non-equivalence was reduced by generating matched cohorts.

As described previously [5], we assembled a cohort of VA enrollees who tested positive for SARS-CoV-2 RNA in a respiratory specimen within the VA system based on polymerase chain reaction (PCR) tests as well as those with evidence of SARS-CoV-2 infection identified outside the VA but documented in VA records as identified by the VA National Surveillance Tool between March 1, 2020 and April 30, 2021. The earliest date of a documented positive test was taken as each patient's date of infection. We included only those Veterans who had an assigned VA primary care team (e.g., Patient Aligned Care Team) or at least one VA primary care clinic visit in the two-year period prior to infection to minimize missingness in EHR-based covariates that are generated from health system interaction. Cohorts were identified sequentially on a monthly basis, with assignment to a particular month for cases based on the date of the positive test or documentation in notes of non-VA evidence of infection. VA-enrolled Veterans without a positive test prior to or during the month who met the same inclusion criteria were considered uninfected potential comparators for that month. The uninfected control group

members were eligible for repeated sampling and matching with replacement until they had a positive test. To avoid misclassification of first infection date based on a positive test, infected Veterans with COVID-19-related diagnostic codes (ICD-10: B97.29, U07.1, U09.9, J12.82, Z86.16) listed in fee-for-service Medicare claims 15 or more days before their VA test were excluded. In addition, Veterans from the uninfected comparator group with any such diagnostic codes were excluded from sampling for matching in the month the COVID-19-related code arose and any months thereafter.

We developed 14 separate monthly patient cohorts—one for each month (March 2020-April 2021) —for the purpose of defining index dates and matching covariates. For example, the March 2020 cohort included all VA enrollees with an initial positive test during March 2020 and all VA enrollees who were alive as of March 1, 2020 and had not been infected prior to April 1, 2020. SARS-CoV-2-infected patients were included as potential comparator patients in months before infection. In a given month, uninfected Veterans could be matched to multiple infected Veterans in that same month and uninfected Veterans could be included in multiple month-specific cohorts as long as they remained uninfected and continued to meet other eligibility criteria. To minimize immortal time bias, the index date was defined as the date of the earliest positive test for SARS-CoV-2- infected Veterans and as the 1st day of the relevant month for uninfected Veterans [6]. Each patient's index date served as the anchor for defining matching covariates (with covariate construction starting 14 days prior to the positive test date for infected patients), based on EHR data from the prior two years.

Matching specification

Our goal was to conduct many-to-one matching that would maximize retention of infected patients for external validity and covariate balance for internal validity. A priori, we defined a suitable matching strategy as one that would result in <5% attrition of the infected cohort and achieve covariate balance among the selected covariates for matching based on standardized differences < 0.1 [7].

Coarsened exact matching (CEM) was initially attempted. Covariates used for matching were derived iteratively at a single point in time (summer 2021) with the understanding that the evidence base about causes and consequences of COVID-19 was (and is) evolving rapidly. In collaboration with clinician-investigators (see left column, Appendix 1), we identified a broad list of demographic, clinical, and health care utilization measures hypothesized to be either risk factors for pre-specified outcomes alone (e.g., survival, depression, total VA costs, disability, healthcare-related financial strain due to high out-of-pocket costs) or confounders associated with both infection and outcomes [8].

To minimize sample loss when attempting to match on many covariates in CEM [9], the five physician principal investigators then worked together to prioritize covariates for the final matching specification (see right column, Appendix 1). Modified coarsened exact matching was then implemented using this prioritized set of covariates. However, a suitable exact match could not be identified for 53.7% of infected Veterans, so we reverted to a form of combined exact and calendar time-specific propensity score matching [10], with cohorts identified by index month.

In a two-step process, infected patients were exact matched to uninfected controls based on index month, sex, immunosuppressive medication use (binary), state of residence, and COVID-19 vaccination status (effective in January-April 2021 cohorts only) because these covariates were strong potential confounders. In the second step, a total of 37 binary, categorical, and continuous covariates were included in the propensity score model, including immunosuppressive medication use (binary), nursing home residence any time in the prior two years, vaccination status (January-April 2021 cohorts), and diagnosed CDC high-risk conditions: [11] coronary heart disease, cancer (excluding non-metastatic skin cancers), chronic kidney disease, congestive heart failure, pulmonary-associated conditions (including asthma, COPD, interstitial lung disease, and cystic fibrosis), dementia, diabetes, hypertension, liver disease, sickle cell/thalassemia, solid organ or blood stem cell transplant, stroke/cerebrovascular disorders, substance use disorder, anxiety disorder, bipolar disorder, major depression, PTSD, and schizophrenia.

Other categorical variables in the propensity score model included sex, race, ethnicity, rurality of the Veteran's home ZIP code, state of residence, smoking status, and categorization of two comorbidity scores (CAN [12], Nosos [13]). Continuous covariates included age, body mass index (BMI), comorbidity score via Gagne index, distance from a Veteran's home to nearest VA hospital, and four VA utilization measures (inpatient admissions, primary care visits, specialty care visits, mental health visits in the prior 2 years).

A caliper of 0.2 times the pooled estimate of the standard deviation of the logit of the propensity score was used to bound which uninfected patients could be matched to each infected patient [14]. To provide the survey team a sufficiently deep pool of matched controls to account for survey non-participation, the 25 matched uninfected patients closest in propensity score were retained for each infected patient. Infected patients with fewer than 25 matched uninfected patients had all their comparator patients selected as eligible matches.

Matching was performed by the PSMATCH procedure from SAS/STAT 15.1 in SAS® 9.4M6 via the VA Informatics and Computing Infrastructure (VINCI) platform.

Outcomes comparisons to be conducted

The EHR-based clinical outcomes that we intend to compare between matched cohorts are mortality, depression, suicide, onset of new clinical diagnoses, exacerbation of prevalent conditions, development of COVID-19 sequelae, and health care use and VA health care costs. The survey-based outcomes to be compared between matched cohorts include disability, healthcare-related financial strain, and health-related quality of life. Our default approach to analyses will be "per-protocol", such that uninfected patients who cross over to become infected will be censored at the time of infection. Future analyses will account for this potentially informative censoring via inverse probability of censoring weights [15] and/or censoring of the entire matched strata at time of censoring. The study team discussed inclusion of negative control outcomes, but an outcome expected to be null between comparators could not be identified due to the ubiquitous effects of SARS-CoV-2 infection and the conditioning of negative control outcomes on health care utilization that might be differential between comparators.

Results

From a sampling frame of 231,160 Veterans who had documentation of at least one SARS-CoV-2 infection between March 2020 and April 2021, and 9,291,822 Veterans without evidence of infection over the same time period, we excluded patients who had neither a CAN comorbidity CAN score (i.e., were not assigned a PACT team) nor primary care use 24 months prior to index, or who had missing or implausible height, weight, or age (Fig. 1). We also excluded Veterans with missing ZIP codes or ZIP codes outside of Washington, D.C. or the 50 states, patients who were uninfected on the 1st of each month but became infected later in the same month (for the uninfected group), or had a prior infection documented in Medicare. Lastly, we excluded 776 (0.4%) of 209,312 infected patients who did not have a suitable match, which generated final matched cohorts of 208,536 infected and 3,014,091 uninfected Veterans (comprising 5,173,400 total person-months of follow-up because of matching with replacement). Unmatched infected patients are described in Appendix Table 2 and exhibited greater rates of missing information than those with suitable matches.

As expected, the cohorts prior to matching were imbalanced in many covariates (Appendix Table 3). After matching, the cohorts were well-balanced on all covariates, based on standardized mean differences (SMD) < 0.1 (Table 1). The cohorts included Veterans from all 50 states and Washington, D.C. Most Veterans in the matched cohorts were male (88.3%), non-Hispanic (87.1%), white (67.2%), and living in urban areas (71.5%), with a mean (standard deviation, SD) age of 60.6 (16.4), BMI of 31.3 (6.6), and mean (SD) straight-line distance to the closest VA medical center of 35.8 (35.2) miles. A minority were current smokers (12.6%), 39.3% had never smoked and 42.5% were former smokers. Comorbidity was assessed several ways, including Gagne score (mean = 1.4, SD = 2.2), count of CDC high-risk conditions (mean = 2.3, SD = 1.9) and count of 5 mental health conditions prevalent in Veterans (mean = 0.9, SD = 1.0). The most common diagnoses were hypertension (61.4%), diabetes (34.3%), major depression (32.2%), coronary heart disease (28.5%), PTSD (25.5%), anxiety (22.5%), and chronic kidney disease (22.5%). Approximately 10% of matched cohort members had been prescribed one or more immunosuppressive medications within 24 months before the index date (qualifying medications listed in the Appendix Table 4). Of the 34.0% of the cohort with index dates between January-April 2021 when vaccinations became available, 3,153 Veterans (1.5% of the entire infected cohort) received at least one dose of a vaccine before their first positive COVID-19 test result.

In the 24 months prior to the index date, cohort members had a mean (SD) of 8.3 (10.2) primary care visits, 13.4 (14.9) specialty care visits, and 7.8 (21.9) mental health visits in VA. Over one-half (53.3%) of infected patients were drawn from three of the 14 months in the observation period (November 2020, December 2020, and January 2021).

Discussion

Despite the very large sample size available for this research with 231,160 infected and just over 9 million uninfected Veterans, a strategy of bias reduction based on coarsened exact matching resulted in lack of an exact match for 53.7% of cases. The large sample loss reduced statistical power and generalizability since the exposure-disease associations may have differed between successfully matched and unmatched populations. The combined exact matching and propensity score approach, on the other hand, resulted in a much lower failure to match frequency at only 0.4%, with a high rate of success as assessed by the SMDs < 0.1 for all matching covariates. The work performed by the CORC to identify important covariates on which to match cases and uninfected controls using propensity score methodology should facilitate the performance of causal research on long COVID-19 etiology in this population. Matched cohorts will be updated from May 2021 forward to be able to generate evidence on Veteran experience after April 2021. Given the ever-changing environment of

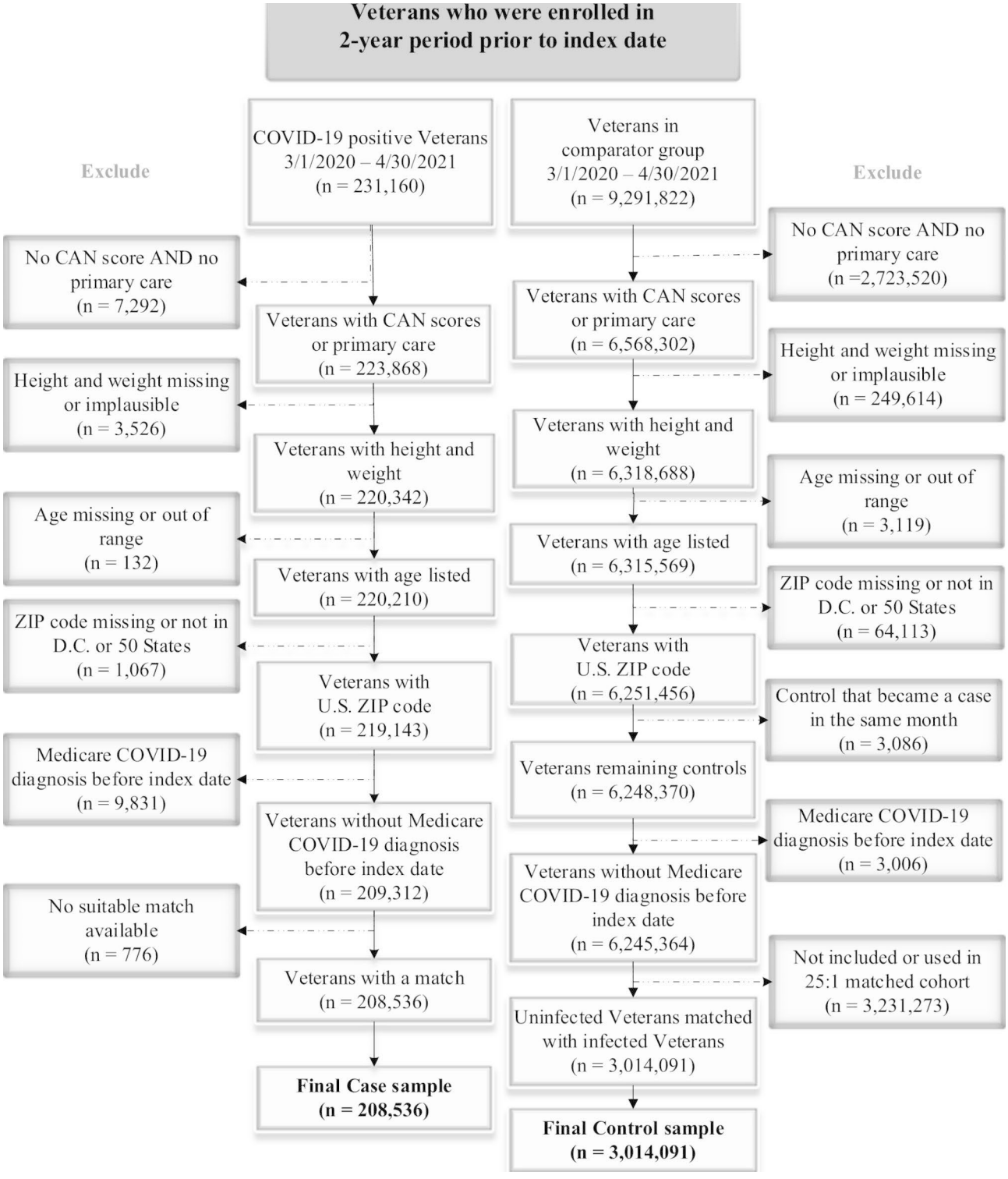

Fig. 1 STROBE Figure of Cohort Derivation

variants, vaccination status, immunogenicity from prior exposure, and tests and treatments available, future analyses will consider period-specific effects and include individuals with antigen test-detected infections in these future cohorts.

As the nation's largest national integrated publicly financed health system, the VA has the unique ability to track long-term outcomes among individuals infected with SARS-CoV-2 because it has a well-established comprehensive EHR that was developed around the mission

Table 1 Descriptive Statistics of Matched Cohorts

Variable Name	*COVID-19* (n = 208,536)	*Uninfected* (n = 5,173,400)	*SMD*
Age, mean (SD)	60.6 (16.2)	60.6 (16.4)	0.00106
BMI, mean (SD)	31.3 (6.4)	31.3 (6.6)	0.01178
Sex, N (%)			
Female	21,949 (10.5)	540,127 (10.4)	0.00988
Male	183,676 (88.1)	4,566,632 (88.3)	
Unknown	2,911 (1.4)	66,641 (1.3)	
Race, N (%)			
American Indian/Alaska Native	1,965 (0.9)	47,784 (0.9)	0.00610
Asian	2,081 (1.0)	51,718 (1.0)	
Black or African American	47,726 (22.9)	1,189,212 (23.0)	
Native Hawaiian or Other Pacific Islander	1,946 (0.9)	48,095 (0.9)	
White	139,952 (67.1)	3,474,658 (67.2)	
Multiple Races	1,978 (0.9)	49,028 (0.9)	
Missing	12,888 (6.2)	312,905 (6.0)	
Hispanic Ethnicity, N (%)			
Yes	20,309 (9.7)	492,607 (9.5)	0.00792
No	181,097 (86.8)	4,506,252 (87.1)	
Missing	7,130 (3.4)	174,541 (3.4)	
Rurality, N (%)			
Urban	149,310 (71.6)	3,700,754 (71.5)	0.00144
Not Urban (incl. missing)	59,226 (28.4)	1,472,646 (28.5)	
Smoking Status, N (%)			
Current	26,188 (12.6)	654,415 (12.6)	0.01379
Former	88,275 (42.3)	2,201,135 (42.5)	
Never	81,916 (39.3)	2,032,442 (39.3)	
Missing	12,157 (5.8)	285,408 (5.5)	
CDC High Risk Conditions, N (%)			
Coronary Heart Disease	59,972 (28.8)	1,472,173 (28.5)	0.00668
Cancer	38,136 (18.3)	942,791 (18.2)	0.00165
Chronic Kidney Disease	47,610 (22.8)	1,161,200 (22.4)	0.00920
Congestive Heart Failure	22,124 (10.6)	532,106 (10.3)	0.01059
Pulmonary	46,178 (22.1)	1,131,201 (21.9)	0.00671
Dementia	10,837 (5.2)	254,136 (4.9)	0.01298
Diabetes	72,020 (34.5)	1,773,699 (34.3)	0.00528
Hypertension	127,761 (61.3)	3,174,235 (61.4)	0.00187
Liver Disease	15,324 (7.3)	370,974 (7.2)	0.00684
Sickle Cell	383 (0.2)	9,131 (0.2)	0.00169
Transplant	687 (0.3)	16,864 (0.3)	0.00061
Stroke/Cerebrovascular disease	13,012 (6.2)	314,234 (6.1)	0.00689
Substance Use Disorder	25,915 (12.4)	645,302 (12.5)	0.00140
Anxiety	47,186 (22.6)	1,165,628 (22.5)	0.00230
Bipolar disorder	8,039 (3.9)	199,388 (3.9)	0.00005
Major Depression	67,383 (32.3)	1,664,825 (32.2)	0.00282
PTSD	52,964 (25.4)	1,317,278 (25.5)	0.00148
Schizophrenia	4,737 (2.3)	115,145 (2.2)	0.00309
# CDC High Risk Conditions, mean (SD)	2.3 (1.9)	2.3 (1.9)	0.01088
# Mental Health Conditions (Anxiety, Bipolar disorder, major depression, PTSD, schizophrenia), mean (SD)	0.9 (1.1)	0.9 (1.0)	0.00200
Gagne score, mean (SD)	1.4 (2.3)	1.4 (2.2)	0.01559
# VA Inpatient Admissions, mean (SD)	0.4 (1.2)	0.4 (1.3)	0.00079
# VA Primary Care Visits, mean (SD)	8.6 (9.7)	8.3 (10.2)	0.02916
# VA Specialty Care Visits, mean (SD)	13.9 (14.1)	13.4 (14.9)	0.03288
Mental Health Care Utilization, mean (SD)	7.9 (22.6)	7.8 (21.9)	0.00457

Table 1 (continued)

Variable Name	***COVID-19*** **(n = 208,536)**	***Uninfected*** **(n = 5,173,400)**	***SMD***
Immunosuppressed in prior 24 months, N (%)	20,366 (9.8)	502,358 (9.7)	0.00188
CLC At Index Date, N (%)	2,176 (1.0)	46,586 (0.9)	0.01457
NOSOS score, N (%)			
NOSOS, missing	4,801 (2.3)	110,729 (2.1)	0.03224
NOSOS Category 1 [0, 0.417)	5,708 (2.7)	129,206 (2.5)	
NOSOS Category 2 [0.417, 0.471)	9,599 (4.6)	227,788 (4.4)	
NOSOS Category 3 [0.471, 0.534)	12,449 (6.0)	305,685 (5.9)	
NOSOS Category 4 [0.534, 0.611)	15,187 (7.3)	378,825 (7.3)	
NOSOS Category 5 [0.611, 0.707)	17,737 (8.5)	450,045 (8.7)	
NOSOS Category 6 [0.707, 0.829)	20,556 (9.9)	525,057 (10.1)	
NOSOS Category 7 [0.829, 0.998)	23,462 (11.3)	599,981 (11.6)	
NOSOS Category 8 [0.998, 1.259)	26,956 (12.9)	686,219 (13.3)	
NOSOS Category 9 [1.259, 1.805)	31,372 (15.0)	786,826 (15.2)	
NOSOS Category 10 [1.805, Inf)	40,709 (19.5)	973,039 (18.8)	
CAN Score, N (%)			
CAN, missing	4,066 (1.9)	86,041 (1.7)	0.03258
CAN Category 1, 0–20	34,427 (16.5)	851,737 (16.5)	
CAN Category 2, 25–40	31,926 (15.3)	801,677 (15.5)	
CAN Category 3, 45–60	38,266 (18.3)	970,378 (18.8)	
CAN Category 4, 65–80	46,894 (22.5)	1,184,134 (22.9)	
CAN Category 5, 85–90	30,745 (14.7)	762,537 (14.7)	
CAN Category 6, 95–99	22,212 (10.7)	516,896 (10.0)	
Vaccinated in January-April 2021, N (%)	3,153 (1.5)	75,353 (1.5)	0.00470
Index Month, N (%)			
March 2020	2,340 (1.1)	58,278 (1.1)	0.00363
April 2020	6,786 (3.3)	168,727 (3.3)	
May 2020	4,402 (2.1)	109,604 (2.1)	
June 2020	6,967 (3.3)	173,776 (3.4)	
July 2020	14,945 (7.2)	373,025 (7.2)	
August 2020	8,376 (4.0)	208,750 (4.0)	
September 2020	6,787 (3.3)	169,048 (3.3)	
October 2020	12,212 (5.9)	302,119 (5.8)	
November 2020	31,004 (14.9)	767,646 (14.8)	
December 2020	43,732 (21.0)	1,085,816 (21.0)	
January 2021	36,282 (17.4)	900,381 (17.4)	
February 2021	15,699 (7.5)	388,553 (7.5)	
March 2021	9,942 (4.8)	244,839 (4.7)	
April 2021	9,062 (4.3)	222,838 (4.3)	
U.S. State of Residence, N (%)			

Table 1 (continued)

Variable Name	*COVID-19* (n = 208,536)	*Uninfected* (n = 5,173,400)	*SMD*
Alaska	694 (0.3)	16,880 (0.3)	0.01515
Alabama	4,089 (2.0)	101,648 (2.0)	
Arkansas	3,075 (1.5)	76,426 (1.5)	
Arizona	6,310 (3.0)	157,105 (3.0)	
California	16,253 (7.8)	405,136 (7.8)	
Colorado	3,278 (1.6)	80,733 (1.6)	
Connecticut	1,895 (0.9)	46,782 (0.9)	
Washington, D.C.	414 (0.2)	10,111 (0.2)	
Delaware	552 (0.3)	13,199 (0.3)	
Florida	16,003 (7.7)	399,175 (7.7)	
Georgia	7,893 (3.8)	196,375 (3.8)	
Hawaii	263 (0.1)	6,295 (0.1)	
Iowa	2,473 (1.2)	59,479 (1.1)	
Idaho	1,582 (0.8)	39,124 (0.8)	
Illinois	6,462 (3.1)	159,840 (3.1)	
Indiana	4,533 (2.2)	112,806 (2.2)	
Kansas	2,471 (1.2)	61,282 (1.2)	
Kentucky	3,352 (1.6)	82,956 (1.6)	
Louisiana	3,754 (1.8)	93,423 (1.8)	
Massachusetts	2,990 (1.4)	73,840 (1.4)	
Maryland	2,640 (1.3)	65,047 (1.3)	
Maine	631 (0.3)	15,442 (0.3)	
Michigan	4,383 (2.1)	108,714 (2.1)	
Minnesota	4,251 (2.0)	105,503 (2.0)	
Missouri	6,609 (3.2)	164,281 (3.2)	
Mississippi	1,804 (0.9)	44,823 (0.9)	
Montana	1,122 (0.5)	27,824 (0.5)	
North Carolina	8,521 (4.1)	212,232 (4.1)	
North Dakota	865 (0.4)	21,154 (0.4)	
Nebraska	1,872 (0.9)	46,270 (0.9)	
New Hampshire	743 (0.4)	18,024 (0.3)	
New Jersey	2,623 (1.3)	64,850 (1.3)	
New Mexico	1,373 (0.7)	34,024 (0.7)	
Nevada	3,049 (1.5)	75,467 (1.5)	
New York	7,739 (3.7)	192,659 (3.7)	
Ohio	9,173 (4.4)	228,564 (4.4)	
Oklahoma	4,030 (1.9)	100,147 (1.9)	
Oregon	1,386 (0.7)	34,259 (0.7)	
Pennsylvania	7,880 (3.8)	196,104 (3.8)	
Rhode Island	762 (0.4)	18,729 (0.4)	
South Carolina	6,352 (3.0)	157,970 (3.1)	
South Dakota	1,408 (0.7)	29,278 (0.6)	
Tennessee	5,848 (2.8)	145,295 (2.8)	
Texas	19,436 (9.3)	485,224 (9.4)	
Utah	1,243 (0.6)	30,706 (0.6)	
Virginia	5,070 (2.4)	125,759 (2.4)	
Vermont	149 (0.1)	3,594 (0.1)	
Washington	1,976 (0.9)	48,855 (0.9)	
Wisconsin	5,087 (2.4)	126,171 (2.4)	
West Virginia	1,621 (0.8)	40,213 (0.8)	
Wyoming	554 (0.3)	13,603 (0.3)	
Distance to Nearest VAMC (miles), mean (SD)	35.6 (36.5)	35.8 (35.1)	0.00452

Table 1 (continued)

Variable Name	*COVID-19* (n = 208,536)	*Uninfected* (n = 5,173,400)	*SMD*

Note: The uninfected cohort represents 3,014,091 unique Veterans with a total of 5,173,400 person-months since they were matched with replacement

of providing lifelong care for Veterans. In addition, Veterans are historically reliant on VA for care if they engage with the health system.

Our matching strategy defines the specific effect that will be estimated from our results. We considered historical controls of Veterans receiving care in the VA before the pandemic, but that would estimate the effect of individual SARS-CoV-2 infection combined with all the many other social and systematic disruptions that accompanied the pandemic. We also considered comparing Veterans hospitalized with SARS-CoV-2 infection to Veterans hospitalized with other conditions (e.g., influenza), which would be analogous to a randomized clinical trial with an active comparator. Such a comparator group asks whether COVID-19 hospitalization is worse than other sorts of hospitalizations. We reasoned that, for most Veterans, had they not developed COVID-19, they may not have been hospitalized with another condition that same month (although we did not exclude those hospitalized, so they do occur at whatever their natural frequency is in the comparator group).

We also did not wish to restrict to only hospitalized COVID-19 patients, as we took as a scientific question the relationship between initial severity of SARS-CoV-2 infection and subsequent outcomes—rather than presuming it by conditioning on initial severity. We also considered Veterans infected with other non-SARS-CoV-2 viruses. However, we noted the substantial body of evidence on sepsis and pneumonia—much of it viral in origin—that suggested such patients also have adverse long-term outcomes caused by non-SARS-CoV-2 viruses, including influenza. As such, we reasoned such comparators might underestimate the total individual effects survivors of COVID-19 would face and health systems would need to support. Each of these comparators may be of great interest to other research groups; they were not, however, our primary focus. Our goal was to generate matched cohorts to support cohort studies of EHR-derived outcomes and sample selection for qualitative interviews and patient surveys.

The retrospective cohort study described here is subject to several limitations. First, cohort matching results in sample loss that may reduce generalizability of results compared to weighting methods, although we were able to retain > 99% of the infected patients in the sample after matching. Second, there is likely some contamination of the uninfected comparator group with Veterans with undiagnosed SARS-CoV-2 infection or who tested positive for SARS-CoV-2 with test results not available from private insurers, Medicare Advantage plans, Medicaid, or other community sources. Third, covariate specification for matching is based on our understanding of risk factors and confounders of SARS-CoV-2 infection as of spring 2022, however, we are unable to measure all risk factors via administrative data. Specifically, unmeasured confounders such as employment, income, or other social vulnerability indicators may be imbalanced between matched groups and could confound the association between infection and outcomes. Fourth, results may not generalize to Veterans who became infected after April 2021 or to non-Veterans. CORC will update matched cohorts from May 2021-March 2022, and that work is ongoing.

Conclusion

Our understanding of the long-term outcomes of Veterans who were infected with SARS-CoV-2 will be gleaned from qualitative interviews, population-based surveys, and cohort studies of outcomes derived from EHRs. This study will explore all these approaches, all framed in the context of the matched cohorts generated from EHR data from the largest integrated health system in the U.S. Due to Veterans' reliance on VA for care and eligibility for care once enrolled, we will be able to evaluate clinical and economic outcomes following their acute SARS-CoV-2 infection, as long-term outcomes two years after the onset of the pandemic are now being realized.

Abbreviations

CAN	Care Assessment Need
CDC	Centers for Disease Control and Prevention
COPD	Chronic obstructive pulmonary disease
CAD	Coronary heart disease
HIV	Human immunodeficiency virus
CKD	Chronic kidney disease
CHF	Congestive heart failure
SUD	Substance use disorder
SMI	Serious mental illness
PTSD	Post-traumatic stress disorder
VA	Veterans Health Administration
CBOC	VA community-based outpatient clinic.

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s12874-023-01882-z.

Supplementary Material 1

Authors' contributions

Conception: VAS, TSZB, PH, ESW, MN, JAP, KB, ASB, DMH, EJB, GNI, TJI, CBB, AMO, MLM. Design of the work: VAS, TSZB, PH, ESW, MN, JAP, ASB, DMH, EJB, GNI, TJI, CBB, AMO, MLM. Data acquisition: TSZB, KB, PG, AK, AF, AB, AH, TAS,

GNI. Analysis: VAS, TSZB, KB, MLM. Interpretation of data: VAS, TSZB, PH, MN, JAP, KB, ASB, DMH, EJB, GNI, TJI, CBB, AMO, MLM. Drafted work: VAS, TSZB, MLM. Substantive revision: VAS, TSZB, PH, ESW, MN, JAP, KB, PG, AK, AF, AB, AH, TAS, ASB, DMH, EJB, GNI, TJI, CBB, AMO, MLM.

Funding/Support
The study was supported by the U.S. Department of Veterans Affairs HSR&D grant C19 21–278 and C19 21–279. MM and DMH were also supported by a senior Research Career Scientist award from the Department of Veterans Affairs (RCS 10–391 to M.M. and RCS 21–136 to D.M.H.) and by the Durham VA Center of Innovation to Accelerate Discovery and Practice Transformation (CIN 13–410).

Data availability
The datasets generated and/or analyzed during the current study are not publicly available due to Department of Veterans Affairs data restrictions prohibiting sharing. Contact the corresponding, Dr. Matthew Maciejewski, for data requests.

Declarations

Ethics approval and consent to participate
Initial and continuing reviews approved by were reviewed and approved by the Durham Veterans Affairs Institutional Review Board and Research and Development Committee. All methods were carried out in accordance with relevant guidelines and regulations. No informed consent was obtained because the project is secondary analysis of data. The Durham VA Institutional Review Board granted waivers for informed consent, HIPAA information, and HIPAA research on decedents.

Consent for publication
Not applicable.

Competing interests
The authors declare that they have no competing interests.

Role of the sponsor
The Health Services Research and Development Service, Department of Veterans Affairs had no role in the design, conduct, collection, management, analysis, or interpretation of the data; or in the preparation, review, or approval of the manuscript. The opinions expressed are those of the authors and not necessarily those of the Department of Veterans Affairs, the United States Government, Duke University, the University of Washington, the University of Michigan, Oregon Health & Science University (OHSU), Portland State University, and Oregon State University.

Author details
[1]Center of Innovation to Accelerate Discovery and Practice Transformation, Durham VA Medical Center, Durham, NC, USA
[2]Department of Population Health Sciences, Duke University, Durham, NC, USA
[3]Division of General Internal Medicine, Department of Medicine, Duke University, Durham, NC, USA
[4]Health Services Research & Development Center of Innovation for Veteran-Centered and Value-Driven Care, and Gastroenterology section, Veterans Affairs Puget Sound Health Care System, Seattle, WA, USA
[5]Department of Health Systems and Population Health, University of Washington, Seattle, WA, USA
[6]Center to Improve Veteran Involvement in Care, VA Portland Health Care System, Portland, OR, USA
[7]Oregon Health & Science University (OHSU), Portland, OR, USA
[8]Portland State University School of Public Health, Portland, OR, USA
[9]Seattle Epidemiologic Research and Information Center, VA Puget Sound, Seattle, WA, USA
[10]Population Health: Health Solutions, Veterans Health Administration, Washington, DC, USA
[11]VA Center for Clinical Management Research, Ann Arbor, VA, MI, USA
[12]Departments of Anesthesiology and Psychiatry, University of Michigan Medical School, Ann Arbor, MI, USA
[13]College of Public Health and Human Sciences, Center for Quantitative Life Sciences, Oregon State University, Corvallis, OR, USA
[14]Division of Gastroenterology, University of Washington, Seattle, WA, USA
[15]National Clinical Scholars Program, University of Michigan Medical School, Ann Arbor, MI, USA
[16]Department of Internal Medicine, University of Michigan Medical School, Ann Arbor, MI, USA
[17]Geriatric Research Education and Clinical Center, Durham VA Medical Center, Durham, NC, USA
[18]Department of Medicine, Duke University, Durham, NC, USA
[19]Division of Nephrology, University of Washington, Seattle, WA, USA

Received: 2 June 2022 / Accepted: 3 March 2023
Published online: 04 April 2023

References

1. Al-Aly Z, Xie Y, Bowe B. High-dimensional characterization of post-acute sequelae of COVID-19. Nature. 2021;594(7862):259–64.
2. Groff D, Sun A, Ssentongo AE, Ba DM, Parsons N, Poudel GR, et al. Short-term and long-term rates of Postacute Sequelae of SARS-CoV-2 infection: a systematic review. JAMA Netw Open. 2021;4(10):e2128568.
3. Hernan MA, Alonso A, Logan R, Grodstein F, Michels KB, Willett WC, et al. Observational studies analyzed like randomized experiments: an application to postmenopausal hormone therapy and coronary heart disease. Epidemiology. 2008;19(6):766–79.
4. Hernan MA, Robins JM. Using Big Data to emulate a target Trial when a Randomized Trial is not available. Am J Epidemiol. 2016;183(8):758–64.
5. Ioannou GN, Ferguson JM, O'Hare AM, Bohnert ASB, Backus LI, Boyko EJ, et al. Changes in the associations of race and rurality with SARS-CoV-2 infection, mortality, and case fatality in the United States from February 2020 to March 2021: a population-based cohort study. PLoS Med. 2021;18(10):e1003807.
6. Suissa S. Immortal time bias in pharmaco-epidemiology. Am J Epidemiol. 2008;167(4):492–9.
7. Austin PC. A Tutorial and Case Study in Propensity score analysis: an application to estimating the Effect of In-Hospital Smoking Cessation Counseling on Mortality. Multivar Behav Res. 2011;46(1):119–51.
8. Brookhart MA, Schneeweiss S, Rothman KJ, Glynn RJ, Avorn J, Sturmer T. Variable selection for propensity score models. Am J Epidemiol. 2006;163(12):1149–56.
9. Ripollone JE, Huybrechts KF, Rothman KJ, Ferguson RE, Franklin JM. Evaluating the utility of coarsened exact matching for Pharmacoepidemiology using real and simulated Claims Data. Am J Epidemiol. 2020;189(6):613–22.
10. Mack CD, Glynn RJ, Brookhart MA, Carpenter WR, Meyer AM, Sandler RS, et al. Calendar time-specific propensity scores and comparative effectiveness research for stage III colon cancer chemotherapy. Pharmacoepidemiol Drug Saf. 2013;22(8):810–8.
11. Prevention CfDCa. Underlying Medical Conditions Associated with Higher Risk for Severe COVID-19: Information for Healthcare Professionals Atlanta, GA: Centers for Disease Control and Prevention. ; 2022 [Available from: https://www.cdc.gov/coronavirus/2019-ncov/hcp/clinical-care/underlying-conditions.html.
12. Wang L, Porter B, Maynard C, Evans G, Bryson C, Sun H, et al. Predicting risk of hospitalization or death among patients receiving primary care in the Veterans Health Administration. Med Care. 2013;51(4):368–73.
13. Wagner TH, Upadhyay A, Cowgill E, Stefos T, Moran E, Asch SM, et al. Risk Adjustment Tools for Learning Health Systems: a comparison of DxCG and CMS-HCC V21. Health Serv Res. 2016;51(5):2002–19.
14. Austin PC. Optimal caliper widths for propensity-score matching when estimating differences in means and differences in proportions in observational studies. Pharm Stat. 2011;10(2):150–61.
15. Buchanan AL, Hudgens MG, Cole SR, Lau B, Adimora AA, Study WsIH. Worth the weight: using inverse probability weighted Cox models in AIDS research. AIDS Res Hum Retroviruses. 2014;30(12):1170–7.

Publisher's note