NOTE: The following protocol was approved by CDC on 2/18/2021. It is being made available for awareness of other researchers primarily and for public transparency. In summary, the aim of the project described in the protocol is to evaluate long COVID-19 symptoms in the **Vaccine Safety Datalink population**, a very large, multisite, demographically and clinically diverse population of all ages, with COVID-19 disease identified in both outpatient and inpatient settings.

VSD Study 1344

Long COVID-19: Changes in Healthcare Utilization Following Infection with SARS-CoV-2

Principal Investigator: Sara Tartof, PhD MPH, Department of Research & Evaluation, Division of Epidemiology

Co-Investigators: Lei Qian, PhD; Bruno Lewin, MD; Debbie Malden, DPhil, MSc; Lina Sy, MPH

Background and Significance

As of November 25, 2020, approximately 60 million people worldwide have been infected with COVID-19, and 1.4 million have died. To date, the United States has experienced the greatest absolute burden of COVID-19-associated cases and deaths worldwide, with more than 12.7 million cases and 260,000 deaths as of December 1, 2020. Moreover, reported figures likely underestimate the true burden of COVID-19 due to either asymptomatic infection or mild illness. As such, it is estimated that for every case of COVID-19 in the United States, there are approximately 7-8 unreported infections.

Recent results from follow-up studies have indicated that a substantial proportion of patients may continue to experience a range of lingering symptoms for many months following the initial acute phase of infection[1-6]. These have been termed "long haulers," or persons with "long COVID-19." Currently, there is no consensus definition of what defines the "long haul" (i.e., post-acute COVID-19 sequelae or "long COVID-19"). In the COVID-19 Symptom Study, in which more than 4 million people in the United States, United Kingdom, and Sweden entered their symptoms after a COVID-19 diagnosis, post-acute COVID-19 was defined as the presence of symptoms extending beyond 3 weeks from the initial onset. Chronic COVID-19 was defined as extending beyond 12 weeks.

Estimates of the proportion of patients with COVID-19 who experience long COVID-19 symptoms (LCS) range widely. One study, which has not yet been published, using an app created by King's College London and Massachusetts General Hospital, found that of more than 4,000 patients with COVID-19, about 10% of those aged 18 to 49 years still experienced symptoms 4 weeks after becoming sick, 4.5% of all ages had symptoms lasting more than 8 weeks, and 2.3% had them for more than 12 weeks[7].

Other studies from the United States and Europe of non-hospitalized COVID-19 patients have reported approximately 25-30% of patients still experience symptoms after 90 days (13 weeks). A study conducted by CDC found that 35% of symptomatic respondents reported residual symptoms at least 2 weeks after testing positive[8]. A study of hospitalized patients in Italy found that in patients who had recovered from COVID-19, 87.4% reported persistence of at least one symptom, particularly fatigue and dyspnea, a mean of 60 days (8 weeks) after initial symptom onset[1].

While other viral infections have been associated with persistence of symptoms[3], what differentiates COVID-19 persistence is the wide-ranging symptoms involving multiple organ systems. LCS can affect many parts of the body, including the heart, lungs, and the digestive and nervous systems[1,2,4, 9]. The most commonly reported symptoms after acute COVID-19 infection are fatigue and dyspnea (difficulty breathing)[4]. However, additional reported symptoms include cardiac inflammation, abdominal pain and diarrhea, anosmia (loss of smell), a condition resembling chronic fatigue syndrome, "brain fog" characterized by difficulty with concentration and memory, psychiatric disorders such as anxiety and depression[10], and dysautonomia (a disorder of autonomic nervous system that regulates nonvoluntary body functions, such as heart rate, blood pressure, and sweating).

The conditions associated with LCS are diverse and complex and may vary by severity of the acute phase of COVID-19 infection or pre-existing health status prior to infection. While LCS may impact all age groups, most studies have been restricted to adult populations[1,2,5,6,8]. Furthermore, the majority of evidence around LCS has been gathered from follow-up studies of previously hospitalized patients that often lack information on symptom history before acute COVID-19 illness or details on symptom severity[1]. Furthermore, the single-center studies have information on a relatively small number of hospitalized patients, and many studies lack unbiased comparison groups of patients discharged for other reasons, limiting the interpretation of the reported associations. More studies are needed to better characterize the duration, frequency, and types of LCS.

The current study will address these limitations by evaluating LCS in a very large, multisite, demographically and clinically diverse population of all ages with COVID-19 identified in outpatient and inpatient settings.

Two-Phase Approach

We propose a two-phase approach that will maximize our potential to publish important findings on LCS as quickly as possible, while also diving more deeply into questions that may incur a longer timeline. For **Phase I**, we will conduct a rapid multi-site interrupted time series (ITS) study to assess changes in high-level utilization counts in the 3 months (and ultimately 6 months and longer, when more follow-up time has accrued) post-index date versus a pre-index date time period in 2019[11]. To adjust for secular confounding (e.g., impact of COVID-19 pandemic), we will compare pre- and post-utilization among patients who test positive for COVID-19 (hereafter, "patients who test positive" versus a matched population of patients who test negative for COVID-19 (hereafter, "patients who test negative"). We will further stratify by characteristics of interest to explore sub-groups at highest risk of LCS.

For **Phase II**, we propose a matched retrospective cohort study to investigate the risk of LCS, by severity of COVID-19 infection (inpatient, Intensive Care Unit (ICU)). In this phase, we will assess the additional burden of utilization associated with COVID-19, beyond that which might occur in the same settings without COVID-19. This is particularly important in the ICU, where post-intensive care syndrome [3] is a well-documented condition, even in the absence of COVID-19.

Study Aims

Phase I

Aim 1a.) We will describe the trajectory of weekly counts (i.e., counts over time) of utilization (combined and stratified by outpatient, inpatient, Emergency Department (ED), virtual settings) at each

participating site, and by age, sex, and race/ethnicity, over the study period for patients who test positive versus patients who test negative, starting from the time of COVID-19 test. We will use the encounter distributions to verify the approach to define the index date. Next, we will focus on utilization across settings for the 3 months following index dates, as well as corresponding (by calendar month and day) pre-periods in 2019.

Aim 1b.) Use an interrupted time series (ITS) approach to estimate the difference in changes in utilization, by healthcare setting (i.e., stratified by outpatient, inpatient, ED, virtual; and outpatient vs. inpatient), in the pre- and post-period for patients who test positive versus patients who test negative.

Aim 1c.) Use ITS approach to estimate the difference in changes in utilization (all settings combined) in the pre- and post-period for patients who test positive versus patients who test negative, stratified by age, sex, race/ethnicity, and clinical characteristics of interest.

Aim 1d.) Use ITS approach to estimate the difference in changes in utilization (all settings combined) in the pre- and post-period for patients who test positive versus patients who test negative, stratified by select LCS types.

Conduct analyses for 3 and 6-month follow-up periods. We will examine longer time periods if data are available.

Phase II (THIS PHASE IS ON PAUSE FOR NOW)

Aim 2a.) Assess risk of LCS by severity of COVID-19 disease

Conduct matched retrospective cohort analyses, stratified by in-patient treatment category: receiving treatment in hospital and ICU settings.

Approach

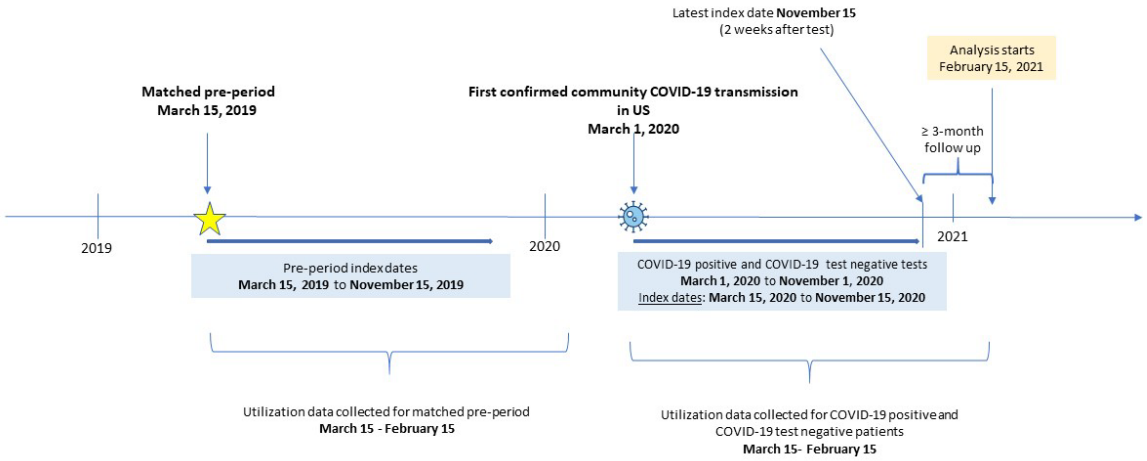
Aim 1

Study Design: Retrospective cohort study.

Environment: Vaccine Safety Datalink (VSD) multi-site study.

Study Period: We will include utilization, comorbidity, clinical, and other data on patients who test positive versus patients who test negative (defined by SARS-CoV-2 laboratory test) from March 1, 2020 to February 15, 2021, and from March 15, 2019 to February 15, 2020. For complete assessment of utilization, we will also require continuous membership (allowing for a 31-day gap) from 12 months prior to the index date and 3 months (and ultimately 6 months or more when more time accumulates, which will extend the end of follow-up to May 15, 2021 or later) after the index date. Patients who test negative will be defined as those who have a negative COVID-19 diagnostic test from March 1, 2020 through 3 months (and ultimately 6 months) following their index date.

Figure 1. Timeline for Cohort Construction



Index Date: The index date will be defined as the date 2 weeks following the date of positive SARS-CoV-2 laboratory test. This allows a 2-week washout window immediately post-positive laboratory test. For example, if a patient is diagnosed with COVID-19 by positive laboratory test on May 1, 2021, the index date for this patient will be May 15, 2021. Patients who test negative may be matched to patients who test positive on VSD site, age, sex and race/ethnicity, and test date, based on data availability.

Study Cohort: The cohort will include patients of all ages who test positive for SARS-CoV-2 from March 1, 2020 to November 1, 2020 and patients who test negative for SARS-CoV-2, matched by test date to a patient who tests positive. We will restrict the study to patients with positive COVID-19 tests until November 1, 2020, to allow at least 3-months of follow-up. This date assumes that study analyses will begin no earlier than February 15, 2021.

Patients who test negative will be matched to patients who test positive using the matching criteria described above.

Covariate Data to Collect: Although our interrupted time series approach removes the need to adjust for comorbidities (pre-post comparison), it is possible that patients who test positive differ from patients who test negative on prevalence of certain factors that are also associated with utilization. Therefore, we plan to collect comorbidity data in the 12 months prior to the index date, including demographic characteristics (age, sex, race/ethnicity, Medicaid), prior healthcare utilization (outpatient, inpatient, ED settings 12 months prior), clinical comorbidities (cancer, chronic kidney disease, cardiac disease, organ transplant, sickle cell disease, diabetes I and II, down syndrome, asthma, cardiovascular disease, cystic fibrosis, hypertension, liver disease, fibrosis, thalassemia, immunodeficiency, and dementia [comorbidities included according to CDC COVID-19 high risk category definition]; and obesity based on body mass index [BMI]), and other potential covariates of interest for stratified subgroup analyses. In addition, for analysis of 6-month follow-up period, we will collect COVID-19 vaccine information among the population of patients who test positive and patients who test negative, since the follow-up period and vaccination period will possibly overlap.

Study Aims

Aim 1a.) ***We will describe the trajectory of weekly counts of utilization at each participating site, and by age, sex, and race/ethnicity, over the study period for patients who test positive versus patients who test negative. We will use the encounter distributions to verify the approach to define index date. Next, we will focus on utilization across settings for the 3 months following index dates, as well as corresponding (by calendar month and day) pre-periods in 2019.***

For this aim we will first identify all patients with a positive SARS-CoV-2 laboratory test from March 1, 2020 to November 1, 2020. We will then match these patients who test positive to patients who test negative on VSD site, and by age, sex, race/ethnicity, and test date based on data availability. We will describe the trajectory of weekly counts (i.e. counts over time) of utilization (combined and stratified by outpatient, inpatient, ED, virtual settings) at each participating site, and by age, sex, and race/ethnicity, over the study period for patients who test positive versus patients who test negative, starting from the date of COVID-19 test.

We propose an index date defined as the date 2 weeks following the date of positive or negative SARS-CoV-2 laboratory test. However, with these first descriptive analyses, we plan to verify this approach with utilization data and therefore plan to capture utilization starting at the time of the laboratory test, rather than only 14 days after.

Next, we will capture utilization data from the identified patients who test positive and patients who test negative during 3 months after a corresponding index date in 2019. We will also estimate and describe the trajectory of weekly counts (i.e., counts over time) of utilization in 2019 (combined and stratified by outpatient, inpatient, ED, virtual settings) at each participating site, and by age, sex, and race/ethnicity.

Based on the index date, which we will initially define as the date 2 weeks following the date of positive or negative SARS-CoV-2 laboratory test, we will define a 3-month post-index time window and a 3-month pre-index time window for each patient. The 3 months in the pre-period will correspond to those in the post-period: for example, if a patient tests positive for COVID-19 on May 1, 2020, the post-index date period will cover May 15, 2020 – August 15, 2020. The pre-index date period will be May 15, 2019 –

August 15, 2019. This logic will also be applied to patients who test negative, anchored on their index date.

During the 3-month pre- and post-index date periods of both patients who test positive and patients who test negative, we will explore the utilization trajectory and possible secular clusters after COVID-19 infection. We will describe the clinical and demographic characteristics of patients who test positive and patients who test negative.

Aim 1b.) ***Use an interrupted time series (ITS) approach to estimate the difference in changes in utilization, by healthcare setting (i.e., stratified by outpatient, inpatient, ED, virtual; and outpatient compared to inpatient), in the pre- and post-period for patients who test positive versus patients who test negative.***

For this aim we will conduct ITS analyses, as illustrated below. Specifically, we plan to use this design because inherent adjustment for secular confounding, including various restrictions on healthcare utilization, other community lockdown measures, and changing rates of SARS-CoV-2 infection over time are built into the design. First, we will start with a Difference-in-Difference (DiD) approach. It is the simplest approach in the ITS family and will be an informative first step, comparing the average encounter rate during the assessment period[13]. The illustration of interrupted time series design and parameter estimates can be found in **Figure 1** and **Table 1.** ITS may model a linear trajectory or even a smoothing spline curve for utilization rate; we will therefore plot encounter rates over time and explore secular trends prior to determining the final model.

Once the final model is identified, we will conduct stratified analyses by individual setting (outpatient, inpatient, ED, virtual), and by inpatient compared to outpatient setting.

Figure 1. Difference-in-Difference Approach as a Special Case of the Interrupted Time Series Design with a Control Group

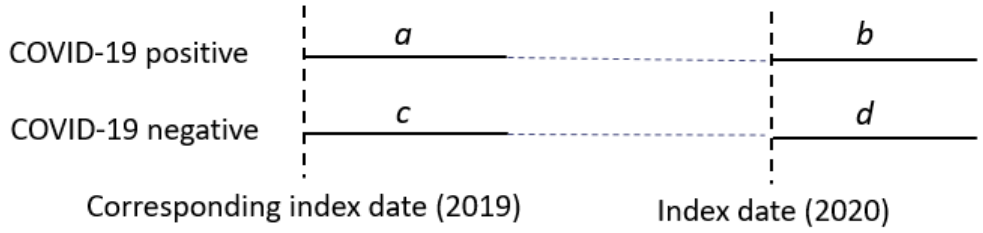


Table 1. Parameter Estimates for the Difference-in-Difference Approach

	Year 2019	Year 2020	Rate Ratio
COVID-19 positive	*a*	*b*	$rr_1 = b/a$
COVID-19 negative	*c*	*d*	$rr_2 = d/c$
Ratio of Rate Ratio			$rrr = bc/ad$

Aim 1c.) ***Use ITS approach to estimate the difference in changes in utilization (all settings combined) in the pre- vs. post-period for patients who test positive versus patients who test negative, stratified by age, sex, race/ethnicity and clinical characteristics of interest (i.e. comorbidities, prior influenza and/or pneumococcal vaccination, etc.).***

This aim will provide high-level insights into populations that may be at higher risk for LCS. We will conduct similar ITS analyses as **Aim 1b** stratified by race/ethnicity, age group, sex, and other limited variables, depending on statistical power. We will have a flag for care setting if additional analyses are needed by setting.

Aim 1d.) ***Use ITS approach to estimate the difference in changes in utilization (all settings combined) in the pre- compared to post-period for patients who test positive versus patients who test negative, stratified by LCS type (i.e. neurologic, respiratory, gastrointestinal, cardiac, psychiatric, other).***

We will conduct analyses for 3 and 6-month follow-up periods. We will examine longer time periods if data are available.

Similar to **Aim 1c**, this aim will provide high-level insights into the burden and frequency of each type of LCS. We will first use a list of 46 LCS outcomes that includes conditions such as arrhythmias, autonomic dysfunction, depression, etc. We may group these conditions into broader categories based on the affected organ system such as neurologic, respiratory, gastrointestinal, cardiac, psychiatric, and other. To define utilization by type, we will identify all visits that have at least one ICD-10 code that falls within each type. For example, for autonomic dysfunction, we will define all visits (from all settings) that have at least one ICD-10 code related to autonomic dysfunction in the pre- and post-periods for both patients who test positive and patients who test negative. We will describe the clinical and demographic characteristics of those with each type of COVID-19 LCS and then conduct similar ITS analyses as **Aim 1b**, stratified by LCS type. We will have a flag for care settings if additional analyses are required by setting. For analysis of 6-month follow-up period, we will conduct a sensitivity analysis by excluding the follow-up time after receiving the COVID-19 vaccine.

Phase II (THIS PHASE IS ON PAUSE FOR NOW)

Aim 2) ***Assess risk of LCS by severity of COVID-19 disease.***

Conduct retrospective matched cohort analyses, stratified by those hospitalized with COVID-19 and those admitted to the ICU with COVID-19.

For **Phase II,** we will assess the excess utilization (i.e., LCS) associated with COVID-19 by severity of COVID-19 disease. In other words, we will assess the additional burden of utilization associated with COVID-19 beyond that which might occur in the same settings without COVID-19. For these analyses, the exposure of interest will be COVID-19 disease, and the outcome of interest will be utilization (all settings combined).

We will create two separate cohorts: inpatient and ICU cohorts. We elected to not include an outpatient cohort for this aim because we felt there was considerable overlap with **Aim 1** analyses (most cases in outpatient setting), and our primary interest is to disentangle the effects of the hospitalization or ICU

visit itself with that of COVID-19. In each cohort, the exposed group will include patients with COVID-19, and the unexposed group will include those without COVID-19 documented by the index date. For example, the ICU cohort will compare adjusted rates of utilization 3 months after the index date in those with COVID-19-associated ICU stays versus utilization rates in patients with ICU stays due to reasons other than COVID-19. For the unexposed groups, we will include patients that may not have been part of the control groups for **Phase I**.

The index date for each of the two analyses will be defined as the date of discharge for the first hospitalization with a COVID-19 positive laboratory test (within 7 days +/- date of admission) for the inpatient cohort, and the date of discharge for the first ICU stay with a COVID-19 positive laboratory test (within 7 days +/- date of admission) for the ICU cohort. We will capture all utilization (i.e., outpatient, inpatient, virtual) in the 3 months following the index date for all cohorts. To assess comorbidities and other covariates, we will require continuous membership (allowing for a 31-day gap) from 12 months prior to the index date.

Similar to **Phase I**, the cohort will include those who had a COVID-19 positive test from March 1, 2020 to November 1, 2020. The follow-up time of those in the unexposed cohort will be censored if and when they develop COVID-19.

To create the unexposed comparison groups, we will identify patients without COVID-19 matched up to 1:5 on age (+/- 1 year), race/ethnicity, sex, timing of discharge date for the hospitalization/ICU (+/- 1 week), and discharge to skilled nursing facility/long term care facility (SNF/LTCF).

To account for the possibility that the characteristics which predict the likelihood of COVID-19 disease may also be associated with utilization post-discharge, we will use propensity score analyses with inverse probability of treatment weighting (IPTW) to balance covariates across exposure groups and adjust for confounding characteristics:

1. First, we will estimate the probability of exposure.
2. Next, the weight for each patient will be calculated as the inverse of the predicted probability of receiving their own exposure and will be normalized by dividing by the mean weight of each group.
3. Following weighting, standardized differences will be used to assess whether balance of covariates is achieved between the comparison groups.
4. Finally, we will use weighted Poisson regression models with robust error variance to estimate the rate ratio of utilization associated with COVID-19.

Note: Phase II approach may be modified based on **Phase I** findings.

Potential Limitations

There are potential limitations to this study. First, we make assumptions that utilization occurring in all settings in the 3- and 6-months post COVID-19 disease are potentially related to or attributable to COVID-19, as we do not require a COVID-19 diagnosis code at every encounter. We do not believe that coding is occurring reliably across settings and across time. To address this concern, we use ITS and matched cohort designs to assess the incremental utilization associated with COVID-19. Second, we

assume that comorbidity status in the 12 months prior to index date will be similar to the distribution in the post-period, although status may change during this time period. Third, the length of LCS may differ by system or symptoms. Further, LCS of a particular type may be longer in time than our analyses allow, given limitations on follow-up time. Fourth, there might be a delay from symptom onset to appointment time, so we may overestimate the duration of symptoms, as reflected by healthcare utilization. Fifth, some patients with COVID-19 may not be diagnosed, resulting in misclassification. Sixth, while CDC includes smoking and pregnancy as potential risk conditions for COVID-19, we do not include these in our study. Finally, there is no ICU file in the VSD infrastructure data. We would need to request that each site create an ICU ancillary file. Most VSD sites have experience identifying ICU admissions using automated data sources.

Data Management

Data files: Constant, Enroll, Inpatient, Outpatient, Vaccine from VSD Dynamic Data Files (DDF) or cycle files, ancillary COVID-19 DxID and COVID-19 lab data, and ancillary Height-Weight and ICU files.

Chart review: We will likely conduct preliminary chart abstractions at KPSC (approximately 100 reviews total) as part of this study, primarily to determine whether providers attribute the neurologic/respiratory/gastrointestinal/cardiac/psychiatric visits to COVID-19. If needed, we can expand the chart review to additional sites.

Data management: The VSD team at KPSC will be responsible for data management activities, including data extraction and consolidation between sites, study documentation, and archival. KPSC will store all electronic documents, KPSC data sets, and files relevant to the project on KPSC computers, which have restricted access. Currently, we propose this study as a multi-site study. SAS programs will be developed at KPSC and sent to participating sites for approval prior to data extraction. Site data managers will be responsible for developing an ancillary Height-Weight and ICU files.

Human Subjects Protection

Human subjects: Privacy and confidentiality will be strictly protected according to VSD standard procedures. There will be minimal risks to patient privacy and confidentiality. All information will be stored on secure KPSC computers and at participating sites. This study will be covered under KPSC's umbrella VSD IRB approval, which includes a waiver for the requirement to obtain HIPAA authorizations. Participating sites may seek IRB approval as needed.

Estimated Timeline for Phase I*

December 2020	Discuss concept on December VSD project call; submit full proposal to CDC and sites for comment and invite sites to participate
March 2021	Revise proposal based on reviewer comments; circulate Distributed Data Model (DDM) programs for review and approval
March 2021	Creation of ancillary Height-Weight file at participating sites
March-April 2021	Data extraction and cleaning, limited chart review
April-May 2021	Data analyses
June 2021	Manuscript preparation
June -July 2021	Submit manuscript to CDC for clearance; submit manuscript for publication

*__Phase II__ may start prior to completion of __Phase I__.

Estimated Timeline for Phase II (THIS PHASE IS ON PAUSE FOR NOW)

TBD	Circulate DDM programs for review and approval
TBD	Creation of ancillary ICU file at participating sites
TBD	Data analyses
TBD	Manuscript preparation
TBD	Submit manuscript to CDC for clearance; submit manuscript for publication

References

1. Carfi A, Bernabei R, Landi F, Gemelli Against C-P-ACSG. Persistent Symptoms in Patients After Acute COVID-19. *JAMA.* 2020;324(6):603-605.
2. Carvalho-Schneider C, Laurent E, Lemaignen A, et al. Follow-up of adults with noncritical COVID-19 two months after symptom onset. *Clin Microbiol Infect.* 2021;27(2):258-263.
3. Datta SD, Talwar A, Lee JT. A Proposed Framework and Timeline of the Spectrum of Disease Due to SARS-CoV-2 Infection: Illness Beyond Acute Infection and Public Health Implications. *JAMA.* 2020;324(22):2251-2252.
4. Del Rio C, Collins LF, Malani P. Long-term Health Consequences of COVID-19. *JAMA.* 2020;324(17):1723–1724
5. Halpin SJ, McIvor C, Whyatt G, et al. Postdischarge symptoms and rehabilitation needs in survivors of COVID-19 infection: A cross-sectional evaluation. *J Med Virol.* 2021;93(2):1013-1022.
6. Huang C, Huang L, Wang Y, et al. 6-month consequences of COVID-19 in patients discharged from hospital: a cohort study. *Lancet.* 2021;397(10270):220-232.
7. Toy S, Reddy S, Hernandez D. Doctors Begin to Crack Covid's Mysterious LongTerm Effects; Severe fatigue, memory lapses, heart problems affect patients who weren't that badly hit initially; 'It's been so long'. *Wall Street Jounal* Nov 1, 2020.
8. Tenforde MW, Kim SS, Lindsell CJ, et al. Symptom Duration and Risk Factors for Delayed Return to Usual Health Among Outpatients with COVID-19 in a Multistate Health Care Systems Network - United States, March-June 2020. *MMWR Morb Mortal Wkly Rep.* 2020;69(30):993-998.
9. Centers for Disease Control and Prevention. Late Sequelae of COVID-19. *COVID-19* 2020; **https://www.cdc.gov/coronavirus/2019-ncov/hcp/clinical-care/late-sequelae.html**. Accessed January 22, 2021.
10. Taquet M, Luciano S, Geddes JR, Harrison PJ. Bidirectional associations between COVID-19 and psychiatric disorder: retrospective cohort studies of 62 354 COVID-19 cases in the USA. *Lancet Psychiatry.* 2021;8(2):130-140.
11. Hechter RC, Qian L, Yan S, Luo Y, Krishnarajah G, Tseng HF. Impact of the change of copay policy in Medicare Part D on zoster vaccine uptake among Medicare beneficiaries in a managed care organization. *BMC Health Serv Res.* 2017;17(1):503.
12. Maley JH, Brewster I, Mayoral I, et al. Resilience in Survivors of Critical Illness in the Context of the Survivors' Experience and Recovery. *Ann Am Thorac Soc.* 2016;13(8):1351-1360.

13. Lopez Bernal J, Cummins S, Gasparrini A. Difference in difference, controlled interrupted time series and synthetic controls. *Int J Epidemiol.* 2019;48(6):2062-2063.