COVID-19 Natural History Master Protocol

W. Katherine Yih, PhD, MPH,[1] Wei Hua, MD, PhD, MHS, MS,[2] Christine Draper, BA,[1] Sarah Dutcher, PhD, MS,[2] Candace Fuller, PhD, MPH,[1] Maria Kempner, BA,[1] Brian Kit, MD, MPH,[2] Jennifer Lyons, PhD, MPH,[1] Meighan Rogers Driscoll, MPH,[1] Darren Toh, ScD,[1] Vincent Lo Re III, MD, MSCE[3]

Affiliations: 1. Department of Population Medicine, Harvard Pilgrim Health Care Institute and Harvard Medical School, Boston, MA; 2. Office of Surveillance and Epidemiology, Center for Drug Evaluation and Research, US Food and Drug Administration, Silver Spring, MD; 3. Division of Infectious Diseases, Department of Medicine and Center for Clinical Epidemiology and Biostatistics, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA.

Version 3.0

October 9, 2020

The Sentinel System is sponsored by the U.S. Food and Drug Administration (FDA) to proactively monitor the safety of FDA-regulated medical products and complements other existing FDA safety surveillance capabilities. The Sentinel System is one piece of FDA's Sentinel Initiative, a long-term, multi-faceted effort to develop a national electronic system. Sentinel Collaborators include Data and Academic Partners that provide access to healthcare data and ongoing scientific, technical, methodological, and organizational expertise. The Sentinel Coordinating Center is funded by the FDA through the Department of Health and Human Services (HHS) Contract number HHSF223201400030I.

COVID-19 Natural History Master Protocol

Table of Contents

History of Modifications

Version	Date	Modification	Author
1.0	08/31/2020	Original Version	As shown on title page
1.1	09/29/2020	First revision	As shown on title page
2.0	10/2/2020	Second revision (minor)	As shown on title page
3.0	10/9/2020	Third revision (minor)	As shown on title page

A. Introduction

Coronavirus disease 2019 (COVID-19) is an emergent acute viral infection attributable to the novel severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) that primarily affects the respiratory system.[1,2] While the proportion of asymptomatic infections remains uncertain, clinical manifestations range from a mild, self-limiting respiratory tract illness to severe progressive pneumonia, multiorgan failure, and death.[2-12] Understanding of COVID-19 and its prognostic factors is evolving. Unadjusted analyses of patients with COVID-19 in China,[4-6,8,11,13] Italy,[14] and the United States[15,16] suggest case fatality rates are higher in the elderly and those with certain pre-existing conditions, and clinical experience points to the importance of timely supportive care for preventing progression to severe disease and death. However, published studies have evaluated predominantly hospitalized patients, have examined few determinants and outcomes of COVID-19, and frequently did not adjust for important confounding variables. As a result, there are major knowledge gaps on the epidemiology of COVID-19 in the United States, especially in certain subgroups, such as children, certain ethnic groups, and pregnant women. Population-representative studies evaluating the epidemiology of COVID-19 are therefore sorely needed to provide insights into its mechanisms and consequences and inform the development of interventions to reduce the risk of adverse outcomes. The Sentinel System affords the unique opportunity to create large, well-characterized, national cohorts of patients diagnosed with COVID-19.[17] Analyses of these cohorts could address key knowledge gaps to improve the understanding of the characteristics and outcomes of COVID-19 patients.

B. Objectives

This protocol presents guidance for identifying cohorts of patients diagnosed with COVID-19 in the Sentinel System, delineates the variables that might be relevant to such analyses, assesses the feasibility with which these variables or data elements can be collected within Sentinel, lists potential limitations, and is accompanied by code lists for the data elements.

The kinds of uses to which these cohorts can be put include:

1. Describing the course of illness and COVID-19-related outcomes in various demographic groups and examining changes over time
2. Determining prognostic factors for COVID-19 based on data available early in the course of illness
3. Determining the incidence of and risk factors for important COVID-19-related complications, such as arterial or venous thrombotic complications, multisystem inflammatory syndrome in children (MIS-C), and bacterial superinfections
4. Studying the effectiveness and safety of interventions to treat COVID-19 and its complications
5. Providing a benchmark, or serving as an external control, for single-arm trials of COVID-19 treatments in relevant patient populations

The specific study questions to be addressed during implementation will have implications for which data elements are needed, which Data Partners may be appropriate to include, and which methods should be used to develop the cohorts and datasets. In implementing this Master Protocol to address specific study questions, coordination with related work conducted under other work orders (e.g., COVID-19 International Classification of Diseases, Tenth Revision, Clinical Modification [ICD-10-CM] diagnosis code validation; incidence and determinants of arterial and venous thrombotic events in COVID-19) is recommended in order to share approaches and avoid duplication of effort.

C. Methods

1. Data sources

Data sources may include claims-based systems, electronic health record (EHR)-based systems (which can be further differentiated into integrated care delivery systems and EHR data aggregators), and systems with linked EHR and claims data. Sentinel's claims-based Data Partners provide longitudinal data on large populations, with well-established data updating and quality assurance procedures, and are used for medical product safety and effectiveness studies.[17] However, for COVID-19 natural history studies, data on vital signs, laboratory results, inpatient medication exposure, and other details of clinical care will be important. EHR systems are expected to contain much more comprehensive data on those features than claims-based systems. A significant caveat is that these kinds of data elements are often not in a consistent format either across or within Sentinel Data Partners and will require considerable effort to clean, standardize, and check. Of note, COVID-19 diagnostic testing and serologic antibody testing will be incorporated into the Sentinel Common Data Model and will be available from both claims-based and EHR-based Data Partners in the future.

Data Partners will be selected during the implementation of each proposed analysis to create cohorts and datasets based on the availability of data elements crucial to the specific question to be studied and other criteria. In using claims-based data sources, enrollment requirements can be selected to ensure capture of medically attended events during specified periods before or after COVID-19 diagnosis. In EHR-based data sources, information about enrollment is not available but researchers can, for example, require eligible patients to have at least one encounter during the specified baseline period to increase the likelihood of capturing key co-morbidities recorded in the same EHR system.

We expect to use variables available in the Sentinel Common Data Model (SCDM), as well as variables in the source data that are currently not in the SCDM. The SCDM captures the values found in the source data. For example, an ICD-10-CM diagnosis code in the source data will show up as the same ICD-10-CM code in the SCDM. When necessary, mapping to standard vocabularies is done and is transparent and traceable. Regardless of whether the SCDM is used, steps will be taken to ensure that the same variable (e.g., a pre-existing comorbidity) is defined in a standardized and consistent way across data sources.

2. COVID-19 case definition

A presumptive case of COVID-19 will be defined broadly, as having any of a set of COVID-19-related ICD-10-CM diagnosis codes *or* any positive result of an eligible laboratory test. Eligible laboratory tests include a reverse transcriptase polymerase chain reaction (RT-PCR) or other nucleic acid amplification test (NAAT) for SARS-CoV-2 or a SARS-CoV-2 antigen test.[18-20] More specific case definitions, e.g., having a positive NAAT result, can be applied in particular implementations of this protocol.

ICD-10-CM diagnosis codes

A validation study of COVID-19 case-finding algorithms consisting of subsets of the ICD-10-CM codes in **Table 1** below is underway, restricted to hospitalized cases and using positive NAAT results for SARS-CoV-2 as the gold standard for confirmation. The ICD-10-CM codes to be used to identify presumptive COVID-19 cases for natural history studies should be selected in light of the results of that validation study, especially positive predictive values.

Table 1. ICD-10-CM diagnosis codes for COVID-19 that have been selected for validation based on positive NAAT result for SARS-CoV-2.

ICD-10-CM code	Description
B97.29	Other coronavirus as the cause of diseases classified elsewhere
U07.1	COVID-19, virus identified [code effective April 1, 2020]
B34.2	Coronavirus infection, unspecified site
B97.21	SARS-associated coronavirus as the cause of diseases classified elsewhere
J12.81	Pneumonia due to SARS-associated coronavirus

A more inclusive approach may be warranted to identify potential cases without laboratory confirmation in the outpatient setting. CDC coding and reporting guidelines for April 1-September 30, 2020 state, "If the provider documents "suspected," "possible," "probable," or "inconclusive" COVID19, do not assign code U07.1. Assign a code(s) explaining the reason for encounter (such as fever) or Z20.828, Contact with and (suspected) exposure to other viral communicable diseases."[21]

SARS-CoV-2 laboratory tests

At present, several different tests are being implemented in clinical practice to identify current or past SARS-CoV-2 infection. These include: 1) RT-PCR or other NAAT assays of respiratory tract specimens, 2) antigen detection assays of respiratory tract specimens, and 3) serologic tests that identify antibodies (such as Immunoglobulin M and Immunoglobulin G) to SARS-CoV-2 within blood or saliva.[18-20] Laboratory codes for SARS-CoV-2 and other Data Partner-specific mnemonics can be used in EHR systems to identify health plan members with positive test results.

For the laboratory component of the presumptive COVID-19 case-finding algorithm, only positive results from a NAAT or antigen assay (#1 or #2 above) should be considered to classify COVID-19 status at this time. Patients with a positive result for SARS-CoV-2 by NAAT or antigen test in any setting will meet the broad case definition.

3. Index date definition

In general, in establishing a COVID-19 cohort, a patient's index date will be defined as the date of specimen collection for the patient's first positive eligible laboratory test for SARS-CoV-2 or the date of the first eligible COVID-19 ICD-10-CM diagnosis. For Data Partners with multiple sources of information, the index date will be defined as the earlier of these dates. Date of symptom onset will be unavailable or inaccurate for the vast majority of COVID-19 cases.

However, the specific study question addressed during implementation may dictate the selection of an alternative index date, for example, hospital admission date or date of initiation of a treatment of interest.

4. Inclusion/exclusion criteria

Demographic characteristics: Any age, any sex

Earliest index date (i.e., start date for identifying COVID-19 cases): January 20, 2020 (date of first laboratory-confirmed case of COVID-19 in the United States[22])

Latest index date: To be determined during implementation. For example, if at least 4 weeks of follow-up are required in order to study a particular question, the last eligible index date should be specified as 28 days prior to the last available date in each participating Data Partner's data.

Enrollment requirements *(needed only for certain kinds of analyses, available in claims-based data sources)*: Pre- and post-index date enrollment requirements to be determined in light of specific study question. Typically, apparent enrollment gaps of up to 45 days are allowed. As mentioned previously, in EHR-only data sources, patients can be required to have at least one encounter during a specified look-back period, roughly approximating an enrollment requirement.

5. Follow-up

For maximum utility and flexibility, follow-up will be open-ended, with no predefined censoring event prior to death, to allow assessment of the entire course of illness, including short- and long-term outcomes and reinfection. Certain study questions may require more narrowly defined follow-up periods.

6. Guidance on formation of cohorts

a. Inclusive cohort

An inclusive cohort of presumptive COVID-19 cases meeting the criteria above will be identified during implementation. This cohort will not be restricted with respect to age, sex, race/ethnicity, setting of diagnosis (e.g., ambulatory, hospital, nursing home), pregnancy status, exposure to medications of interest, or any feature other than those specified in the sections above. This approach will provide an integrated view of the epidemiology of COVID-19, capturing patients' transitions from one setting to another, e.g., from ambulatory care to hospital setting, from unmedicated to initiation with Drug X, etc. The unbounded follow-up time will allow longer-term outcomes in any medical setting to be captured, including those occurring after discharge from the hospital. From this inclusive cohort, subcohorts can then be identified, as needed, to study the natural history of the disease or related topics in relevant subgroups—for example, hospitalized children, pregnant women, adults treated with a particular medication of interest, etc. Identification of certain subcohorts may require the use of data from certain Sentinel Data Partners and exclude the use of others. For example, identifying a subcohort of patients receiving Drug X may require access to data on in-hospital medications, which will not be available within claims data sources.

Identifying a subcohort of pregnant women will likely be more involved than for most other subcohorts of interest, particularly as it will be desirable to estimate gestational age. Considerations regarding pregnancy are discussed in greater detail below.

In accordance with the Sentinel System's distributed data model, the datasets for the cohort and subcohorts will reside behind the firewalls of the respective participating Data Partners.

b. Pregnant women

The identification of a pregnant subcohort within the inclusive cohort of COVID-19 patients will depend on the data sources used. Data sources vary in their ability to identify both pregnant patients and pregnancy episodes, with additional differences in capabilities for describing baseline characteristics; gestational age at COVID-19 diagnosis; and maternal, pregnancy, and infant outcomes. In Sentinel, pregnancies can be identified via general or specific code lists depending on whether one wishes to: 1) create pregnancy episodes by

estimating pregnancy start among women with a live-born infant, or 2) identify a potential pregnancy regardless of pregnancy outcome or whether pregnancy start can be estimated.

1) Creating pregnancy episodes by estimating pregnancy start among women with a live-born infant

In this approach, Sentinel uses a curated code list for live deliveries, from which gestational age at delivery is estimated via a previously developed algorithm (**Appendix 1**). Gestational age at COVID diagnosis and maternal and infant outcomes can be evaluated. Patient demographics, health conditions, medical product use, and other characteristics during the pre-pregnancy period and each trimester can also be identified. Certain characteristics, such as timing of COVID-19 diagnosis, can be reported by estimated gestational week. Maternal outcomes of COVID-19 can be assessed for all live-birth pregnancy episodes. All Sentinel Data Partners, including claims-based insurers and integrated delivery systems, have data that can be used to assess maternal outcomes. For infant outcomes, a subset of pregnancy episodes identified by the algorithm can be linked to the infant record as part of the SCDM Mother-Infant Linkage (MIL) Table (**Appendix 2**). Four Sentinel Data Partners, including national insurers and a smaller regional insurer, populate and refresh the MIL Table. Approximately 70% of live birth deliveries identified via the algorithm are successfully linked to an infant.

2) Identifying potential pregnancy regardless of pregnancy outcome or whether pregnancy start can be estimated

In this second approach, which does not require evidence of a live birth, a potentially pregnant cohort can be identified via investigator-defined code lists referencing prenatal tests, exams, or procedures; delivery codes; or other pregnancy outcomes.[23] With this method, we are unable to determine gestational age with the existing tool and therefore cannot with certainty describe the timing of characteristics or events, such as COVID-19 diagnosis, in relation to pregnancy start. The method for identifying pregnant patients and pregnancy episodes will depend on the data source. For instance, in a hospitalization-based data source, such as Hospital Corporation of America (HCA), pregnant COVID-19 patients can be identified via diagnosis and procedure codes as long as a pregnancy or delivery code is recorded for the same hospitalization as a COVID-19 code. But patient and pregnancy characteristics, including outcomes that occur outside of the hospitalization of interest, may not be observable. TriNetX, a global health research network that includes EHR data, can also identify pregnant patients using diagnosis and procedure codes. Timing of COVID-19 diagnosis and other pregnancy characteristics can be observed in relation to the index code of interest; however, in the TriNetX platform, this code may not identify pregnancy start and only refers to the first identified pregnancy in patients with multiple pregnancies. Availability of pregnancy data in other systems, whether claims-based, EHR-based, or integrated delivery systems, depends on the data elements and capabilities of the specific data source.

These approaches will potentially support FDA's participation in the ICMRA pregnancy study under development (http://www.icmra.info/drupal/en/news/22july2020/summary).

7. Data elements

Data elements deemed important for describing the clinical characteristics, course, complications, and outcomes of COVID-19 and identifying risk factors or prognostic factors related to these complications and outcomes are presented in the accompanying document,

"COVID-19 nat hx data elements 2020-09.xlsx." More variables can be added to this list as needed in order to address specific study questions.

"COVID-19 nat hx data elements 2020-09.xlsx" includes assessment of the availability of each data element in claims-based and EHR-based data systems. Three tiers of data availability have been defined. Tier 1 data are variables that are commonly coded or otherwise fairly easy to collect. In contrast, Tier 3 data are variables that are not readily available and are challenging to ascertain, such as those requiring manual or technology-enabled medical record review. Tier 2 data represent variables that do not require medical record review but have not to date been coded or routinely used by Sentinel and would require work to standardize and clean.

In **Table 2** below, we present the general rubrics of variables to be collected or calculated, with caveats (particularly about availability) and comments about temporal aspects.

Table 2. Caveats and temporal aspects of categories of data elements

Rubric	Examples	Caveats/limitations	Temporal aspects
Demographics	Age, sex, ethnicity, geographic location	Race/ethnicity often missing from claims	Collect most recent values prior to or on index date
Medical history, pre-existing conditions	Smoking, alcohol use, chronic conditions	Substance use under-captured; some conditions might be missed in look-back period; EHR-only data sources (not linked to claims) do not include enrollment data so uniform look-back periods to capture medical history and co-morbidities cannot be employed	Collect evidence of each condition at any time during specified look-back period prior to index date (i.e., baseline period)
Medications for pre-existing conditions	Antiviral therapy, angiotensin converting enzyme (ACE) inhibitor/angiotensin receptor blocker (ARB), immunosuppressive drug therapy	Over-the-counter medications, such as non-steroidal anti-inflammatory drugs (NSAIDs), will be missed (except in EHR data if administered during hospitalization); duration of inpatient medications after index date probably not reliably ascertainable—illness or hospitalization might affect use of usual medications	Ensure days' supply overlaps index date to avoid capturing terminated prescriptions
Onset timing	Time of any symptom onset	Important for studying COVID-19 natural history but likely often absent or inaccurate	Seek to obtain absolute date or date relative to index date

Rubric	Examples	Caveats/limitations	Temporal aspects
Signs, symptoms	Fever, cough, headache; symptoms can be used in algorithms to identify some conditions (e.g., MIS-C)	Absence of codes for a COVID-19 symptom does not mean symptom was absent; mild symptoms may be undercoded; recording of signs and symptoms likely inconsistent across institutions, individual coders, and time	Obtain on index date
Vital signs	Respiratory rate, heart rate, blood pressure, oxygen saturation	Vital signs not available in claims data (unless linked to another data source); if present in EHR data, they require standardization and cleaning	Collect at multiple points, with dates, to help characterize disease progression
COVID-19 diagnosis	Positive SARS-CoV-2 NAAT or antigen test	Claims data may not include test results; in some Data Partners, LOINC codes or other hospital-/laboratory-specific naming strategies can be used to identify specific tests and results but requires custom work, standardization, and cleaning	Either use once to confirm case or consider collecting at multiple points, with dates, to help characterize virus persistence
Laboratory results	Blood count, electrolytes, liver function tests	Claims data do not include laboratory results (unless linked to another data source); laboratory data require standardization and cleaning	Consider collecting at baseline or at multiple points (e.g., first and last), with dates, to help characterize disease severity and progression
Chest radiographic imaging	Reports from chest x-rays, computed tomography, and magnetic resonance imaging	These data include reports from chest radiographic imaging. Data from these reports can only be accessed via manual review of reports. These reports are prepared by the radiologist from each ambulatory radiology center or hospital.	Consider collecting reports at COVID-19 diagnosis or at time points after diagnosis.

Rubric	Examples	Caveats/limitations	Temporal aspects
Clinical progression	Non-pharmaceutical treatments and procedures, such as oxygen support, ICU admission; 7-category ordinal scale	Oxygen support (other than, possibly, mechanical ventilation) undercoded; duration since symptom onset likely unavailable or inaccurate; other durations likely unavailable, especially in claims data, in part because inpatient procedures occurring during hospitalization may be rolled up to admission date, but also because cessation of certain procedures /treatments may not be recorded; categories of ordinal scale may not be available, especially towards less severe end of scale	Collect date of procedure of interest (e.g., oxygen support, ICU admission); durations (e.g., from hospital admission to ICU admission) can be calculated if dates available; use ordinal scale to define status at time points of interest to help categorize improvement or deterioration
COVID-19-related complications and outcomes	Respiratory failure, shock, thrombotic events, bacterial infections, multi-system inflammatory syndrome	Some conditions of concern (e.g., multi-organ dysfunction[24]) do not have specific codes, so misclassification likely; some conditions may only be ascertained from progress notes	Collect at first occurrence, with date; include all settings as some events might occur in outpatient setting, whether or not patient previously hospitalized
Improvement of signs, symptoms	Increased oxygen saturation, fever normalization	Vital signs not available in claims data (unless linked to another data source); if present in EHR data, they require standardization and cleaning; assessing improvement requires comparing values from at least two time points	Collect at multiple points (or just first and last), with dates, to be able to document improvement
COVID-19-specific treatments, supportive medications, and medications for complications	Remdesivir, convalescent plasma, vasopressors, antithrombotics	Capture of treatments and their timing depends on where treatment is administered and type of treatment; in claims data, inpatient medication use is often not well captured; durations of treatment may require medical record review	Collect dates and treatment durations to help characterize initiation, switching, concomitant use, and other drug use patterns

Rubric	Examples	Caveats/limitations	Temporal aspects
Disposition	Live discharge, readmission to hospital, death	If death not medically attended, capture and timeliness of capture will vary	Collect dates of events to calculate durations
Pregnancy, gestational age		In Sentinel, pregnancy episodes can be determined retrospectively from live births but *not if no live birth resulted*; gestational stage by *trimester* quite accurate but gestational age by *week* less so Potential pregnancies may also be identified by a generalized code list referencing any prenatal test, exam or procedure. This approach will identify more potential pregnancies than using live birth delivery codes alone; however, weeks of gestation cannot be estimated.	From codes associated with live births, gestational age of pregnancy at index date can be calculated
Pregnancy outcome	Maternal outcomes, such as pre-eclampsia, pregnancy-related thrombosis	Pregnancy characteristics and maternal outcomes can be described in all Sentinel claims-based Data Partners; however, non-live birth pregnancy outcomes (e.g., stillbirth, termination) cannot be evaluated using the current CIDA tool	In defining a general cohort of potential pregnancies that include non-live birth outcomes, cannot calculate gestational age or determine temporality relative to COVID-19 diagnosis
Pregnancy outcomes, continued	Outcomes for fetus/infant, such as intra-uterine growth retardation, fetal demise	As of 10/1/2020, can be determined by four Data Partners that use mother-infant linkage but only outcomes in *live* infants; determining relationship between COVID-19 and miscarriage or stillbirth affected by uncertainties in timing of pregnancy and of COVID-19	From codes associated with live births, weeks of gestation at index date can be calculated; condition of infant (e.g., re congenital conditions) can be assessed in some specified post-birth time window after linking infant to mother; some fetal anomalies can be diagnosed in utero via anatomic scan

Code lists have been developed for all variables (e.g., diagnoses, medications, and procedures) in "COVID-19 nat hx data elements 2020-09.xlsx" (**Appendix 3**). Information made available via the Reagan Udall Foundation Therapeutics Evidence Accelerator Parallel Analysis Workgroup was leveraged with a goal of aligning code lists (where appropriate) with those being used by other researchers. Code lists may need to be updated over time due to the rapid evolution of coding terminologies.

8. Classification of level of severity of COVID-19

Two schemes for classifying severity of COVID-19 cases are proposed, for studies requiring such classification. The first, as described in the FDA guidance "COVID-19: Developing Drugs and Biological Products for Treatment or Prevention Guidance for Industry,"[25] is based on the clinical concepts of the condition and requires access to certain vital signs data such as respiratory rate, heart rate, and oxygen saturation. This classification system is presented in the left column of **Table 3** below. The second scheme, in the right column, is derived from the first scheme and represents a trade-off between accuracy and feasibility. It does not use vital signs and can be used with data sources lacking ready access to such data. This "practical" definition has not yet been validated and is subject to change. To facilitate the interpretation of study results, researchers should consider evaluating the performance of the suggested practical definition against the FDA conceptual definition in the selected data source, time period, and geographic region.

The severity classification allows capture of patients' transitions from one state to another in the clinical course and can be used for measuring baseline status or selecting patients. To monitor disease trajectory or characterize disease progression as related to prognosis, more detailed information on time-varying vital signs and biomarkers may be needed.

Table 3. Two schemes for classifying severity of COVID-19 cases

#	Conceptual severity definition[25]	Suggested practical severity definition*
1	**Asymptomatic** [SARS-CoV-2 testing] Positive testing by standard RT-PCR assay or equivalent test AND [Symptoms] No symptoms in Row 2	**Asymptomatic or very mild** Positive laboratory test for SARS-CoV-2 (consistent with inclusion criteria of specific study) AND No symptoms from the Conceptual Definition, Row 2 list and no pneumonia AND No initiation of any oxygen therapy or hospitalization

#	Conceptual severity definition[25]	Suggested practical severity definition*
2	**Mild illness** [SARS-CoV-2 testing] Positive testing by standard RT-PCR assay or equivalent test AND [Symptoms] Could include fever, cough, sore throat, malaise, headache, muscle pain, gastrointestinal symptoms, without shortness of breath or dyspnea AND [No clinical signs indicative of Moderate, Severe, or Critical Severity]	**Mild illness** Positive laboratory test for SARS-CoV-2 (consistent with inclusion criteria of specific study) AND Any symptom from the Conceptual Definition, Row 2 list (*excluding* new-onset dyspnea) but no pneumonia AND No initiation of any oxygen therapy or hospitalization
3	**Moderate illness** [SARS-CoV-2 testing] Positive testing by standard RT-PCR assay or equivalent test AND [Symptoms] Could include any symptom of Mild illness or shortness of breath with exertion AND [Clinical signs] Such as respiratory rate ≥ 20 breaths per minute, saturation of oxygen (SpO2) > 93% on room air at sea level, heart rate ≥ 90 beats per minute AND [No clinical signs indicative of Severe or Critical Severity]	**Moderate illness** *[In contrast with Mild, must have dyspnea or pneumonia or hospitalization (without ICU)]* Positive laboratory test for SARS-CoV-2 (consistent with inclusion criteria of specific study) AND Any symptom from the Conceptual Definition, Row 2 list (*including* new-onset dyspnea) or pneumonia AND No initiation of any oxygen therapy or ICU admission or organ failure

#	Conceptual severity definition[25]	Suggested practical severity definition*
4	**Severe illness** [SARS-CoV-2 testing] Positive testing by standard RT-PCR assay or equivalent test AND [Symptoms] Could include any symptom of Moderate illness or shortness of breath at rest, or respiratory distress AND [Clinical signs] Such as respiratory rate ≥ 30 per minute, heart rate ≥ 125 per minute, SpO2 ≤ 93% on room air at sea level or PaO2/FiO2 <300 AND [No criteria for Critical Severity]	**Severe illness** *[In contrast with Moderate, must have initiation of any oxygen therapy (without intubation/IMV or ECMO) or admission to ICU]* Positive laboratory test for SARS-CoV-2 (consistent with inclusion criteria of specific study) AND Any symptom from the Conceptual Definition, Row 2 list (*including* new-onset dyspnea) or pneumonia AND Hospitalization or admission to ICU but no intubation/IMV or ECMO and no organ failure
5	**Critical illness** [SARS-CoV-2 testing] Positive testing by standard RT-PCR assay or equivalent test AND Evidence of critical illness, at least one of the following: [1. Respiratory failure] At least one: endotracheal intubation and mechanical ventilation, oxygen delivered by high-flow nasal cannula (heated, humidified, oxygen delivered via reinforced nasal cannula at flow rates > 20 L/min with fraction of delivered oxygen ≥ 0.5), noninvasive positive pressure ventilation, ECMO, or clinical diagnosis of respiratory failure (i.e., clinical need for one of the preceding therapies, but preceding therapies not able to be administered in setting of resource limitation) OR [2. Shock] Systolic blood pressure < 90 mmHg, or diastolic blood pressure < 60 mmHg or requiring vasopressors OR [3. Multi-organ dysfunction/failure]	**Critical illness** Positive laboratory test for SARS-CoV-2 (consistent with inclusion criteria of specific study) AND Organ failure recorded during a hospitalization or evidence of intubation/IMV or ECMO

Abbreviations: ECMO=extracorporeal membrane oxygenation; IMV=intermittent mandatory ventilation; RT-PCR=reverse transcriptase polymerase chain reaction; SARS-CoV-2=severe acute respiratory syndrome coronavirus 2

* The severity definition should be tailored to the study question, population, and time period of interest. Since the severity of a patient's illness can vary over time, the index date

and time period for assessing the criteria will depend on the study question. For example, a study designed to describe the characteristics of patients with ultimately asymptomatic COVID-19 might reasonably use an assessment period of 30 days after collection of the positive specimen, while a study evaluating the effectiveness of early treatment in then-asymptomatic patients might choose an assessment period of, say, the 2 weeks before the start of treatment.

Severity scales not requiring a positive laboratory test result can be developed for studies that will include cases identified via ICD-10-CM codes.

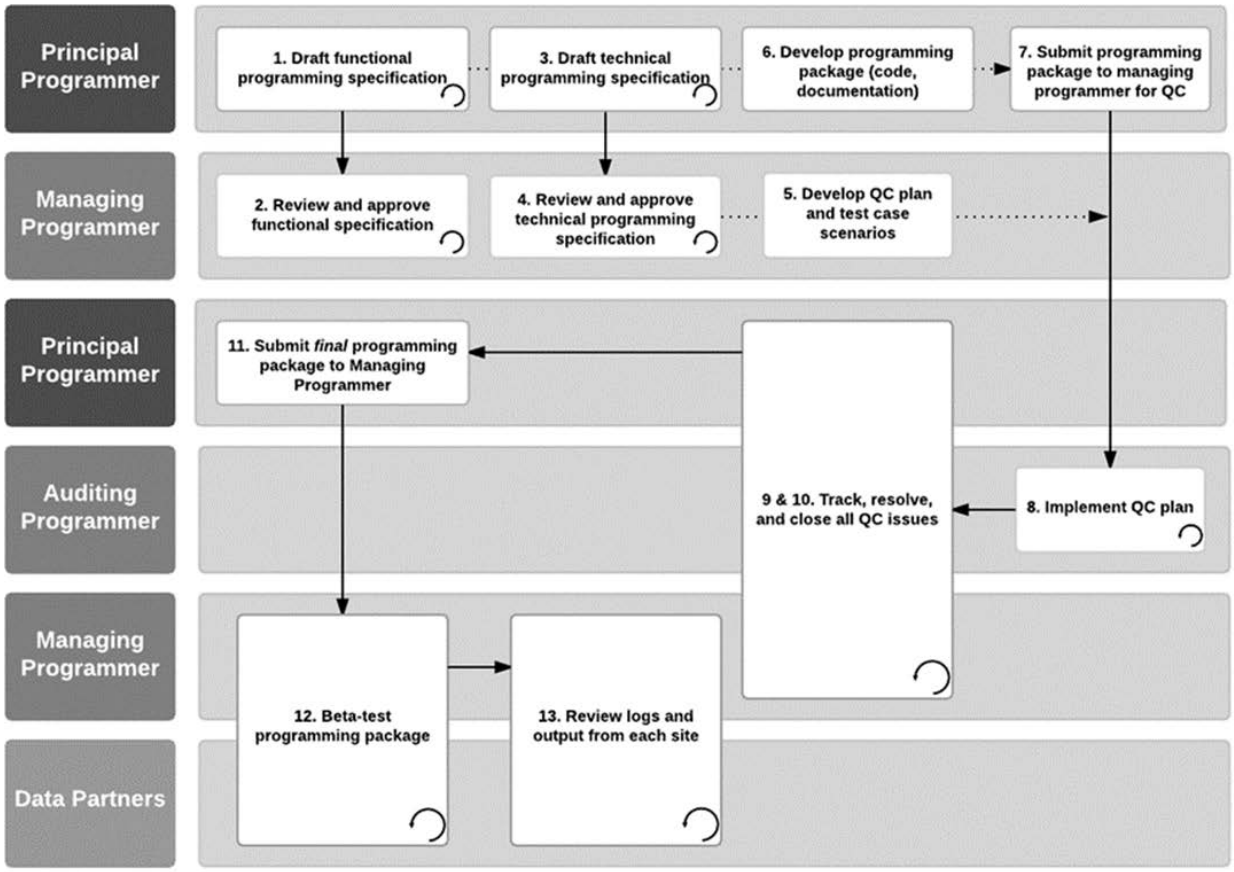
9. Data management and quality control

The Sentinel Operations Center will be responsible for writing and distributing SAS programs that can be used to evaluate data from participating Data Partners. The distributed network will allow Data Partners to maintain physical and operational control of their data while allowing use of the data to meet the study needs. The Sentinel Operations Center will maintain a secure distributed querying web-based portal to enable secure distribution of analytic queries, data transfer, and document storage. The system will meet all required State and Federal security guidelines for health data (e.g., Federal Information Security Management Act [FISMA], Health Insurance Portability and Accountability Act of 1996), specifically FISMA compliant for FISMA security controls as specified in the National Institute of Standards and Technology (NIST) Special Publication 800-53 (NIST and Joint Task Force Transformation Initiative 2017).

Sentinel's standard data quality assurance (QA) procedures will not be usable for all data elements and data sources since those QA procedures are conducted on data formatted in the Sentinel Common Data Model. Thus, ad hoc QA procedures will be developed for specific studies requiring the use of such data elements or data sources. The standard Sentinel quality assurance approach, which may be usable for other study questions, assesses consistency with the Sentinel Common Data Model, evaluates adherence to data model requirements and definitions, examines logical relationships between data model tables, and reviews trends in medical and pharmacy services use within and across Data Partners. Data curation will be consistent with guidance set forth by the FDA in its current recommendations for data quality assurance, specifically, "Guidance for Industry and FDA Staff: Best Practices for Conducting and Reporting Pharmacoepidemiologic Safety Studies Using Electronic Healthcare Data" (Guidance), section IV.E "Best Practices – Data Sources: QA and Quality Control", published in May 2013 (FDA 2013).

In addition to quality assurance of data elements, the Sentinel Operations Center adopts standard SAS programming quality assurance and quality control processes to check SAS programs and deliverables. **Figure 1** illustrates the Standard Operating Procedures for SAS programming quality assurance and quality control within the Sentinel System.

Figure 1: Standard Operating Procedures for SAS Programming Quality Assurance and Quality Control in the Sentinel System



(Circular arrow in some boxes indicates iterative process, incorporating feedback.)

10. Limitations to consider and methods to address

a. Misclassification:

Misclassification of COVID-19 status is possible. Within both claims and EHR data, cases of COVID-19 in the populations of the Data Partners may be missed because: 1) people with asymptomatic, very mild, or mild COVID-19 may not seek medical attention or testing; 2) COVID-19 patients may be unable to access SARS-CoV-2 testing (due to their personal situation or to the lack of testing in their area) or unwilling to undergo SARS-CoV-2 testing; or 3) false-negative SARS-CoV-2 assays may occur due to improper sampling procedures or test insensitivity. Conversely, non-COVID-19 cases may be misclassified as COVID-19 due to false-positive test results or ICD-10-CM COVID-19 diagnoses reflecting "rule out" diagnoses rather than true cases of COVID-19. Validation of ICD-10-CM-based diagnostic coding algorithms for confirmed diagnoses based on a positive NAAT result for SARS-CoV-2 is ongoing. Analyses of the natural history of COVID-19 should therefore consider conducting parallel analyses using different COVID-19 definitions (e.g., based on positive laboratory tests for SARS-CoV-2 and claims-based diagnoses of COVID-19). Consistent results across different methods of COVID-19 classification will enhance the validity of the results.

Some data elements within claims and EHR databases will be challenging to ascertain and may also be subject to misclassification. Day of symptom onset will likely be unavailable or inaccurate without chart review or patient-reported data. Vital signs, height, weight, and laboratory data are not available within claims data and may vary in format in EHR systems, requiring substantial effort to standardize and clean. Inpatient medication use is not well-

captured within claims data. Oxygen support, important for describing the progression of COVID-19 disease, is often undercoded, and durations of different kinds of oxygen support may not be accurate. Codes for signs and symptoms are likely inconsistent across settings, individual coders, and time points, and mild symptoms may tend to be undercoded, with implications for disease-severity classification. Some important complications of COVID-19, such as multiorgan dysfunction or multisystem inflammatory syndrome in children (MIS-C), do not have specific diagnosis codes yet, and any algorithms created to detect them are as yet unvalidated.

There is also potential for misclassification regarding pregnancy and gestational age. The pregnancy identification and gestational age estimation algorithm was developed for health insurance claims data because in these data there is no information about pregnancy start or the pregnancy period;[23] gestational age at delivery must be derived. Pregnancy episodes that do not end in a live birth (e.g., stillbirth, termination, etc.) cannot be captured with the existing tool, and we will need to implement alternative algorithms to identify these pregnancy outcomes. Investigators can describe stillbirth and other non-live birth outcomes in a cohort of patients not identified by a delivery code but would not be able to determine with certainty whether the diagnoses or characteristics of interest occurred within the pregnancy period. As we rely on an algorithm to estimate pregnancy start and define the pregnancy episode, there may be misclassification of gestational age. This will be especially important to consider when evaluating precise timing of patient characteristics or diagnoses, such as COVID-19. If using claims-based data, reporting these characteristics by trimester is less likely to result in misclassification of timing of diagnosis.

b. Selection bias:

There is the potential for selection bias due to variability in COVID-19 testing over time, by geographic location, and by severity of disease. Investigators could consider sensitivity analyses whereby they: 1) repeat analyses including only persons tested after a specific date, e.g., April 1, 2020, after the specific ICD-10-CM diagnosis code for COVID-19 and better access to testing became available; 2) condition analyses on geographic region, since this might affect COVID-19 exposure and selection into the cohort; or 3) conduct analyses according to severity of illness at presentation (categorized in Table 3, above) or setting of care (e.g., ambulatory versus hospital-based versus nursing home).

c. Unmeasured confounding:

Unmeasured confounding represents a primary limitation of observational study designs. For that reason, investigators should consider performing sensitivity analyses to determine the potential impact of unmeasured confounders on observed associations between exposures of interest and outcomes.[26,27]

d. Protopathic bias:

Analyses evaluating the comparative effectiveness of treatments on the natural history of COVID-19 diagnosis should consider the potential for protopathic bias,[28] which occurs when treatment exposure is influenced by early stage of disease. A medication may be prescribed for early symptoms of disease not yet diagnosed. Failure to account for reverse causality may create a false association between a medication and the outcome of interest.[28] For example, NSAIDs may be prescribed for early symptoms of COVID-19. Analyses examining associations between NSAIDs and severity of COVID-19 might examine different exposure windows prior to COVID-19 diagnosis to assess for possible protopathic bias.

e. Issues in ascertaining initiation and duration of medications of interest:

Ascertainment of exposure to medications of interest and their duration will depend on the setting in which the medication is administered (e.g., hospital, ambulatory) and the type of drug. Within HCA and similar data systems, inpatient medication administration dates and

times are often available, but this level of detail may not be consistently available in other systems. For example, inpatient treatments may not be captured in claims-based systems, and when they are captured, treatment timing may not be available. If medication initiation occurs outside the hospital setting, outpatient dispensing data may be used, but assumptions about the timing of treatment initiation must be made based on dispensing date. Integrated care delivery systems would likely include data from both settings. The best data sources to use will depend on the treatments of interest and their typical administration/dispensing setting. Obtaining duration of treatment may require medical record review.

f. Issues with healthcare utilization variables:

During the COVID-19 pandemic, healthcare utilization variables, such as ambulatory encounters, Emergency Department visits, hospital admission, ICU admission, duration of ICU stay, and mechanical ventilation, might not appropriately classify patients' true disease severity. Health care utilization may be affected by state and local "stay at home" orders and local health care resource demand and supply and may vary by time, region, and even hospital in the same region. Surges in COVID-19 cases and shifting of the epidemic center as the pandemic continues will add more complexity to determining how the impact of the uncertainty of these variables can be adequately addressed in sensitivity analyses.

g. Generalizability:

The set of COVID-19 patients assembled may not be fully representative of the population. For example, in commercial insurance data sources, publicly insured people may be underrepresented, and uninsured people are not represented. Certain ethnic groups might be underrepresented.

D. Human subjects considerations

The Sentinel Initiative is a public health surveillance activity that is not considered research under the Federal Policy for the Protection of Human Subjects -- the "Common Rule" (see 42 CFR 46.02(I)(2)).[29] Thus, neither development nor implementation of this protocol is under the purview of institutional review boards.

E. References

1. Zhu N, Zhang D, Wang W, et al. A Novel Coronavirus from Patients with Pneumonia in China, 2019. *N Engl J Med* 2020; **382**(8): 727-33.
2. Huang C, Wang Y, Li X, et al. Clinical features of patients infected with 2019 novel coronavirus in Wuhan, China. *Lancet* 2020; **395**(10223): 497-506.
3. Chen N, Zhou M, Dong X, et al. Epidemiological and clinical characteristics of 99 cases of 2019 novel coronavirus pneumonia in Wuhan, China: a descriptive study. *Lancet* 2020; **395**(10223): 507-13.
4. Guan WJ, Ni ZY, Hu Y, et al. Clinical Characteristics of Coronavirus Disease 2019 in China. *N Engl J Med* 2020.
5. Wang D, Hu B, Hu C, et al. Clinical characteristics of 138 hospitalized patients with 2019 novel coronavirus-infected pneumonia in Wuhan, China. *JAMA* 2020.
6. Wu C, Chen X, Cai Y, et al. Risk factors associated with acute respiratory distress syndrome and death in patients With Coronavirus Disease 2019 pneumonia in Wuhan, China. *JAMA Intern Med* 2020.
7. Young BE, Ong SWX, Kalimuddin S, et al. Epidemiologic Features and Clinical Course of Patients Infected With SARS-CoV-2 in Singapore. *JAMA* 2020.
8. Zhou F, Yu T, Du R, et al. Clinical course and risk factors for mortality of adult inpatients with COVID-19 in Wuhan, China: a retrospective cohort study. *Lancet* 2020; **395**(10229): 1054-62.

9. Grasselli G, Pesenti A, Cecconi M. Critical care utilization for the COVID-19 outbreak in Lombardy, Italy: Early experience and forecast furing an emergency response. *N Engl J Med* 2020: In Press.
10. Shi H, Han X, Jiang N, et al. Radiological findings from 81 patients with COVID-19 pneumonia in Wuhan, China: a descriptive study. *Lancet Infect Dis* 2020; **20**(4): 425-34.
11. Wu Z, McGoogan JM. Characteristics of and Important Lessons From the Coronavirus Disease 2019 (COVID-19) Outbreak in China: Summary of a Report of 72314 Cases From the Chinese Center for Disease Control and Prevention. *JAMA* 2020.
12. Xu XW, Wu XX, Jiang XG, et al. Clinical findings in a group of patients infected with the 2019 novel coronavirus (SARS-Cov-2) outside of Wuhan, China: retrospective case series. *BMJ* 2020; **368**: m606.
13. Yang J, Zheng Y, Gou X, et al. Prevalence of comorbidities in the novel Wuhan coronavirus (COVID-19) infection: a systematic review and meta-analysis. *Int J Infect Dis* 2020.
14. Grasselli G, Zangrillo A, Zanella A, et al. Baseline Characteristics and Outcomes of 1591 Patients Infected With SARS-CoV-2 Admitted to ICUs of the Lombardy Region, Italy. *JAMA* 2020.
15. Team CC-R. Severe outcomes among patients with Coronavirus Disease 2019 (COVID-19) - United States, February 12-March 16, 2020. *MMWR Morb Mortal Wkly Rep* 2020; **69**(12): 343-6.
16. Preliminary Estimates of the Prevalence of Selected Underlying Health Conditions Among Patients with Coronavirus Disease 2019 — United States, February 12–March 28, 2020. MMWR Morb Mortal Wkly Rep. ePub: 31 March 2020. DOI: http://dx.doi.org/10.15585/mmwr.mm6913e2.
17. Platt R, Brown JS, Robb M, et al. The FDA Sentinel Initiative - An Evolving National Resource. *N Engl J Med* 2018; **379**(22): 2091-3.
18. Cheng MP, Papenburg J, Desjardins M, et al. Diagnostic Testing for Severe Acute Respiratory Syndrome-Related Coronavirus 2: A Narrative Review. *Ann Intern Med* 2020; **172**(11): 726-34.
19. Sethuraman N, Jeremiah SS, Ryo A. Interpreting Diagnostic Tests for SARS-CoV-2. *JAMA* 2020.
20. Weissleder R, Lee H, Ko J, Pittet MJ. COVID-19 diagnostics in context. *Sci Transl Med* 2020; **12**(546).
21. Centers for Disease Control and Prevention. ICD-10-CM Official Coding and Reporting Guidelines: April 1, 2020 through September 30, 2020. Accessed at: https://www.cdc.gov/nchs/data/icd/COVID-19-guidelines-final.pdf?fbclid=IwAR06h6zP8KkehNzjEqlpGvACzMBhP26Khp9WG1JqbfLUQpYOt_LCqwGVHxU. Accessed on September 29, 2020.
22. Team CCR, Jorden MA, Rudman SL, et al. Evidence for Limited Early Spread of COVID-19 Within the United States, January-February 2020. *MMWR Morb Mortal Wkly Rep* 2020; **69**(22): 680-4.
23. MacDonald SC, Cohen JM, Panchaud A, McElrath TF, Huybrechts KF, Hernandez-Diaz S. Identifying pregnancies in insurance claims data: Methods and application to retinoid teratogenic surveillance. *Pharmacoepidemiol Drug Saf* 2019; **28**(9): 1211-21.
24. Nickson C. Multiple Organ Dysfunction Syndrome. Life in the Fast Lane website. Accessed at: https://litfl.com/multiple-organ-dysfunction-syndrome/. Accessed on August 29, 2020.
25. US Food and Drug Administration. COVID-19: Developing Drugs and Biological Products for Treatment or Prevention Guidance for Industry. May 2020. Accessed at: https://www.fda.gov/media/137926/download. Accessed on August 29, 2020.
26. Haneuse S, VanderWeele TJ, Arterburn D. Using the E-Value to Assess the Potential Effect of Unmeasured Confounding in Observational Studies. *JAMA* 2019; **321**(6): 602-3.

27. Lin DY, Psaty BM, Kronmal RA. Assessing the sensitivity of regression results to unmeasured confounders in observational studies. *Biometrics* 1998; **54**(3): 948-63.
28. Horwitz RI, Feinstein AR. The problem of "protopathic bias" in case-control studies. *Am J Med* 1980; **68**(2): 255-8.
29. U.S. Department of Health and Human Services, Office for Human Research Protections. Federal Policy for the Protection of Human Subjects ('Common Rule'). Accessed at: https://www.hhs.gov/ohrp/regulations-and-policy/regulations/common-rule/index.html. Accessed on October 9, 2020.

F. Appendices [1-4]

Appendix 1: ICD-10-CM Algorithm for Gestational Age

The ICD-10 algorithm for gestational age incorporates codes for (1) gestational age in weekly increments from gestational week 20 through gestational week 42 or greater (codes Z3A20-Z3A49, referred to as the "Z codes"), (2) preterm delivery (other than the Z codes), and (3) postterm delivery (other than the Z codes). We identified ICD-10 codes for preterm delivery and postterm delivery by implementing forward-backward mapping of ICD-9-CM codes included in the initial version of the pregnancy tool. Of note, there are no codes equivalent to the Z codes in the ICD-9-CM coding scheme.

Priorities:

If multiple codes for specific weeks of gestation (Z codes), preterm delivery, and/or postterm delivery are available, the ICD-10 algorithm for gestational age prioritizes the following codes:

(1) **Codes that specify weeks of gestation, including all Z codes ranging from 20 weeks through >=42 weeks of gestation in one-week increments, and codes that indicate preterm delivery with weeks of gestation specified in one-week increments (other than Z codes)**. If multiple codes are observed, codes indicating longer gestational age are prioritized over those indicating shorter gestational age. We assume the approximate mid-point of the specified gestational age [e.g., 263 days (37 weeks and 4 days) for 37 weeks gestation].
(2) **Codes that indicate preterm delivery without specifying weeks of gestation.** If multiple codes are observed, codes with more specificity (e.g., preterm delivery, 2nd trimester of pregnancy or 'extreme immaturity') are prioritized over those with less specificity (e.g., preterm newborn, unspecified weeks of gestation). Further, codes indicating longer gestational age are prioritized over those indicating shorter gestational age.
(3) **Codes that indicate postterm delivery without specifying weeks of gestation.** If multiple codes are observed, codes indicating longer gestational age are prioritized over those indicating shorter gestational age.

If no codes for preterm or postterm delivery are observed, then the default assumption for gestational age is 273 days. However, this assumption is user specified and can be modified.

Gestational age code lookup file

Code	Description	Code-type	Dura-tion	Prior-ity group 1	Prior-ity group 2	Prior-ity
Z3A49	Greater than 42 weeks gestation of pregnancy	10	301	1	0	1
Z3A42	42 weeks gestation of pregnancy	10	298	1	0	2
Z3A41	41 weeks gestation of pregnancy	10	291	1	0	3
Z3A40	40 weeks gestation of pregnancy	10	284	1	0	4
Z3A39	39 weeks gestation of pregnancy	10	277	1	0	5
Z3A38	38 weeks gestation of pregnancy	10	270	1	0	6
Z3A37	37 weeks gestation of pregnancy	10	263	1	0	7
P0739	Preterm newborn, gestational age 36 completed weeks	10	256	1	0	8
Z3A36	36 weeks gestation of pregnancy	10	256	1	0	8
P0738	Preterm newborn, gestational age 35 completed weeks	10	249	1	0	9
Z3A35	35 weeks gestation of pregnancy	10	249	1	0	9
P0737	Preterm newborn, gestational age 34 completed weeks	10	242	1	0	10
Z3A34	34 weeks gestation of pregnancy	10	242	1	0	10
P0736	Preterm newborn, gestational age 33 completed weeks	10	235	1	0	11
Z3A33	33 weeks gestation of pregnancy	10	235	1	0	11
P0735	Preterm newborn, gestational age 32 completed weeks	10	228	1	0	12
Z3A32	32 weeks gestation of pregnancy	10	228	1	0	12
P0734	Preterm newborn, gestational age 31 completed weeks	10	221	1	0	13
Z3A31	31 weeks gestation of pregnancy	10	221	1	0	13
P0733	Preterm newborn, gestational age 30 completed weeks	10	214	1	0	14
Z3A30	30 weeks gestation of pregnancy	10	214	1	0	14
P0732	Preterm newborn, gestational age 29 completed weeks	10	207	1	0	15

Z3A29	29 weeks gestation of pregnancy	10	207	1	0	15
P0731	Preterm newborn, gestational age 28 completed weeks	10	200	1	0	16
Z3A28	28 weeks gestation of pregnancy	10	200	1	0	16
P0726	Extreme immaturity of newborn, gestational age 27 completed weeks	10	193	1	0	17
Z3A27	27 weeks gestation of pregnancy	10	193	1	0	17
P0725	Extreme immaturity of newborn, gestational age 26 completed weeks	10	186	1	0	18
Z3A26	26 weeks gestation of pregnancy	10	186	1	0	18
P0724	Extreme immaturity of newborn, gestational age 25 completed weeks	10	179	1	0	19
Z3A25	25 weeks gestation of pregnancy	10	179	1	0	19
P0723	Extreme immaturity of newborn, gestational age 24 completed weeks	10	172	1	0	20
Z3A24	24 weeks gestation of pregnancy	10	172	1	0	20
P0721	Extreme immaturity of newborn, gestational age less than 23 completed weeks	10	158	1	0	22
P0722	Extreme immaturity of newborn, gestational age 23 completed weeks	10	165	1	0	21
Z3A23	23 weeks gestation of pregnancy	10	165	1	0	21
Z3A22	22 weeks gestation of pregnancy	10	158	1	0	22
Z3A21	21 weeks gestation of pregnancy	10	151	1	0	23
Z3A20	20 weeks gestation of pregnancy	10	144	1	0	24
P0720	Extreme immaturity of newborn, unspecified weeks of gestation	10	196	1	0	25
O6013*	Preterm labor second trimester with preterm delivery third trimester	10	245	1	0	26
O6014*	Preterm labor third trimester with preterm delivery third trimester	10	245	1	0	26
P0730	Preterm newborn, unspecified weeks of gestation	10	245	1	0	26
O6012*	Preterm labor second trimester with preterm delivery second trimester	10	168	1	0	27
76528	35-36 completed weeks of gestation	09	252	1	0	28

76527	33-34 completed weeks of gestation	09	238	1	0	29
76526	31-32 completed weeks of gestation	09	224	1	0	30
76525	29-30 completed weeks of gestation	09	210	1	0	31
76524	27-28 completed weeks of gestation	09	196	1	0	32
76523	25-26 completed weeks of gestation	09	182	1	0	33
76522	24 completed weeks of gestation	09	168	1	0	34
76521	Less than 24 completed weeks of gestation	09	168	1	0	34
7650*	Extreme immaturity, by weight	09	196	1	0	35
64421	Early onset of delivery, delivered, with or without mention of antepartum condition	09	245	1	0	36
7651*	Other preterm infants, by weight	09	245	1	0	36
76520	Unspecified weeks of gestation	09	245	1	0	36
O481	Prolonged pregnancy	10	294	0	1	1
P0822	Prolonged gestation of newborn	10	294	0	1	1
O480	Post-term pregnancy	10	287	0	1	2
P0821	Post-term newborn	10	287	0	1	2
6452*	Prolonged pregnancy	09	294	0	1	28
76622	Prolonged gestation of infant	09	294	0	1	28
6451*	Post term pregnancy	09	287	0	1	29
76621	Post-term infant	09	287	0	1	29

Appendix 2: Sentinel Common Data Model Mother-Infant Linkage criteria

Mothers/deliveries

Deliveries were identified using the following criteria:

1. An encounter from one year later than the start date of the Data Partner's data availability through the end date of the DP's data
2. An encounter contained a delivery code of interest (over 700 diagnosis and procedure codes)
3. Women were between 10 years and 54 years of age inclusive on the admission date of the delivery encounter
4. No evidence of delivery for 180 days prior to any identified delivery

5. Mothers must have had medical coverage from 180 days prior to date of delivery through the delivery date. A single enrollment gap of up to 45 days within the enrollment period is allowed.

Note that more than one delivery per mother may be identified within the Data Partner's Sentinel Distributed Database.

Infants

1. Identified with a date of birth from one year later than the start date of the DP's data availability through the end date of the DP's data
2. An infant must have at least one day of enrollment with medical coverage during their first year of life.

A family subscriber number was used for most of the linkages. (For some commercial plans when a family is all covered under a single subscriber (e.g., a parent), a family subscriber number is assigned to each member in the family.) Further, matching on family subscriber number is virtually standard across Data Partners with potentially slight variation. However, linking methods for those not initially matched on subscriber ID may vary depending on the identifying information each DP has in their source data, such as names and addresses. In most cases, Data Partners used deterministic matching to link subscriber IDs and other variables to each other.

Most linkages are restricted to the infant's date of birth being within the interval of three days prior to delivery visit admission date through delivery visit discharge date. This three-day window is built in to allow for out of hospital deliveries to be recorded and linked and is based on experience from the MEPREP and PRISM projects. In the case of multiples, more than one infant can be linked to the same mother/delivery if they have separate subscriber IDs; however, current capabilities allow for analysis of singletons only.

Appendix 3: Data elements

[separate document, COVID-19 Natural History Data Elements_*[date]*.xlsx]

Appendix 4: Code lists

[separate document, COVID Natural History Code List_*[date]*.xlsx]