Surveillance for Emerging Threats to Mothers and Babies Network

Sampling and Weighting Methods For COVID-19 End of Pregnancy Medical Record Abstraction

Date: October 2021

Division of Birth Defects and Infant Disorders

National Center on Birth Defects and Developmental Disabilities

Table of Contents

Background

SET-NET

The Surveillance for Emerging Threats to Mothers and Babies Network, or SET-NET, is mother-baby linked longitudinal surveillance to understand the impact of emerging and reemerging threats on pregnant people and their infants. SET-NET is a preparedness network that can be further expanded if new threats emerge for mothers and babies. In early 2020 it was rapidly adapted for COVID-19 surveillance.[1]

Medical records abstraction (MRA) to capture key clinical information on mothers and infants is fundamental to the SET-NET surveillance approach. However, for COVID-19 surveillance, given high case counts in some parts of the United States and limited resources, jurisdictions requested assistance from CDC to prioritize surveillance resources by conducting MRA on a representative sample of included pregnancies as opposed to the entire cohort. Given the need to quickly report data for clinical decision-making and public health action and minimize the burden on already strained health departments, CDC developed a sampling approach to support health departments in the collection of population-based data. The sampling approach may be expanded to other SET-NET exposures of interest in the future.

SET-NET's longitudinal design indicates that there are two main components for MRA: end-of-pregnancy abstraction, which captures maternal, pregnancy, and birth records, and infant follow-up (IFU) abstraction, which captures infants' development over time. This document is intended for a technical audience and describes the methodology of the COVID-19 sampling approach for MRA of end of pregnancy, and calculations of sampling weights and population estimates. Additional documentation for the COVID-19 IFU sampling methodology and weight calculations, and the sampling methodology for additional SET-NET exposures, will be provided at a later date.

Surveillance Cohort

The population considered for surveillance through SET-NET varies by exposure. The COVID-19 inclusion criteria are provided in Table 1.

For SET-NET COVID-19 surveillance, data are collected on pregnant women with laboratory evidence of COVID-19 during pregnancy. Infants born to women with COVID-19 during pregnancy may be monitored for follow-up through SET-NET, even if the infant has no confirmed congenital infection, in order to support detection of long-term outcomes.

Modules and Data Sources

The SET-NET data system is organized into *general* variables and exposure-specific *modular* variables. The general variables pertain to all mother–baby pairs, regardless of the exposure of interest. Exposure-specific or modular variables complement the general variables by providing information for mother–baby pairs about the exposure of interest. Modular variables were selected to align with existing data sources and published literature and reviewed by a team of experts in obstetrics, pediatrics, epidemiology, and informatics with

[1] Woodworth KR, Reynolds MR, Burkel V, Gates C, Eckert V, McDermott C, Barton J, Wilburn A, Halai UA, Brown CM, Bocour A, Longcore N, Orkis L, Delgado Lopez C, Sizemore L, Ellis EM, Schillie S, Gupta N, Bowen VB, Torrone E, Ellington SR, Delaney A, Olson SM, Roth NM, Whitehill F, Zambrano LD, Meaney-Delman D, Fehrenbach SN, Honein MA, Tong VT, Gilboa SM. A Preparedness Model for Mother-Baby Linked Longitudinal Surveillance for Emerging Threats. *Matern Child Health J* 2021;25(2):198-206. https://doi.org/10.1007/s10995-020-03106-y

consideration for potential data capture. Together, general and modular variables align with key surveillance questions for each exposure, while striving to minimize burden and ensure quality data.[1]

The surveillance protocol focused health department data collection efforts on hospital medical records and healthcare providers' offices (e.g., prenatal care records, maternal and newborn hospitalization records, and infant care medical records). Other data sources may include records from routine case investigations and reports and vital statistics (birth and fetal death certificates). Linkage to data sources such as birth certificates is a common strategy to identify pregnancy status retrospectively for reported cases of infectious diseases.

Inclusion Criteria

The COVID-19-specific inclusion criteria are described in Table 1. This table will be expanded in the future to include additional SET-NET exposures.

Table 1. Inclusion criteria for COVID-19 SET-NET cases

Exposure	Inclusion criteria
COVID-19	• Pregnant women who are SARS-CoV-2 RNA positive (laboratory-confirmed) in at least one clinical specimen at any point during pregnancy, up to and including the day of delivery **AND** • Who reside in a participating jurisdiction **AND** • Who test positive from January 1-December 31, 2020

Case Ascertainment

All SET-NET exposures are nationally notifiable diseases (CSTE case definitions: COVID-19, hepatitis C, syphilis) and as such, case data are submitted through the National Notifiable Disease Surveillance System, or NNDSS, on an electronic report form specific to exposure. The NNDSS case report forms include a pregnancy checkbox to identify pregnant cases; however, pregnancy status ascertainment typically requires case interview or medical chart review. The quality and accuracy of pregnancy status varies by exposure and jurisdiction. As such, most jurisdictions rely on linkages between case surveillance and other available data sources to fully ascertain case counts. These additional data sources may include linkages of case surveillance systems to vital statistics data (such as birth certificates or fetal death certificates), linkages of case surveillance to prenatal screening records, or administrative data including hospital discharge data. For jurisdictions that are sampling for MRA, the complete list of ascertained cases becomes the sampling frame. The unit of sample selection for the end of pregnancy sampling approach is the **pregnancy**.

Sampling Methods

Objective

The objective of this sampling design is for jurisdictions to collect a probability sample to obtain precise estimates of maternal, pregnancy, and birth characteristics and outcomes among pregnancies with the exposure of interest.

Reporting Jurisdiction

Sampling occurs at the jurisdictional level. Cities or counties reporting exposure-specific pregnancy surveillance data separately from the state (e.g., California and Los Angeles County, Pennsylvania and City of Philadelphia, Illinois and City of Chicago) sample their target populations separately from the larger jurisdictional region. State jurisdictions remove cases reported through city or county jurisdictions from their sampling frames.

Target Population and Sampling Frame

Target Population

The target population for the SET-NET COVID-19 population should include all pregnancies meeting the inclusion criteria described in Table 1.

Sampling Frame of Pregnancies with Exposures of Interest

CDC allows jurisdictions to make local decisions regarding the development of their sampling frame based on the accuracy and completeness of their data sources. Jurisdictional sampling frames are constructed according to the inclusion criteria shown in Table 1 and comprise the full list of cases ascertained through jurisdictional surveillance approaches including data linkages. There are several ways jurisdictions can identify pregnancies and jurisdictions are not limited to one method exclusively. Some examples are

1. Jurisdictional case surveillance data for pregnant cases (pregnancy status directly indicated on NNDSS case report form or reportable disease registries)
2. Data linkages to confirm pregnancy status (linking billing data, prenatal screening, or other data sources to case surveillance data)
3. Data linkages to confirm pregnancy completion (linking to birth outcomes such as birth certificates, fetal death certificates, administrative databases, or other data sources to case surveillance data)

Gaps in the Sampling Frame

Jurisdictions considered the implications for their generalizability from each ascertainment method including timeliness and completeness of case ascertainment. Approaches that linked to datasets for births may introduce a time lag. Jurisdictions confirming pregnancy status using data linkages determined whether cases that did not link were part of their sampling frames (e.g., persons indicated as pregnant on the case report form who do not link to vital records). If jurisdictions decided to exclude these unlinked cases from their sampling frame, this was noted as a limitation of their generalizability. If a jurisdiction decided to keep these unlinked cases their sampling frame, then the jurisdiction was required to pursue additional processes or linkages to ascertain pregnancy on these unlinked cases.

Requirements

There are four requirements related to the selection and documentation of the sample:

1. Random selection must be used in each sampling step so that every eligible pregnancy in the sampling frame has a non-zero chance of being selected into the sample.
2. The probability of selection for every pregnancy must be known and retained on the final analytic files for the study.

3. Jurisdictions must have unique identifiers for every sampled pregnancy and these identifiers must be retained in the final analytic file.
4. The sample must be selected using simple random sampling, without replacement, at one or more separate time intervals during the surveillance period. Details of this type of sample design are provided in the next section.

Selection of the Sample

Sample Size

The sample size is determined by each jurisdiction and based on their capacity to conduct MRA.

Sampling Intervals

Sampling of the data for MRA are performed at regular time intervals throughout the reporting period, rather than waiting until data collection has ended. Based on their capacity, jurisdictions that are sampling have determined regular, appropriate intervals for selecting a sample. The sampling frame will be partitioned across the time intervals. Once the sample has been selected for a time interval, all remaining cases from that interval's sampling frame become ineligible for sampling in any subsequent intervals. Therefore, any given case only has one opportunity to be sampled. Each jurisdiction, in consultation with CDC, decides on the number or proportion of cases to be selected for MRA based on expected total cases for the jurisdiction and capacity for MRA.

Priority Pregnancies

CDC requested ***end of pregnancy MRA on <u>all</u> ascertained cases with select priority outcomes.*** Priority outcomes are selected based on low frequency and their importance to answer surveillance questions and should be identifiable prior to MRA. These priority outcomes might include stillbirths, maternal deaths, infant deaths, and postnatal infection in infants. Priority outcomes for COVID-19 in SET-NET are noted in Table 2.

Table 2. Inclusion criteria and priority outcomes for COVID-19 SET-NET cases.

Exposure	Inclusion criteria	Priority Outcome Pregnancies
COVID-19	• Pregnant women who are SARS-CoV-2 RNA positive (laboratory-confirmed) in at least one clinical specimen at any point during pregnancy, up to and including the day of delivery **AND** • Who reside in a participating jurisdiction **AND** • Who test positive from January 1-December 31, 2020	• Pregnancies resulting in one or more neonates who test positive for SARS-CoV-2 infection during the birth hospitalization or within 14 days of birth **AND** who are born to women who meet the inclusion criteria

For pregnancies with priority outcomes, (i.e., priority pregnancies), the expectation for MRA was complete ascertainment (i.e., a sampling fraction of 1.0). Sampling of pregnancies with other outcomes (i.e., nonpriority outcomes) is conducted after the priority pregnancies are removed.

Random Sample and Stratified Random Sample

Once priority pregnancies are identified and removed from the frame, jurisdictions use simple random sampling without replacement to select additional pregnancies for MRA. The probability of selection for any nonpriority pregnancy is the number selected out of the number of possible. Random selection provides the best method to obtain a representative sample and the sampling weight is the inverse probability of selection.

A small number of jurisdictions conduct stratified random sampling in order to ensure adequate representation of a given subgroup of cases. For example, a jurisdiction may be interested in stratified random sampling by maternal race and ethnic subgroups and may choose to apply a higher sampling fraction to some maternal race and ethnic subgroups that are less prevalent in the sampling frame, and a lower sampling fraction to.

Weighting

Weighted datasets are created by CDC quarterly. The goal of the sampling weights is to adjust each record such that, as a whole, the submitted records from a jurisdiction represent the total sampling frame of that jurisdiction. All records will receive a sampling weight, even records from jurisdictions that conduct MRA on all of their cases (i.e., census approach, see below). Jurisdictions conducting sampling for MRA submit documentation of their sampling interval, which includes total cases, number of priority pregnancies, number of records eligible for selection, number of pregnancies selected, and whether medical records were available and abstracted.

Census Approach

This is the approach for jurisdictions not sampling for MRA. Jurisdictions implementing the census approach do not need to send additional documentation to CDC; all pregnancies have an analytic weight *(w)* of 1.0. CDC assumes any missingness is at random, and each record represents one case. The jurisdictional stratum is also 1 for all records. For jurisdictions conducting a full census approach for MRA, each case is weighted to represent only itself and the sum of the cases is the size of the population of interest in the jurisdiction.

$$w = 1$$

Simple Random Sample and Stratified Random Sample Approaches

For all jurisdictions that are sampling for end of pregnancy MRA, the general weight calculation is the same for all nonpriority pregnancies. Most jurisdictions are conducting simple random sampling and not stratified random sampling. For these jurisdictions, there is only one jurisdictional stratum *j* for all pregnancies within the jurisdiction. For jurisdictions that conduct stratified random sampling, the selection and the nonresponse weights will be unique to each stratum *j* per interval *i*.

Formula 1. Sampling weight

$$w_{ij} = w1_{ij}w2_{ij}$$

Where $w1_{ij}$ is the inverse probability of selection for cases in sampling interval *i* and jurisdictional stratum *j* (i.e., stratum *j*) and $w2_{ij}$ is the nonresponse weight for sampling interval *i* and jurisdictional stratum *j*

Formula 2. w1, the inverse probability of selection

$$p1_{ij} = \mathrm{Prob}(\mathrm{selection}\ |\mathrm{interval}\ i\ \mathrm{and\ stratum}\ j) = \frac{number\ of\ cases\ selected\ for\ MRA\ for\ interval\ i\ and\ stratumj}{number\ of\ total\ cases\ in\ sampling\ frame\ for\ interval\ i\ and\ stratumj}$$

therefore

$$w1_{ij} = \frac{1}{Prob(selection\ |interval\ i\ and\ stratum\ j)}$$

Formula 3. w2, the nonresponse weight

And *w*2 is the nonresponse weight for observations in sampling interval *i* and stratum *j*

$$w2_{ij} = \frac{number\ of\ cases\ selected\ for\ MRA\ for\ interval\ i\ and\ stratum\ j}{number\ of\ cases\ with\ MRA\ completed\ for\ interval\ i\ and\ stratum\ j}$$

Sampling weights are applied only to records for which MRA is completed. The sum of the sampling weights assigned to records with completed MRA in interval *i* and stratum *j* is the number of total cases in the sampling frame for interval *i* and stratum *j*.

Priority Pregnancies

Jurisdictions should select all priority pregnancies with certainty such that the probability of selection is 1.0.

Sampling weight. All priority pregnancies that are identified before MRA occurs have a probability of selection of 1.0, and hence have *w1*=1.

Nonresponse weight. Although all priority pregnancies have *w1*=1, it is still possible that these cases may be lost to follow-up (e.g., medical records cannot be located or MRA has not been started). Therefore, priority pregnancies are subject to a nonresponse weight to ensure the number of priority pregnancies in the dataset represent the true number of priority pregnancies in the jurisdiction. The nonresponse weight (*w2*) for priority pregnancies uses the same formula shown in Formula 3, except the nonresponse weight is calculated for the full priority frame rather than individually for each interval. That is, the priority pregnancy nonresponse weight is pooled for all intervals within a jurisdiction. Priority cases are identified prior to stratified random sampling, so the nonresponse weights for priority cases do not consider strata.

Formula 4. Nonresponse weight for priority cases identified prior to MRA

$$w2 \quad = \frac{number\ of\ priority\ pregnancies\ in\ jurisdiction\ (all\ intervals)}{number\ of\ priority\ pregnancies\ with\ MRA\ completed\ (all\ intervals)}$$

Population Estimates

When analyzing weighted SET-NET data, a finite population correction (FPC) is used to adjust standard error estimates. A FPC is appropriate when participating jurisdictions sample data for >5% of the target population (for SET-NET COVID-19 surveillance some jurisdictions may be reporting >99% of the target population). As such, it is more appropriate to analyze the data as a population without replacement and with population totals included for each jurisdiction and strata. Ideally, jurisdictions provide their total case count; however, for COVID-19 these population estimates may not be finalized until well after the sampling/weighting and analyses have begun. While case ascertainment for 2020 SET-NET COVID-19 infections was still underway, we estimated the case counts of COVID-19 in pregnancy for all participating jurisdictions.

Example of calculating population estimates for COVID-19 surveillance

CDC SET-NET estimated the population totals (i.e., the number of cases meeting inclusion criteria) for each jurisdiction using the following process. The total population and total women aged 15-44 years, termed women of reproductive age (WRA), by jurisdiction was based on the 2019 Census data. Next, the total WRA was multiplied by 0.05 to provide the number of pregnant women at any given time ().[2] The total number of COVID-19 cases in 2020 was based on USA Facts data . The cumulative incidence of COVID-19 in 2020 was calculated by jurisdiction and then the total number of pregnant women with COVID-19 in 2020 was estimated. Calculations of each variable can be found in Table 3.

Table 3. Calculation of total pregnant women with COVID-19 by jurisdiction in 2020.

Variables	Explanation/Calculation
Total population	Obtained from US Census data: https://data.census.gov/cedsci/table?g=0100000US&tid=ACSST1Y2019.S0101&hidePreview=true
Total WRA	Obtained from US Census data: https://data.census.gov/cedsci/table?g=0100000US&tid=ACSST1Y2019.S0101&hidePreview=true
Total Pregnant Women	Total WRA * 0.05
Total COVID-19 cases in 2020	Obtained from USA Facts: https://usafacts.org/visualizations/coronavirus-covid-19-spread-map/
Cumulative Incidence of COVID-19	Total COVID-19/Total population * 100,000
Total Pregnant Women with COVID-19	Total Pregnant Women / 100,000 * Cumulative Incidence

[2] Reference: Estimating the Number of Pregnant Women in a Geographic Area (cdc.gov), URL: https://www.cdc.gov/reproductivehealth/emergency/pdfs/PregEstimator_PointIntime-8_2013.pdf. Accessed October 12, 2021.

Limitations

This sampling approach has some limitations. First, CDC and jurisdictions continue to collaborate to ensure the sampling frames are as complete as possible so that these data represent the population of interest under surveillance. However, given issues with completeness and accuracy of pregnancy status across data sources and jurisdictions, CDC allows jurisdictions to make local decisions regarding the development of their sampling frame based on the accuracy and completeness of their data sources. CDC will assess the completeness of sampling using estimations described previously, but gaps may still exist. CDC completed a preliminary sensitivity analysis for two jurisdictions that sent linked vital statistics data for all cases in their jurisdiction and conducted MRA on a subset of sampled cases. The results of this sensitivity analysis showed that weighted 95% confidence intervals included the population estimate from the full cohort (e.g., all cases identified through linked birth certificates) for 92% of the maternal variables, suggesting that the weighted estimates and confidence intervals (CIs) were appropriately capturing the true population estimates.

Second, because CDC guided jurisdictions to partition their sampling into intervals to enable staff to begin MRA, intervals with partial MRA are adjusted for nonresponse at the time of weighting, and these same intervals may be updated later when medical records are returned. In addition, intervals without any MRA cannot be weighted and are omitted from weighted datasets until abstraction begins. Thus, reports using the interim weighted dataset are considered preliminary, and findings may be updated as MRA are completed for the entire interval and subsequently for the entire cohort for an exposure. However, these interim analyses are critical for informing clinical decision-making and public health action, and CDC will continue to monitor this approach and ensure conclusions are based on the best available data.

Summary

The COVID-19 pandemic stretched health department capacity to conduct medical record abstraction on all cases meeting SET-NET inclusion criteria. Although this approach was originally developed for COVID-19, its application is being adapted to address hepatitis C surveillance. The sampling approach enables the collection of population-based data while balancing the capacity and resources of health departments to conduct quality data collection from medical records. The approach also allows flexibility, so that the jurisdictions, in consultation with CDC, decide their sampling approach, including identification of priority pregnancies of interest or stratifications that might be useful to inform their local programmatic needs. As SET-NET was developed as a preparedness network, this sampling approach to collect population-based mother-baby linked longitudinal surveillance may have applications to other emerging threats and future responses.

Sources:

https://www.cdc.gov/reproductivehealth/emergency/pdfs/pregnacyestimatobrochure508.pdf

https://data.census.gov/cedsci/table?g=0100000US&tid=ACSST1Y2019.S0101&hidePreview=true

https://usafacts.org/visualizations/coronavirus-covid-19-spread-map/

Appendix A: Glossary of Terms

The following terms describe SET-NET sampling methods in this report.

Case: A pregnant person with the exposure of interest during pregnancy based on inclusion criteria.

Case Report Form: An electronic report form specific to exposure used by jurisdictions to report information on cases to CDC via the National Notifiable Disease Surveillance System (NNDSS). The CRF for selected exposures may capture pregnancy status, although quality and completeness may vary by exposure and jurisdictional capacity to conduct interviews or medical chart review.

Census: The total amount of cases in a given interval.

CSTE: Council of State and Territorial Epidemiologists, an organization of member states and territories representing public health epidemiologists. This organization developed the standard case definitions for all exposures included in SET-NET surveillance.

Infant: The live birth resulting from the pregnancy meeting inclusion criteria for surveillance. For the purpose of this document, stillborn infants are not included when the term infant is used.

Infant Follow-Up: Data collected at specified time intervals from the medical records of an infant's well child visit.

Interval: Specific time point that a jurisdiction sets as their time frame for selection of sampled cases.

Jurisdiction: State, local, and territorial health departments.

Medical record abstraction: MRA, collecting data from a medical record.

Mother: The pregnant person included in the surveillance.

Nonresponse: When a case has been selected for MRA, if that record cannot be found or any other reason why that case could not be abstracted.

Priority outcome: a certain selected outcome of interest that is based on: relatively low frequency, importance to answer surveillance questions, and can be identified prior to MRA. All pregnancies with this priority outcome are selected for medical record abstraction (i.e., probability of selection is 1.0).

Priority pregnancy: a case with a priority outcome such that all pregnancies with this outcome are selected for medical record abstraction (i.e., probability of selection is 1.0).

Sampling frame: The list of eligible cases from which the sample is selected for a specified interval.

Sampling weight: The inverse of the probability of selection multiplied by the nonresponse weight.

Target population: The entire population that the sampled data represent.