# Defining COVID-19 cohorts in real-world data

Noelle M. Cocoros, DSc, MPH

August 19, 2020

# Considerations for cohort identification in claims and EHR data

- In general, study question must guide definition and approach
    - e.g., descriptive, medical product safety, medical product effectiveness
    - Informs how definition must perform – sensitive, specificity, etc.
    - But data source/type must be considered – e.g., multi vs single site

- COVID-19 specifically
  - Many changes in a short time
    - Diagnostic coding recommendations and practices; likely varies by care setting; may vary by patient characteristics
    - Diagnostic lab test availability, performance, use; likely varies by geography, care setting, patient characteristics, disease severity
      - Tests for infection: molecular amplification (e.g., PCR); antigen
      - Test for prior infection: serology

# Context – applied public health surveillance

For local and national case reporting, case classification is highly specified, not for diagnosis

- **Probable**
  - Meets clinical criteria **AND** epidemiologic evidence with no confirmatory laboratory testing performed for COVID-19.
  - Meets presumptive laboratory evidence **AND** either clinical criteria **OR** epidemiologic evidence.
  - Meets vital records criteria with no confirmatory laboratory testing performed for COVID-19.
- **Confirmed**
  - Meets confirmatory laboratory evidence.
- **Vital Records Criteria**
  - A death certificate that lists COVID-19 disease or SARS-CoV-2 as a cause of death or a significant condition contributing to death.

CDC considers the following labs confirmatory:

*Detection of SARS-CoV-2 RNA in a clinical specimen using a molecular amplification detection test*

https://wwwn.cdc.gov/nndss/conditions/coronavirus-disease-2019-covid-19/case-definition/2020/

# COVID-19 real-world data studies - considerations

- Care setting – inpatient vs outpatient
  - Identify a cohort vs identification of events of interest
    - Likely need to use tests done outpatient to identify hospitalized COVID-19 patients
  - If restricting analysis to hospitalized patients, a positive antigen test alone may be acceptable
  - Could require a COVID-19 specific or applicable symptom diagnosis code with positive antigen
- Consider “any” positive a positive – i.e., do not limit to first test performed
  - Those with a negative antigen test often having a follow up PCR
- Lab related dates
  - Use specimen collection date or order date (test result date will often be days later)
- Symptom onset date not captured in claims or in structured EHR
  - Study follow up time start likely needs to be set to first positive test

# COVID-19 real-world data studies – considerations continued

- If need to identify those with high probability of confirmed disease, consider limiting to molecular amplification/PCR test positive
- If intent is to use COVID-19 as an outcome, to describe its prevalence, or incidence of a certain care pattern or complication amongst those with COVID, could consider a sensitivity analysis:
  - Primary result using PCR alone, secondary result using (PCR or antigen positive)
    - Reasoning: Those tested by antigen (point of care) may be different (e.g., milder illness, regionally specific,  non-English speakers/uninsured more likely to get tested at convenience clinics rather than hospitals, etc.)
  - Or require a diagnosis code if antigen positive only

# Select Sentinel COVID-19 work underway

- Algorithm validation
  - Assess the performance of diagnosis codes to identify COVID-19 where molecular test results are the gold standard
  - Important because not all data sources useful for COVID-19 studies will have complete lab data
- Natural history of COVID-19
  - Descriptive assessments of patients in a large network of hospitals
- Coagulopathy events
  - Examine the incidence of venous and arterial thrombotic events among COVID-19 patients

# Data Type Considerations in Data Networks

Jeffrey Brown, PhD and Judith C. Maro, PhD

August 19, 2020

# Commonly used electronic data types

- Insurance claims data
  - Open v. Closed Claims
- Electronic health records (inpatient and outpatient)
- Registries and patient-reported data

# Claims Data

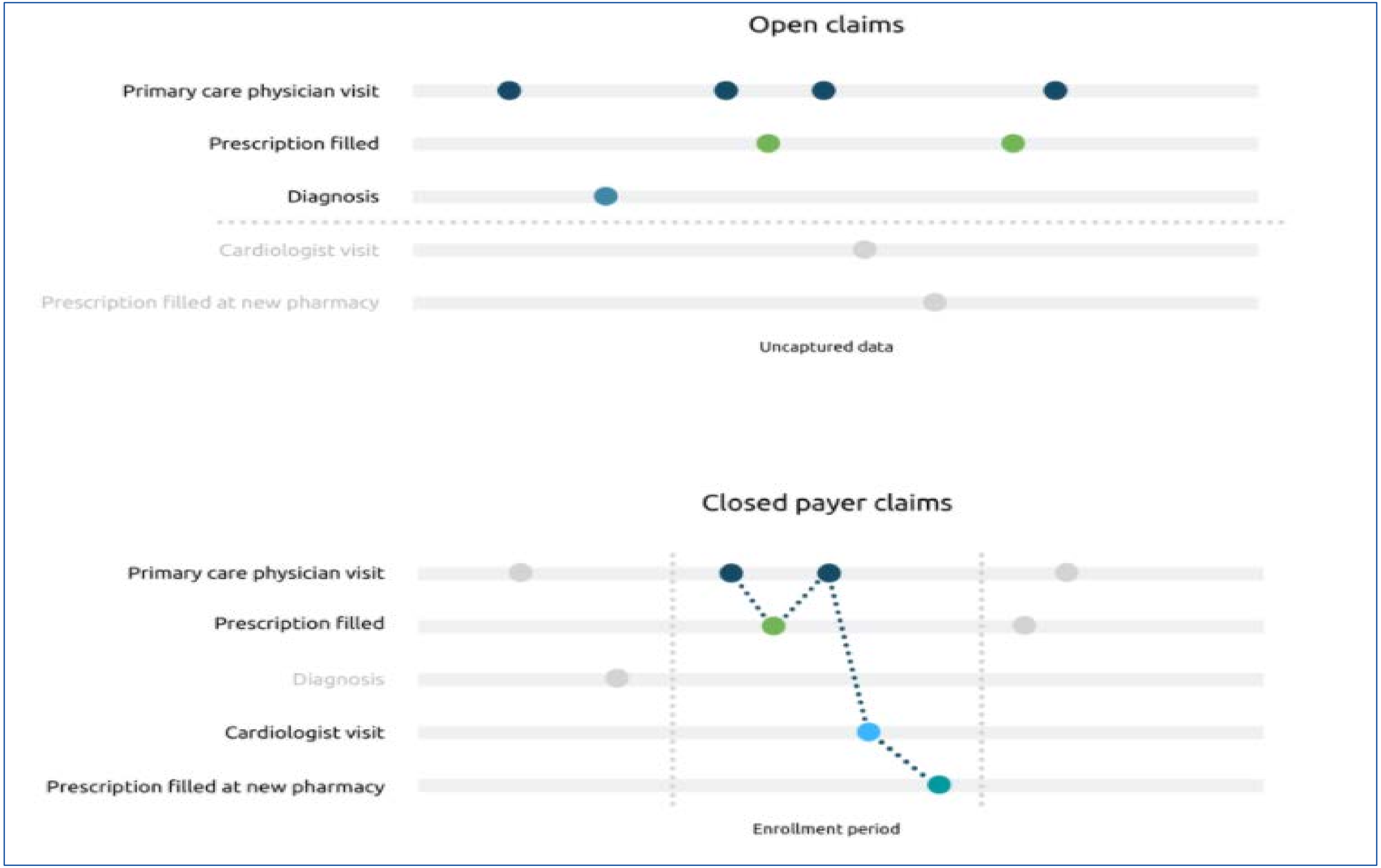
- Advantages
  - Nearly complete longitudinal capture of medically attended events
    - Absence of care is meaningful – it likely did not happen.
    - May not include paid-out-of-pocket dispensings or unattended illnesses (eg, migraine, nausea)
  - Well-defined person-time between enrollment start and end dates
  - Use of standardized coding systems
  - Large, geographically diverse populations

- Disadvantages
  - Inpatient pharmacy detail is bundled and therefore mostly unavailable
  - Clinical laboratory results (when available) are often missing not at random
  - Information only for insured individuals

# Open v Closed Claims

- Open Claims – Originate from clearinghouses (switches), pharmacies, software platforms
  - Advantages: Recency, broad coverage of individuals, not tied to a specific insurer
  - Disadvantages: Data instability related to lack of adjudication, unknown missingness of care, no enrollment information so lack of defined person-time
- Closed – Originate from insurance providers
  - Advantages: Defined person-time preserved via enrollment information, access to adjudicated information
  - Disadvantages: Timeliness

# Open v Closed Claims Illustration



**https://blog.healthverity.com/open-claims-vs-closed-payer-claims**

# Standalone EHR Systems (Outpatient or Inpatient)

- Advantages
  - Per encounter, deep insight into **structured** clinical data (e.g., laboratory results, other vital signs measures, allergies, etc.)
  - Availability of **unstructured** data within systems (harder to use across systems in a network)
  - Inpatient medication administration is well-defined
  - Date-Time stamps within an encounter are generally available

- Disadvantages
  - Lack of defined person-time: the absence of evidence **is not** evidence of absence
    - Means that look-back and follow-up windows are difficult to interpret
    - “Can’t do a rehospitalization study in a hospital database”

# How Does Missingness Impact Study Protocols?

- Take Aim 2: **Evaluate patient characteristics present prior to COVID-19 diagnosis as risk factors for arterial and venous thrombotic events (evaluated separately**)
  - What was the patient’s journey prior to hospitalization? Did they travel for treatment options? Availability of ICU beds? Did they go out of network? Had they moved in with loved ones?
  - If they are “new” to hospital EHR, have to “count” on hospital record (notes) to capture relevant recorded history, at time of urgency when healthcare providers have acknowledged incompleteness in clinical notetaking due to prioritizing patient care.
  - Patients may also be more incapacitated and fail to recall pertinent history.

- **Differential information records can lead to misclassification and introduce bias.**

# How Does Missingness Impact Study Protocols?

- Take Aim 3: **Compare the 90-day risk of arterial and venous thrombotic events (evaluated separately) between health plan members diagnosed with COVID-19 and those diagnosed with influenza.**
  - Inferential observational studies should use a target trail emulation paradigm (If I could do an RCT….)
  - In an RCT, randomization takes care of patient differences
  - In an observational study, inappropriate attention to patient differences (due to missing information) can cause patients to be sorted into inappropriate risk categories such that one cannot distinguish whether the outcome is indeed due to the exposure or confounding factors
    - Examples: Use of comorbidity scores and other utilization metrics to sort high and low risk patients
    - Remove patients with prior history of COVID treatment
    - Particularly important to understand history of TEE events and treatment for TEE events (pharmacy streams)

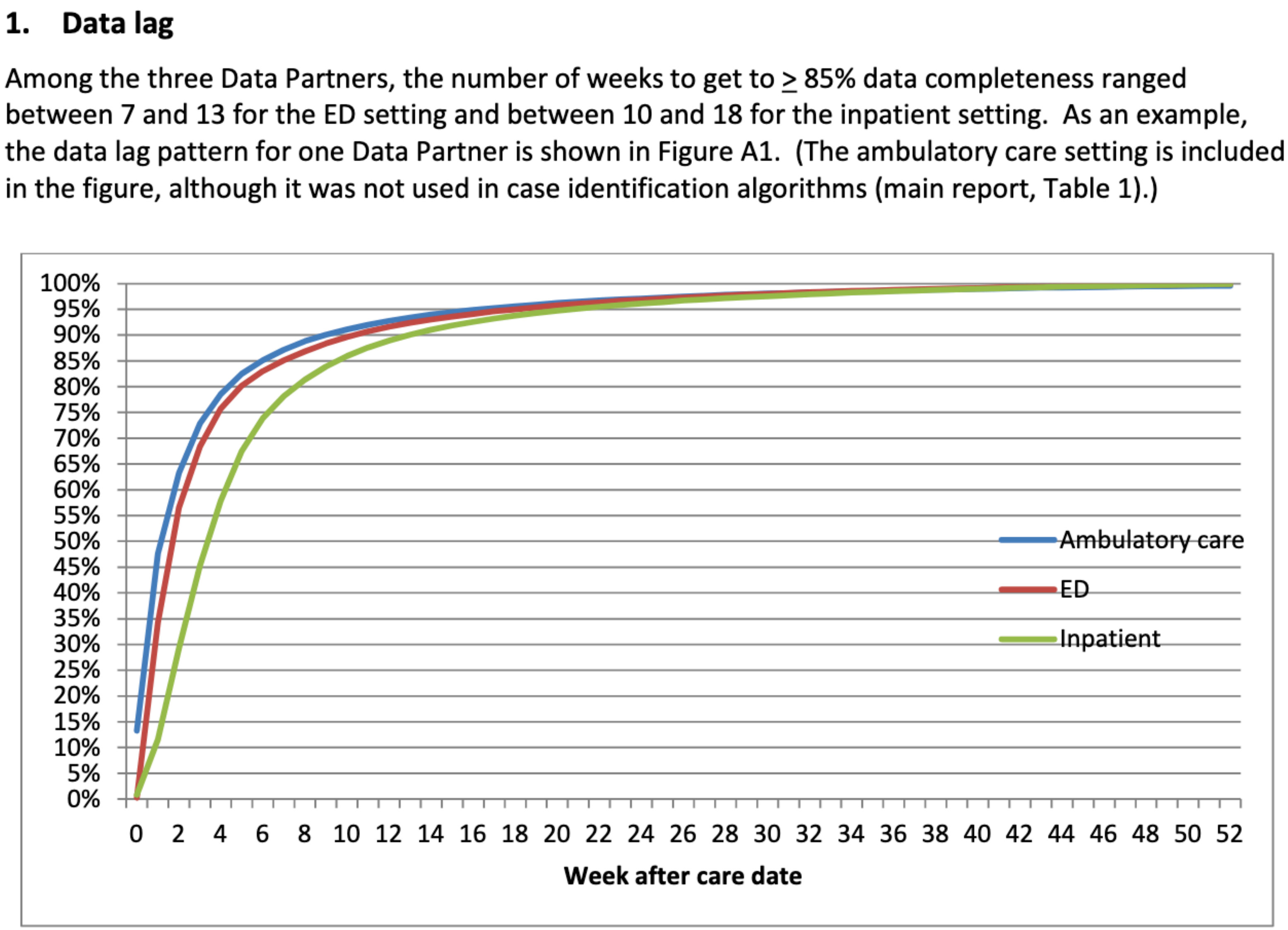
# Linked EHR-Claims Data

- Advantages
  - Breadth and depth for every patient especially in systems where patient is observed for their entire life
  - Holy grail of longitudinal observational data
- Disadvantages
  - Small sample sizes available with this data
  - Direct and on-demand access to unstructured clinical text is not widely available because of identifiability issues

# Registry Data / Patient Reported Data

- Advantages
  - Directly collected for a particular purpose (i.e., not secondary use)
  - Can get information not captured anywhere else (e.g., quality of life measures)
  - Can maintain relationship with patients
- Disadvantages
  - Only includes those that opt in
  - Possible recall bias because some information is based on patient self-report

# Considerations using the “freshest possible” claims data



### 1. Data lag

Among the three Data Partners, the number of weeks to get to ≥ 85% data completeness ranged between 7 and 13 for the ED setting and between 10 and 18 for the inpatient setting. As an example, the data lag pattern for one Data Partner is shown in Figure A1. (The ambulatory care setting is included in the figure, although it was not used in case identification algorithms (main report, Table 1).)

https://www.sentinelinitiative.org/sites/default/files/vaccines-blood-biologics/assessments/Mini-Sentinel_PRISM_Active-Influenza-Vaccine-Safety-Surveillance-Report_0.pdf

# Risks with “Rapid” Data: Strength of Measurement Error?

- Data are not “settled”
  - Incomplete picture of clinical experience
    - More pronounced with claims data that typically arrives in multiple streams with different lag points
    - Unadjudicated claims or open claims are subject to revision
    - EHR data is not immune: Post-discharge updates v. within-hospitalization updates
    - “Daily” feeds can capture intermediate variables (e.g., differential diagnosis subject to change)
  - Exposures may be assessed more completely than Outcomes (that may be later arriving)
    - Bias will depend on study design (e.g., active new user cohort v self-controlled) and whether or not delays are non-differential
    - There are some analytic solutions for partially-elapsed data capture but they do not apply to all designs
  - Trend analyses may be misleading without appropriate data lockpoints