# Data cleaning notes

This keeps notes of the shared dataset RSV\_deid\_2024\_may in the shared one drive folder. This dataset in excel sheet was shared with me by Dr. Carlos Oliveira on May 15, 2024.

This excel sheet contains 4 sheets

1. Dataset sheet: contains all the main data that we care about. This is the testing dataset, so the same patient can be tested more than once. This main file has 3108 unique patients and 96 duplicates, meaning some patients had more than one test within the study period..
2. Nirsevimab sheet: although in the dataset sheet there was nirsevimab immunization status, this nirsevimab sheet contains more complete records of the immunization status. If there is a conflict between these two, according to Carlos’ suggestion, trust this sheet. This sheet doesn’t have any duplicates, so one patient per row.
3. Oxygen sheet: this contains the oxygen flow given to the patient, the most useful variable is the extreme flow value.
4. Dictionary sheet

Here are some notes on the variables:

* StudyID: id for patient, but there might be some duplicates as one patient can be tested several times.
  + There are several patients who were tested multiple times on the same day.
* Collection date: sample collection date
* Accession #: this is used to link the specimen to Nate’s sequencing data.
* Viral test:
* RESPIRATORY VIRUS PCR PANEL  (YH VERIGENE)(LAB ORDER ONLY) [LAB10852] tests for 10 different viruses:

1. Rhinovirus
2. Human Metapneumovirus (HMPV)
3. Adenovirus
4. Influenza A
5. influenza B
6. Parainfluenza Virus 1
7. Parainfluenza Virus 2
8. Parainfluenza Virus 3
9. Parainfluenza Virus 4
10. RSV

* RESPIRATORY VIRUS PCR PANEL (BH GH LMW Q YH) [LAB3444] tests for 15 viruses: (what are the 15 viruses? Now assuming these tested for influenza AB and RSV in the recoding, NEED TO CONFIRM)

There are only 5 records in the cat. Virus detected are:

1. Coronavirus NL63
2. Parainfluenza Virus 3
3. Parainfluenza Virus 3 Adenovirus
4. Rhinovirus
5. Rhinovirus

* Encounter type: department where the patient was seen. Might need to recode this to “inpatient” and “outpatient”. (TBD: recoding of this)
* Admit date: there is some uncertainty about this variable. Cross check with sample collection date. Didn’t do much about this variable
* Collection department, Admit Location (didn’t use this), Admission Location:

Recoded to sample\_source, TBC.

* + Sample\_source is the same as Admission Location
  + If Admission location is missing, using Collection location as sample\_source.
  + Admission Location is “NE 45 WELLS ST”, this is a lab? (need to use this to fill in the missing encounter type)
* Encounter Type:
  + There are 35 missings for encounter type, how to fill in these?
  + Recode Encounter type to fewer categories: outpatient, inpatient (TBC)
    - Emergency 🡪 outpatient
    - Inpatient 🡪 inpatient
    - Specimen 🡪 outpatient
    - Observation 🡪 inpatient? (TBC)
    - Newborn 🡪 inpatient? (TBC)
    - Outpatient 🡪 outpatient
    - Hospital Outpatient Surgery 🡪 outpatient
* ICU admission date:
  + Generated date\_icu\_admit: the same as ICU Admission Date, N/A as NA;
  + Generated date\_icu\_admit\_related: date of ICU admission that can be linked with the testing result (within a week of Collection Date? TBC)
  + Generated icu\_admitted\_related: binary, indicating this record was admitted to the ICU and the ICU admission is related to the testing result.
* ICU LOS:
  + Generated two new variables: icu\_los\_hours, icu\_los\_days
  + There are several patients with los longer than 100 days.
* Gest age:
  + 718 missing this var
  + CDC’s classification: (generated gest\_age\_cat)
    - >=42: post term
    - 41: late term
    - 39-40: full term
    - 37-38: early term
    - < 37: preterm
* Birth weight: what is the unit?
* Race and ethnicity:

The answer for race is messy, several (conflicting) answers at the same time.

* Primary insurance & payor

This can be used to classify SES(?). But how to classify this? There are about 89 records that have different entries for Primary Insurance and Payor, use which one? Currently, generate new variable: insurance\_type:

* + Medicare / Medicaid 🡪 public insurance
  + If missing for both 🡪 uninsured
  + Some notes about some specific insurance classification:
    - LOCAL 1199SEIU BENEFIT FUNDS 🡪 uninsured
    - Assume that those with “MCD MGD” (Medicare Coverage Database managed ??) are Medicare (public insurance) 🡪 public
* Last PCP visit & Last well child visit & recent contact (pending)

Variables related to the medical history and diagnosis:

1. Medical history (past)
2. Problem list: running list (past + current, more from past)
3. Primary problem: is this related to this visit/testing (current)
4. ED Dx: Current
5. Admit Dx: current
6. Hospital problem list: current+past hospital problem list (more from current).

From 1-6 variables, extracted potential risk factrs (e.g. congenital heart disease, preterm …); from 3-6 variables, extracted related respiratory symptoms (difficulty breathing, bronchiolitis,…)

TBC: There are several records with more than one testing results on the same day?

==============================================================================

# Inclusion and exclusion criteria

The study population is patients with suspected respiratory infection who are eligible for administration of Nirsevimab. The first step is to examine all the testing records and confirm that the records are encounters that are potentially due to respiratory infection.

1. Age:

Age at testing < 12 months? who are eligible for mAb in this dataset? What if took the mab at age of 8 months, and then tested at the end of the season? Do we really need to remove ?

I checked the birth date of records, there are 1049 records with birth date before 2.1.2023 (8 months before October 2023 when mAb became available). For these records, there are 818 that I can’t be sure whether they are “at high risk” thus eligible for mAb immunization. Among these 818, there are 213 cases (I assume not deleting these?).

1. Infants with suspected respiratory infection:

There are ~200 records that no patterns of respiratory symptoms can be detected (table 20).

==============================================================================

# Case and control selection

Reference: <https://pubmed.ncbi.nlm.nih.gov/36454733/>

## Test-positive cases:

A case of test-positive is defined as a positive RSV test.

## Test-negative controls:

We define test-negative controls as negative RSV tests

* collected at any time if this patient was tested only once, OR
* collected > 7 days prior this patient's next test AND collected > 7 days after this patient's last test (if the patient was tested multiple times)

## Other notes:

* An individual patient is allowed to contribute as both case and control, as the analysis is to be conducted at test-level, instead of patient-level.
* Each patient can contribute to up to three negative tests (controls).
* If a person had more than one negative test within 7-day period, one random test was selected during the period as a control.

==============================================================================

# Exposure

* Immunized with mAb or not
* Time between immunization and sample collection
  + Get how protection wanes over time.

==============================================================================

# Potential risk factors and underlying conditions

(Reference: <https://doi.org/10.1067/S0022-3476(03)00511-0>)

* Male sex
* Age < 6 months (age at sample collection)
  + Most infants were immunized around birth time.
* Birth in the first half of the RSV season
  + (August-November vs. December-July?)
* Gestational age (need to classify this)
* SES related:
  + Insurance type
  + Race
* Low birth weight (TBC)
* Underlying conditions:
  + Congenital heart disease
  + History of CVD?
  + History of pulmonary diseases?
  + History of GI diseases?
  + (History of) Malnutrition? (haven’t extracted this)

==============================================================================

# Measures of outcome

Related variables:

* RSV testing results: positive or negative
* Encounter\_type
  + Inpatient
  + Outpatient
* Hosp LoS: range from 1-287 days.
* ICU:
  + Whether admitted into ICU
  + ICU LoS
* Symptoms extracted (symp\_xx):
  + URT distress
  + LRT distress
  + Fever
  + Cough
  + Wheezing
  + Breathing problems
  + Bronchiolitis
  + Sepsis (combined with breathing problems?)
  + Apnea (combined with breathing problems?)
* Maximum oxygen flow rate
  + Categorical (which threshold to use?)

Potential outcomes:

* Based on RSV test result only:
  + Case: RSV+
  + Control: RSV-
* Base on RSV result and encounter type:
  + Case: RSV+ & encounter\_type == inpatient
  + Control: All others
* Based on RSV test result and hosp LoS:
  + RSV+ & hosp LoS > ?? days
* RSV test result + ICU admission
  + Case: RSV+ & ICU admission
  + Control: all others
* RSV test results + severity based on symptoms (TBC)
  + RSV+ & only URT distress reported
  + RSV+ & LRT distress reported
* RSV test results + severity based on maximum oxygen flow rate
  + RSV+ & flowrate > ? L/min