# Overview of the Project

CDC’s [Surveillance for Emerging Threats to Mothers and Babies Network (SET-NET)](https://www.cdc.gov/ncbddd/set-net/index.html) detects the effects of health threats on pregnant people and their babies by collecting data during pregnancy through childhood on pregnancies exposed to infectious diseases and uses evidence-based, actionable information to help save and improve the lives of pregnant people, mothers and babies. State, local, and territorial health departments work with CDC to identify exposures to infectious diseases during pregnancy and link them with health outcomes of pregnant people and infants.

As part of SET-NET’s COVID-19 surveillance, information on the clinical course of illness was collected from the medical records of pregnant people. To classify the severity of illness, the SET-NET natural language processing (NLP) code uses regular expressions (string searching algorithms) to identify COVID-19 symptoms, drugs, and other relevant information in the source data.

# Code Development Process

The source data file is in CSV format and contains Boolean, numeric, and textual field values. The regular expressions were developed by extracting the text fields from a subset of the source data, writing all texts for a given field to a separate file, and studying the forms of expression found in each file. Usage examples were collected for each symptom or medication and arranged in a text editor so that common forms of expression became evident. A few special-case regular expressions were needed to handle uncommon forms and to prevent some erroneous text captures such as ICU admissions and health information not pertaining to the pregnant person. The regular expressions were also expanded to include examples not found in the review of the subset of text, but which were deemed to be valid forms of expression for the particular symptom or drug in question.

Regular expressions for capturing oxygen saturation information were adapted from those developed for the open-source clinical phenotyping project [ClarityNLP](https://github.com/ClarityNLP/ClarityNLP). These regular expressions capture information on the use of oxygen devices, flow rates, and saturations. The code applies standard conversion formulas to the captured values to determine whether the patient is using a high-flow Oxygen device, and, if so, to quantify the flow rate.

An iterative process was used for tuning the regular expressions and improving the accuracy of the code. Iterations were performed on a given input file as well as on input files released on subsequent dates. Differences between the results of the NLP code and the existing CDC algorithm were captured and reviewed to determine the source of the discrepancy. Over a period of several months, this process of iterative improvement and tuning reduced the error rate to an acceptable level (<1%).

# Installation Instructions

The SET-NET NLP code has been implemented in the python language and therefore requires a python runtime environment. We recommend the use of [Anaconda](https://www.anaconda.com/) environments (command name *conda*) for python development. Python has a standard package manager called *pip*, but conda is able to resolve dependencies and package version conflicts much better than pip. Conda also provides a conda-compatible replacement for pip.

There are two Anaconda distributions: *conda*, a full-featured numerical computing, machine learning, and statistical software stack; and *miniconda*, a “lite” version of the full Anaconda installation. Only one of these distributions should be installed at a time, and they both use the *conda* command for configuration. The following instructions assume the use of miniconda.

This software requires python version 3.7 or greater.

## Install the Miniconda Python Distribution

### For a Windows installation:

1. Download the latest miniconda installer from <https://docs.conda.io/en/latest/miniconda.html>. Scroll through the list of packages and find a recent version for python 3.7 or greater.
2. Run the installer and accept the defaults unless you have specific reasons for changing them.
3. From the Start menu, find the Anaconda entry and run the item called “Anaconda Prompt (miniconda3)”.
4. Update the miniconda installation:

conda update conda

### For a Linux or Mac installation:

1. Download the latest miniconda package:

wget <https://repo.anaconda.com/miniconda/Miniconda3-latest-Linux-x86_64.sh>

1. Install miniconda:

bash Miniconda3-latest-Linux-x86\_64.sh

<Accept the default install location and answer yes to all questions.>

1. Activate the installer’s modifications to your .bash\_profile file by closing the terminal window and starting a new terminal window.
2. Test the installation by listing the installed conda packages:

conda list

1. If your system cannot find the conda executable, then something went wrong with the modifications to your PATH environment variable. Either edit the path by hand or consult the Anaconda documentation for further instructions.
2. Update the installation:

conda update conda

## Create a Conda Environment

A dedicated conda “environment” will be created for running the code. Conda environments are isolated from each other, can be activated and deactivated easily, and can be configured and updated independently from other environments. The isolation helps prevent incompatible software upgrades and other problems caused by shared system library folders.

The environment will be called setnet and must be explicitly activated to run the code.

From either a command terminal on Mac or Linux, or the miniconda prompt on Windows, run this command and accept the defaults when prompted:

conda create –-name setnet

## Install Required Packages

The next step is to activate the setnet environment and download and install the required python packages into it:

conda activate setnet

conda install -c anaconda jupyter

conda install -c anaconda spacy

With the required packages installed, download and install an English language data file for the SpaCy NLP library:

python -m spacy download en\_core\_web\_md

Note: if installation fails, try running the command again.

## Download the SET-NET NLP Code

The code is housed in a Github repository which can be downloaded to a local hard drive. Open a command terminal (or Miniconda prompt on Windows) and browse to a disk location where the SET-NET code should be downloaded. Clone the git repository with this command:

git clone https://<TBD>

## Run the Code

Change directories to the location of the cloned repo on your system.

cd <path>

Activate the setnet environment and launch Jupyter with this command:

conda activate setnet

jupyter notebook

From the main Jupyter window, open the Jupyter notebook file SetNet\_csv.ipynb.

When running the code for the first time, set the path to the input file and output folder appropriate for your system.

Run the notebook by selecting Restart & Clear Output from the Kernel menu, then Run All from the Cell menu. The notebook should run to completion.

# Sample Data

A dataset with 200 rows of synthetic data is provided. These observations are simulated and should not be treated as real data.

## Synthetic Data Codebook

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| # | Variable | Type | Values | Label |
| 1 | mg\_death\_dx | Char | Text | If yes, cause(s) of maternal death: |
| 2 | mv\_sx | Num | 1=yes, 0=no | Symptoms present during course of illness: |
| 3 | mv\_sx\_fever | Num | 1=yes, 0=no | Fever >100.4F (38C) |
| 4 | mv\_sx\_sfever | Num | 1=yes, 0=no | Subjective fever (felt feverish) |
| 5 | mv\_sx\_chills | Num | 1=yes, 0=no | Chills |
| 6 | mv\_sx\_rigors | Num | 1=yes, 0=no | Rigors |
| 7 | mv\_sx\_myalgia | Num | 1=yes, 0=no | Muscle aches (myalgia) |
| 8 | mv\_sx\_runnose | Num | 1=yes, 0=no | Runny nose (rhinorrhea) |
| 9 | mv\_sx\_sthroat | Num | 1=yes, 0=no | Sore throat |
| 10 | mv\_sx\_taste | Num | 1=yes, 0=no | New olfactory and taste disorder(s) |
| 11 | mv\_sx\_fatigue | Num | 1=yes, 0=no | Fatigue |
| 12 | mv\_sx\_cough | Num | 1=yes, 0=no | Cough (New onset or worsening of chronic cough) |
| 13 | mv\_sx\_wheezing | Num | 1=yes, 0=no | Wheezing |
| 14 | mv\_sx\_sob | Num | 1=yes, 0=no | Shortness of breath (dyspnea) |
| 15 | mv\_sx\_breath | Num | 1=yes, 0=no | Difficulty breathing |
| 16 | mv\_sx\_chest | Num | 1=yes, 0=no | Chest pain |
| 17 | mv\_sx\_nauvom | Num | 1=yes, 0=no | Nausea or vomiting |
| 18 | mv\_sx\_head | Num | 1=yes, 0=no | Headache |
| 19 | mv\_sx\_abdom | Num | 1=yes, 0=no | Abdominal pain |
| 20 | mv\_sx\_diarrhea | Num | 1=yes, 0=no | Diarrhea (>= 3 loose/looser than normal stools/24hr period) |
| 21 | mv\_sx\_oth | Num | 1=yes, 0=no | Other symptoms |
| 22 | mv\_sx\_oth\_sp | Char | Text | If other, specify: |
| 23 | mv\_comp\_pna | Num | 1=yes, 0=no | Pneumonia? |
| 24 | mv\_comp\_ards | Num | 1=yes, 0=no | Acute respiratory distress syndrome? |
| 25 | mv\_comp\_mv | Num | 1=yes, 0=no | Mechanical ventilation (MV)/intubation? |
| 26 | mv\_comp\_ecmo | Num | 1=yes, 0=no | Extracorporeal membrane oxygenation (ECMO)? |
| 27 | mv\_comp\_oth\_sp | Char | Text | If other, specify: |
| 28 | mv\_icu | Num | 1=yes, 0=no | Was the mother admitted to an intensive care unit (ICU) for COVID-19? |
| 29 | cv\_sn\_pos\_spec1 | Num | Date | Date of first SARS-CoV-2 positive result |
| 30 | mv\_sx\_dt | Num | Date | Date of symptom onset |
| 31 | mg\_decon\_icuadm\_dt | Num | Date | Date of ICU admission |
| 32 | mg\_notes | Char | Text | Abstractor notes: |
| 33 | mv\_tx\_rem | Num | 1=yes, 0=no | Remdesivir |
| 34 | mv\_tx\_oth\_sp1 | Char | 1=yes, 0=no | If yes, specify medication 1: |
| 35 | mv\_tx\_oth\_sp2 | Char | 1=yes, 0=no | If yes, specify medication 2: |
| 36 | mv\_tx\_oth\_sp3 | Char | 1=yes, 0=no | If yes, specify medication 3: |

# **Licenses and Disclaimers**

## **Public Domain**

This repository constitutes a work of the United States Government and is not subject to domestic copyright protection under 17 USC § 105. This repository is in the public domain within the United States, and copyright and related rights in the work worldwide are waived through the [CC0 1.0 Universal public domain dedication](https://creativecommons.org/publicdomain/zero/1.0/). All contributions to this repository will be released under the CC0 dedication. By submitting a pull request you are agreeing to comply with this waiver of copyright interest.

## **License**

The repository utilizes code licensed under the terms of the Apache Software License and therefore is licensed under ASL v2 or later.

This source code in this repository is free: you can redistribute it and/or modify it under the terms of the Apache Software License version 2, or (at your option) any later version.

This source code in this repository is distributed in the hope that it will be useful, but WITHOUT ANY WARRANTY; without even the implied warranty of MERCHANTABILITY or FITNESS FOR A PARTICULAR PURPOSE. See the Apache Software License for more details.

You should have received a copy of the Apache Software License along with this program. If not, see <http://www.apache.org/licenses/LICENSE-2.0.html>

The source code forked from other open source projects will inherit its license.

## **Privacy**

This repository contains only non-sensitive, publicly available data and information. All material and community participation is covered by the Surveillance Platform [Disclaimer](https://github.com/CDCgov/template/blob/master/DISCLAIMER.md) and [Code of Conduct](https://github.com/CDCgov/template/blob/master/code-of-conduct.md). For more information about CDC's privacy policy, please visit <http://www.cdc.gov/privacy.html>.

## **Contributing**

Anyone is encouraged to contribute to the repository by [forking](https://help.github.com/articles/fork-a-repo) and submitting a pull request. (If you are new to GitHub, you might start with a [basic tutorial](https://help.github.com/articles/set-up-git).) By contributing to this project, you grant a world-wide, royalty-free, perpetual, irrevocable, non-exclusive, transferable license to all users under the terms of the [Apache Software License v2](http://www.apache.org/licenses/LICENSE-2.0.html) or later.

All comments, messages, pull requests, and other submissions received through CDC including this GitHub page are subject to the [Presidential Records Act](http://www.archives.gov/about/laws/presidential-records.html) and may be archived. Learn more at <http://www.cdc.gov/other/privacy.html>.

## **Records**

This repository is not a source of government records, but is a copy to increase collaboration and collaborative potential. All government records will be published through the [CDC web site](http://www.cdc.gov/).

## **Notices**

Please refer to [CDC's Template Repository](https://github.com/CDCgov/template) for more information about [contributing to this repository](https://github.com/CDCgov/template/blob/master/CONTRIBUTING.md), [public domain notices and disclaimers](https://github.com/CDCgov/template/blob/master/DISCLAIMER.md), and [code of conduct](https://github.com/CDCgov/template/blob/master/code-of-conduct.md).