Shedding Hub: An Open Science Portal for Existing Pathogen Shedding Data and Models

Yuke (Andrew) Wang, PhD, MSPH [1,2,3]
yuke.wang@emory.edu

Till Hoffmann, PhD, MPhys [4]
thoffmann@hsph.harvard.edu

[1] Hubert Department of Global Health, Rollins School of Public Health, Emory University
[2] Emory Center for Infectious Disease Modeling and Analytics and Training Hub, Emory University
[3] Center of Global Safe Water, Sanitation, and Hygiene, Emory University
[4] Department of Biostatistics, Harvard T.H. Chan School of Public Health, Harvard University

IDM Annual Symposium, October 1st, 2024

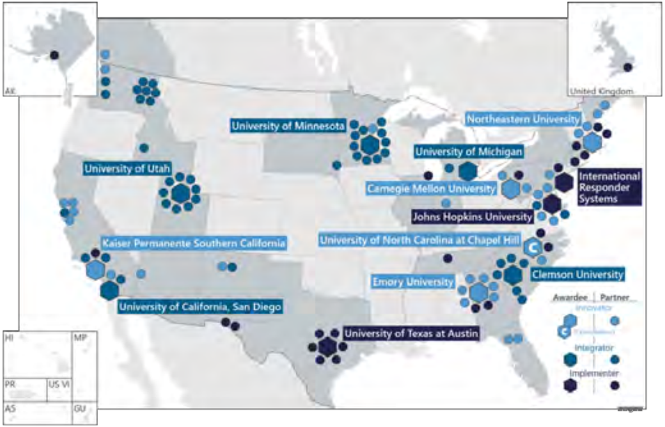
Center for Infectious Disease Modeling and Analytics & Training Hub (CIDMATH)

InsightNet (National Outbreak Analytics & Disease Modeling Network)

Illustration of CFA's partners working to detect and control an infectious disease outbreak. https://www.cdc.gov/insight-net/php/about/index.html

- Established in 2023 by CDC Center for Forecasting and Outbreak Analytics (CFA)
- Focuses on training, analytical tool development, and advancing the analysis and use of data about infectious disease spread
- brings together >100 academic and private partners and health departments

CIDMATH

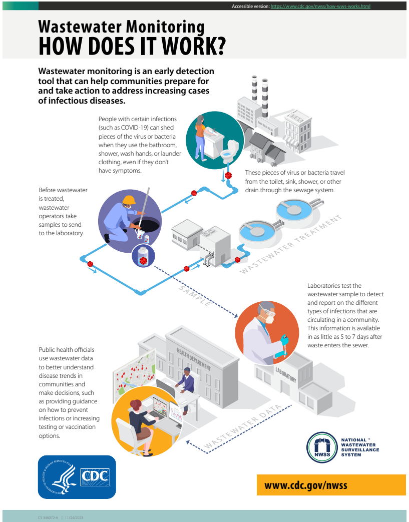


InsightNet's partners. https://www.cdc.gov/insight-net/php/about/index.html

- 13 centers funded through the CFA
- Emory CIDMATH is a Center of Innovation
- Partners include Georgia DPH, Kaiser Permanente of GA, and the Georgia Emerging Infections Program (EIP)

Motivation

Wastewater Surveillance



Accessible version: https://www.cdc.gov/nwss/how-wws-works.html

Wastewater Monitoring
HOW DOES IT WORK?

Wastewater monitoring is an early detection tool that can help communities prepare for and take action to address increasing cases of infectious diseases.

People with certain infections (such as COVID-19) can shed pieces of the virus or bacteria when they use the bathroom, shower, wash hands, or launder clothing, even if they don't have symptoms.

These pieces of virus or bacteria travel from the toilet, sink, shower, or other drain through the sewage system.

Before wastewater is treated, wastewater operators take samples to send to the laboratory.

Laboratories test the wastewater sample to detect and report on the different types of infections that are circulating in a community. This information is available in as little as 5 to 7 days after waste enters the sewer.

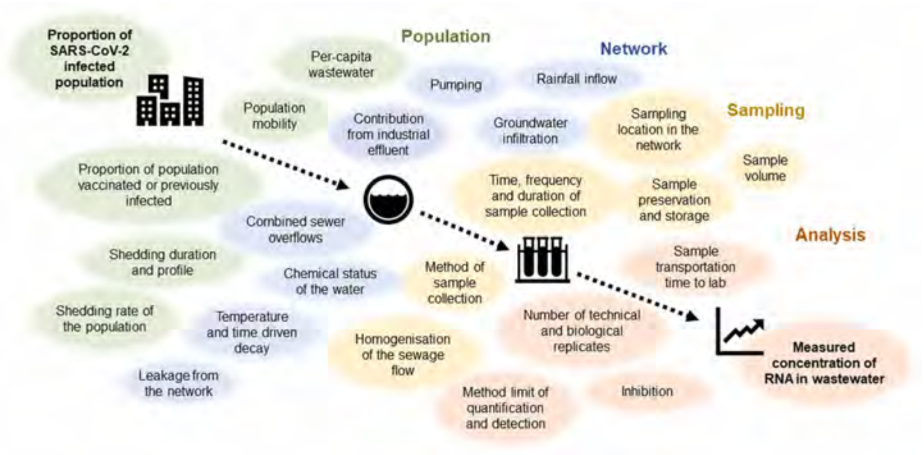
Public health officials use wastewater data to better understand disease trends in communities and make decisions, such as providing guidance on how to prevent infections or increasing testing or vaccination options.

www.cdc.gov/nwss

CS 346072-A | 11/24/2023

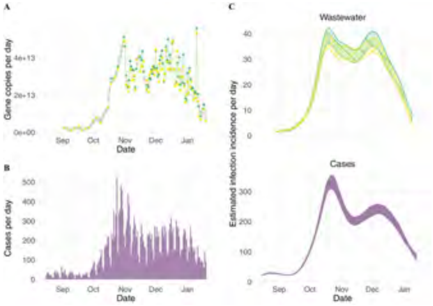
- Wastewater surveillance (WWS) is an approach for monitoring specific pathogen(s) circulating in a population by examining sewage samples.
- Pathogens shed in feces, urine, sputum, and vomit are aggregated in the sewage system.
- A well-designed WWS provides actionable information, including **certification of elimination, early warning, nowcasting/predicting trends, and identification of hotspots**.

Sources of Uncertainty for WWS


Proportion of SARS-CoV-2 infected population
Population
Per-capita wastewater
Network
Pumping
Rainfall inflow
Population mobility
Contribution from industrial effluent
Groundwater infiltration
Sampling location in the network
Sampling
Sample volume
Proportion of population vaccinated or previously infected
Time, frequency and duration of sample collection
Sample preservation and storage
Combined sewer overflows
Analysis
Sample transportation time to lab
Shedding duration and profile
Chemical status of the water
Method of sample collection
Shedding rate of the population
Temperature and time driven decay
Number of technical and biological replicates
Homogenisation of the sewage flow
Measured concentration of RNA in wastewater
Leakage from the network
Method limit of quantification and detection
Inhibition

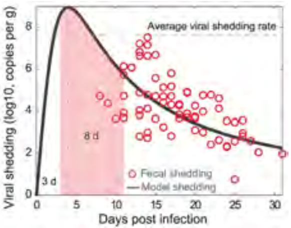
Wade et al., 2022

Wastewater-based Epidemiology



Huisman et al., 2022

- The interpretation of WWS results in terms of clinical cases depends on quantitatively well-characterized shedding information for pathogens and biomarkers



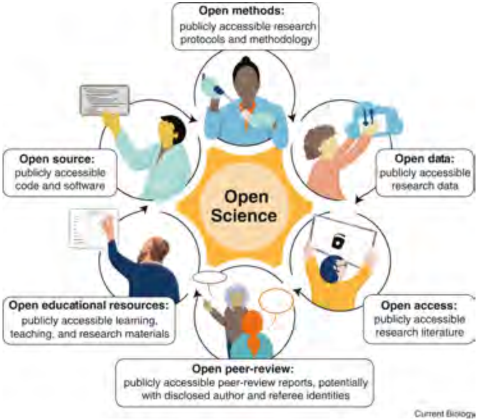
Phan et al., 2023

Current Knowledge of Shedding

- Most shedding data have been collected in clinical studies or human challenging studies
- Limited raw shedding data are openly available
- Shedding data are not standardized
- No public accessible tutorial or feasible tool for modeling shedding dynamics
- No community portal for learning and contributing shedding knowledge

Shedding Hub

Open Science

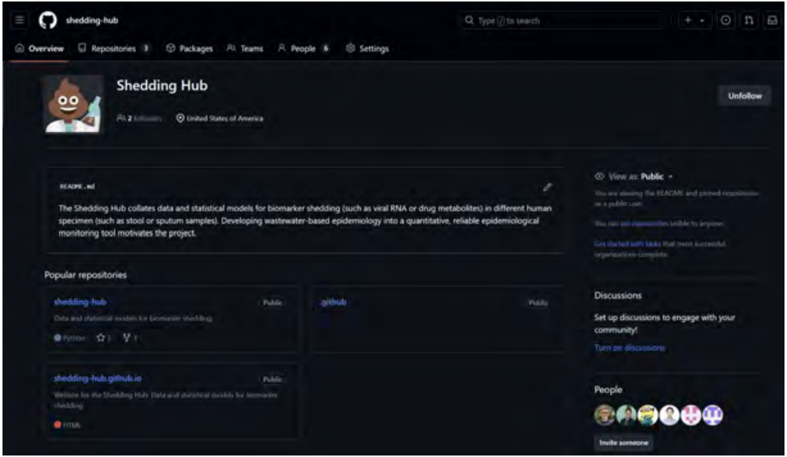

Open methods: publicly accessible research protocols and methodology
Open data: publicly accessible research data
Open access: publicly accessible research literature
Open peer-review: publicly accessible peer-review reports, potentially with disclosed author and referee identities
Open educational resources: publicly accessible learning, teaching, and research materials
Open source: publicly accessible code and software
Open Science
Current Biology

Principles of Open Science:

- Accessible
- Verfiable
- Reliable
- Reproducible
- Sustainable

Bertram et al., 2023

Shedding Hub Organization



Shedding Hub

2 followers · United States of America

README.md

The Shedding Hub collates data and statistical models for biomarker shedding (such as viral RNA or drug metabolites) in different human specimen (such as stool or sputum samples). Developing wastewater-based epidemiology into a quantitative, reliable epidemiological monitoring tool motivates the project.

Popular repositories

shedding-hub · Public
Data and statistical models for biomarker shedding.
Python · ☆ 1 · 1

.github · Public

shedding-hub.github.io · Public
Website for the Shedding Hub: Data and statistical models for biomarker shedding
HTML

View as: Public

You are viewing the README and pinned repositories as a public user.

You can pin repositories visible to anyone.

Get started with tasks that most successful organizations complete.

Discussions

Set up discussions to engage with your community!

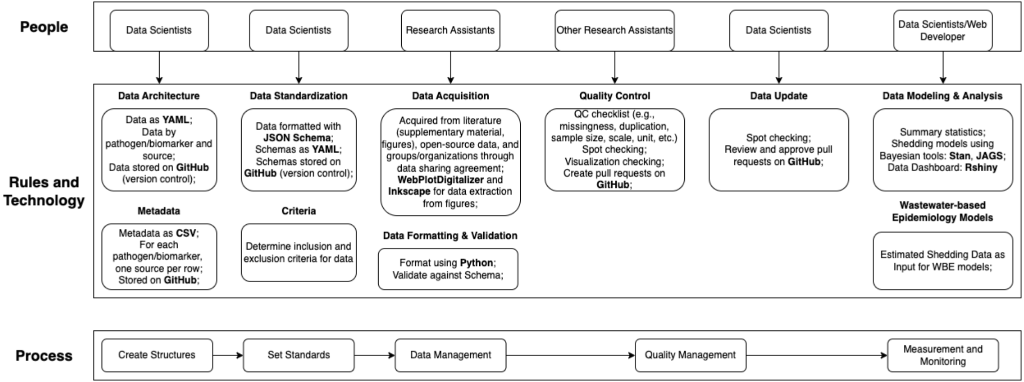
Turn on discussions

People

Invite someone

https://github.com/shedding-hub

Shedding Hub Data Governance Framework



People

Rules and Technology

Process

Data Scientists → **Data Architecture**

Data as **YAML**; Data by pathogen/biomarker and source; Data stored on **GitHub** (version control);

Metadata

Metadata as **CSV**; For each pathogen/biomarker, one source per row; Stored on **GitHub**;

Data Scientists → **Data Standardization**

Data formatted with **JSON Schema**; Schemas as **YAML**; Schemas stored on **GitHub** (version control);

Criteria

Determine inclusion and exclusion criteria for data

Research Assistants → **Data Acquisition**

Acquired from literature (supplementary material, figures), open-source data, and groups/organizations through data sharing agreement; **WebPlotDigitalizer** and **Inkscape** for data extraction from figures;

Data Formatting & Validation

Format using **Python**; Validate against Schema;

Other Research Assistants → **Quality Control**

QC checklist (e.g., missingness, duplication, sample size, scale, unit, etc.) Spot checking; Visualization checking; Create pull requests on **GitHub**;

Data Scientists → **Data Update**

Spot checking; Review and approve pull requests on **GitHub**;

Data Scientists/Web Developer → **Data Modeling & Analysis**

Summary statistics; Shedding models using Bayesian tools: **Stan**, **JAGS**; Data Dashboard: **Rshiny**

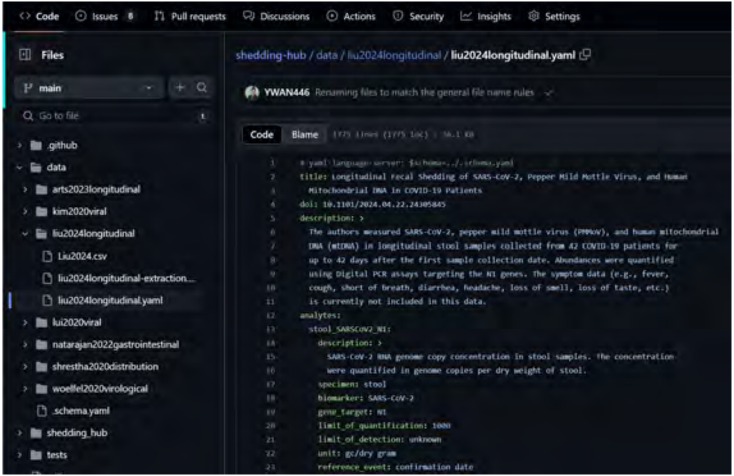
Wastewater-based Epidemiology Models

Estimated Shedding Data as Input for WBE models;

Create Structures → Set Standards → Data Management → Quality Management → Measurement and Monitoring

This Data Governance Framework is an extension of the tillahoffmann/shedding repository.

Data Structure



- .github
- data
  - arts2023longitudinal
  - kim2020viral
  - liu2024longitudinal
    - Liu2024.csv
    - liu2024longitudinal-extraction...
    - liu2024longitudinal.yaml
  - lui2020viral
  - natarajan2022gastrointestinal
  - shrestha2020distribution
  - woelfel2020virological
  - .schema.yaml
- shedding_hub
- tests

shedding-hub / data / liu2024longitudinal / liu2024longitudinal.yaml

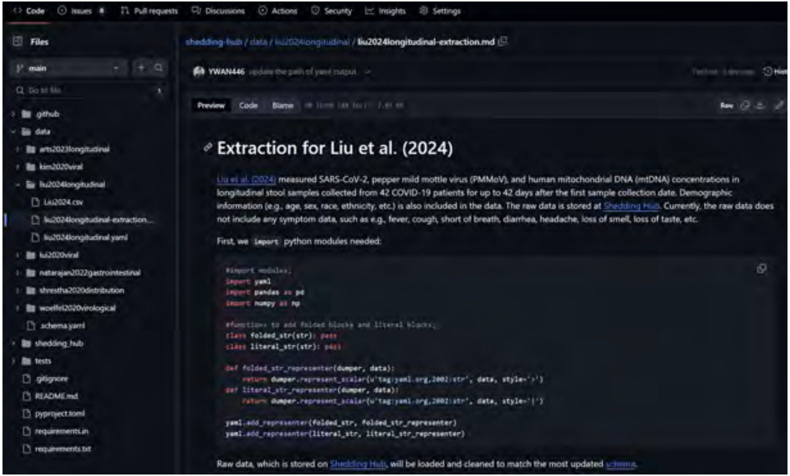
YWAN446 Renaming files to match the general file name rules

1775 lines (1775 loc) · 56.1 KB

```yaml
# yaml-language-server: $schema=../.schema.yaml
title: Longitudinal Fecal Shedding of SARS-CoV-2, Pepper Mild Mottle Virus, and Human
  Mitochondrial DNA in COVID-19 Patients
doi: 10.1101/2024.04.22.24305845
description: >
  The authors measured SARS-CoV-2, pepper mild mottle virus (PMMoV), and human mitochondrial
  DNA (mtDNA) in longitudinal stool samples collected from 42 COVID-19 patients for
  up to 42 days after the first sample collection date. Abundances were quantified
  using Digital PCR assays targeting the N1 genes. The symptom data (e.g., fever,
  cough, short of breath, diarrhea, headache, loss of smell, loss of taste, etc.)
  is currently not included in this data.
analytes:
  stool_SARSCoV2_N1:
    description: >
      SARS-CoV-2 RNA genome copy concentration in stool samples. The concentration
      were quantified in genome copies per dry weight of stool.
    specimen: stool
    biomarker: SARS-CoV-2
    gene_target: N1
    limit_of_quantification: 1000
    limit_of_detection: unknown
    unit: gc/dry gram
    reference_event: confirmation date
```

- Includes raw data files, markdown files to process data, standardized data in YAML format, and a schema file
- Public **accessible** on GitHub

Data Processing



Extraction for Liu et al. (2024)

Liu et al. (2024) measured SARS-CoV-2, pepper mild mottle virus (PMMoV), and human mitochondrial DNA (mtDNA) concentrations in longitudinal stool samples collected from 42 COVID-19 patients for up to 42 days after the first sample collection date. Demographic information (e.g., age, sex, race, ethnicity, etc.) is also included in the data. The raw data is stored at Shedding Hub. Currently, the raw data does not include any symptom data, such as e.g., fever, cough, short of breath, diarrhea, headache, loss of smell, loss of taste, etc.

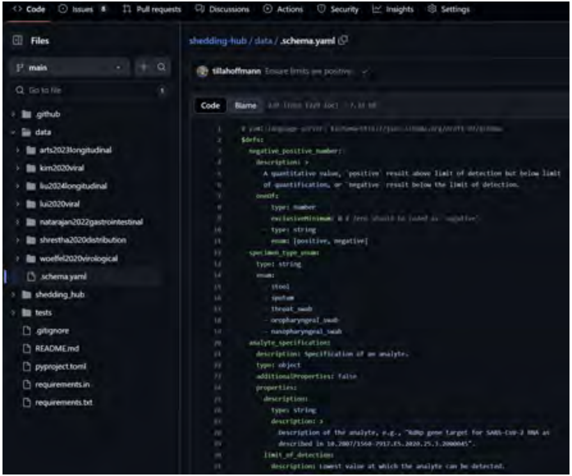
First, we `import` python modules needed:

```
#import modules;
import yaml
import pandas as pd
import numpy as np

#functions to add folded blocks and literal blocks;
class folded_str(str): pass
class literal_str(str): pass

def folded_str_representer(dumper, data):
    return dumper.represent_scalar(u'tag:yaml.org,2002:str', data, style='>')
def literal_str_representer(dumper, data):
    return dumper.represent_scalar(u'tag:yaml.org,2002:str', data, style='|')

yaml.add_representer(folded_str, folded_str_representer)
yaml.add_representer(literal_str, literal_str_representer)
```

Raw data, which is stored on Shedding Hub, will be loaded and cleaned to match the most updated schema.

- **Verifiable** with information sources
- **Reproducible** with open code

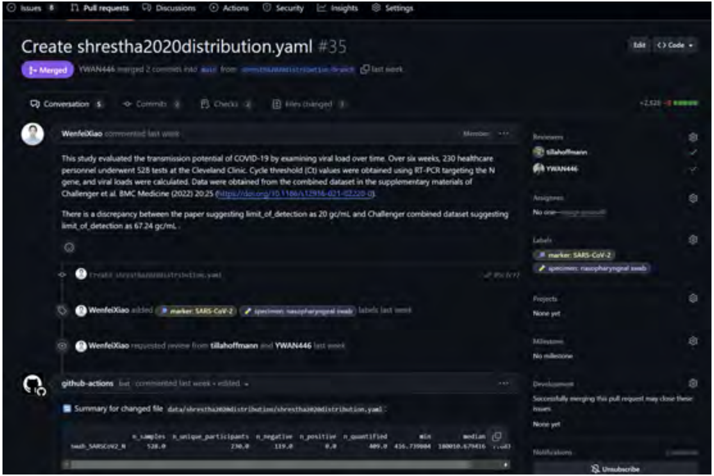
Data Standardization



```yaml
# yaml-language-server: $schema=https://json-schema.org/draft/2020-12/schema
$defs:
  negative_positive_number:
    description: >
      A quantitative value, `positive` result above limit of detection but below limit
      of quantification, or `negative` result below the limit of detection.
    oneOf:
      - type: number
        exclusiveMinimum: 0 # Zero should be coded as `negative`.
      - type: string
        enum: [positive, negative]
  specimen_type_enum:
    type: string
    enum:
      - stool
      - sputum
      - throat_swab
      - oropharyngeal_swab
      - nasopharyngeal_swab
  analyte_specification:
    description: Specification of an analyte.
    type: object
    additionalProperties: false
    properties:
      description:
        type: string
        description: >
          Description of the analyte, e.g., "RdRp gene target for SARS-CoV-2 RNA as
          described in 10.2807/1560-7917.ES.2020.25.3.2000045".
      limit_of_detection:
        description: Lowest value at which the analyte can be detected.
```

- JSON schema (json-schema.org) is a standard to specify the structure of data, validate it, and include documentation about each field
- All data uploaded will be validated against the schema automatically

Data Validation



Create shrestha2020distribution.yaml #35

YWAN446 merged 2 commits into main from [illegible] last week

WenfeiXiao commented last week

This study evaluated the transmission potential of COVID-19 by examining viral load over time. Over six weeks, 230 healthcare personnel underwent 528 tests at the Cleveland Clinic. Cycle threshold (Ct) values were obtained using RT-PCR targeting the N gene, and viral loads were calculated. Data were obtained from the combined dataset in the supplementary materials of Challenger et al. BMC Medicine (2022) 20:25 (https://doi.org/10.1186/s12916-021-02220-0).

There is a discrepancy between the paper suggesting limit_of_detection as 20 gc/mL and Challenger combined dataset suggesting limit_of_detection as 67.24 gc/mL .

Create shrestha2020distribution.yaml

WenfeiXiao added marker: SARS-CoV-2 specimen: nasopharyngeal swab labels last week

WenfeiXiao requested review from tillahoffmann and YWAN446 last week

github-actions bot commented last week • edited

Summary for changed file data/shrestha2020distribution/shrestha2020distribution.yaml :

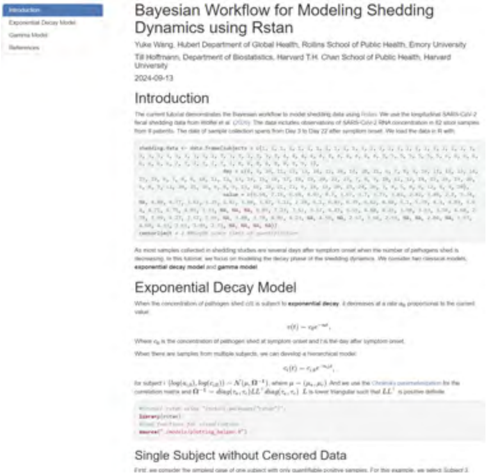
```
               n_samples  n_unique_participants  n_negative  n_positive  n_quantified        min          median
swab_SARSCoV2_N    528.0                  230.0       119.0         0.0         409.0  416.739004  180010.679416  ...
```

Successfully merging this pull request may close these issues.

To generate **reliable** data:

- Data will be checked and reviewed by at least two reviewers
- Functions were created to generate automatic data summary
- Conversations and decisions will be documented on GitHub

Modeling Tutorial and Code



Bayesian Workflow for Modeling Shedding Dynamics using Rstan

Yuke Wang, Hubert Department of Global Health, Rollins School of Public Health, Emory University

Till Hoffmann, Department of Biostatistics, Harvard T.H. Chan School of Public Health, Harvard University

2024-09-13

Introduction

The current tutorial demonstrates the Bayesian workflow to model shedding data using Rstan. We use the longitudinal SARS-CoV-2 fecal shedding data from Wölfel et al. (2020). The data includes observations of SARS-CoV-2 RNA concentration in 82 stool samples from 8 patients. The date of sample collection spans from Day 3 to Day 22 after symptom onset. We load the data in R with:

```
shedding.data <- data.frame(subjects = c(1, 1, 1, 1, 1, [illegible]),
                            day = c(6, 9, 10, 11, 13, [illegible]),
                            value = c([illegible] NA, NA, NA, NA))
[illegible]
```

As most samples collected in shedding studies are several days after symptom onset when the number of pathogens shed is decreasing, in this tutorial, we focus on modelling the decay phase of the shedding dynamics. We consider two classical models, **exponential decay model** and **gamma model**.

Exponential Decay Model

When the concentration of pathogen shed c(t) is subject to **exponential decay**, it decreases at a rate a_0 proportional to the current value:

$$c(t) = c_0 e^{-a_0 t},$$

Where c_0 is the concentration of pathogen shed at symptom onset and t is the day after symptom onset.

When there are samples from multiple subjects, we can develop a hierarchical model:

$$c_i(t) = c_{i,0} e^{-a_{i,0} t},$$

for subject i. $(\log(a_{i,0}), \log(c_{i,0})) \sim N(\mu, \Omega^{-1})$, where $\mu = (\mu_a, \mu_c)$. And we use the Cholesky parameterization for the correlation matrix and $\Omega^{-1} = diag(\tau_a, \tau_c) LL^T diag(\tau_a, \tau_c)$. L is lower triangular such that LL^T is positive definite.

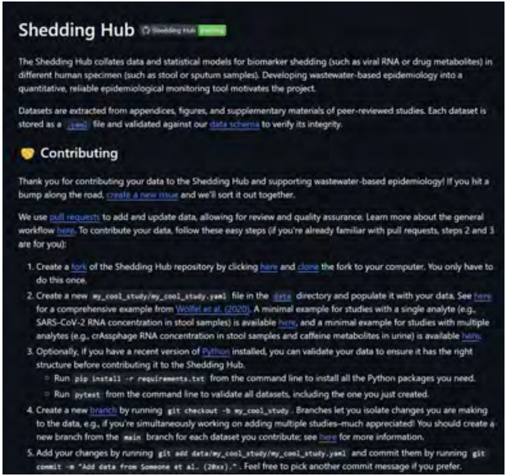
```
#install rstan using "install.packages("rstan")"
library(rstan)
#load functions for visualization
source("./models/plotting_helper.R")
```

Single Subject without Censored Data

First, we consider the simplest case of one subject with only quantifiable positive samples. For this example, we select Subject 3

- Tutorial for modeling shedding data in Python/R
- **Reproducible** using markdown files

Contribution



Shedding Hub

The Shedding Hub collates data and statistical models for biomarker shedding (such as viral RNA or drug metabolites) in different human specimen (such as stool or sputum samples). Developing wastewater-based epidemiology into a quantitative, reliable epidemiological monitoring tool motivates the project.

Datasets are extracted from appendices, figures, and supplementary materials of peer-reviewed studies. Each dataset is stored as a `.yaml` file and validated against our data schema to verify its integrity.

Contributing

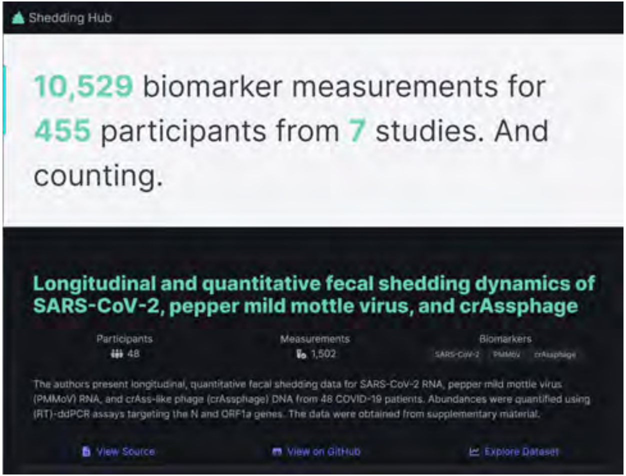
Thank you for contributing your data to the Shedding Hub and supporting wastewater-based epidemiology! If you hit a bump along the road, create a new issue and we'll sort it out together.

We use pull requests to add and update data, allowing for review and quality assurance. Learn more about the general workflow here. To contribute your data, follow these easy steps (if you're already familiar with pull requests, steps 2 and 3 are for you):

1. Create a fork of the Shedding Hub repository by clicking here and clone the fork to your computer. You only have to do this once.
2. Create a new `my_cool_study/my_cool_study.yaml` file in the `data` directory and populate it with your data. See here for a comprehensive example from Wölfel et al. (2020). A minimal example for studies with a single analyte (e.g., SARS-CoV-2 RNA concentration in stool samples) is available here, and a minimal example for studies with multiple analytes (e.g., crAssphage RNA concentration in stool samples and caffeine metabolites in urine) is available here.
3. Optionally, if you have a recent version of Python installed, you can validate your data to ensure it has the right structure before contributing it to the Shedding Hub.
   - Run `pip install -r requirements.txt` from the command line to install all the Python packages you need.
   - Run `pytest` from the command line to validate all datasets, including the one you just created.
4. Create a new branch by running `git checkout -b my_cool_study`. Branches let you isolate changes you are making to the data, e.g., if you're simultaneously working on adding multiple studies–much appreciated! You should create a new branch from the `main` branch for each dataset you contribute; see here for more information.
5. Add your changes by running `git add data/my_cool_study/my_cool_study.yaml` and commit them by running `git commit -m "Add data from Someone et al. (20xx)."`. Feel free to pick another commit message if you prefer.

- **Sustainable** with partnerships and contributions from the research community

Shedding Hub Website



Shedding Hub

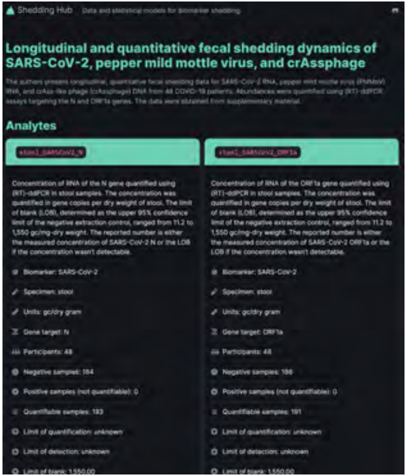
10,529 biomarker measurements for 455 participants from 7 studies. And counting.

Longitudinal and quantitative fecal shedding dynamics of SARS-CoV-2, pepper mild mottle virus, and crAssphage

Participants	Measurements	Biomarkers
48	1,502	SARS-CoV-2 PMMoV crAssphage

The authors present longitudinal, quantitative fecal shedding data for SARS-CoV-2 RNA, pepper mild mottle virus (PMMoV) RNA, and crAss-like phage (crAssphage) DNA from 48 COVID-19 patients. Abundances were quantified using (RT)-ddPCR assays targeting the N and ORF1a genes. The data were obtained from supplementary material.

shedding-hub.github.io



Shedding Hub Data and statistical models for biomarker shedding

Longitudinal and quantitative fecal shedding dynamics of SARS-CoV-2, pepper mild mottle virus, and crAssphage

The authors present longitudinal, quantitative fecal shedding data for SARS-CoV-2 RNA, pepper mild mottle virus (PMMoV) RNA, and crAss-like phage (crAssphage) DNA from 48 COVID-19 patients. Abundances were quantified using (RT)-ddPCR assays targeting the N and ORF1a genes. The data were obtained from supplementary material.

Analytes

stool_SARSCoV2_N

Concentration of RNA of the N gene quantified using (RT)-ddPCR in stool samples. The concentration was quantified in gene copies per dry weight of stool. The limit of blank (LOB), determined as the upper 95% confidence limit of the negative extraction control, ranged from 11.2 to 1,550 gc/mg-dry weight. The reported number is either the measured concentration of SARS-CoV-2 N or the LOB if the concentration wasn't detectable.

- Biomarker: SARS-CoV-2
- Specimen: stool
- Units: gc/dry gram
- Gene target: N
- Participants: 48
- Negative sampled: 184
- Positive samples (not quantifiable): 0
- Quantifiable samples: 193
- Limit of quantification: unknown
- Limit of detection: unknown
- Limit of blank: 1,550.00

stool_SARSCoV2_ORF1a

Concentration of RNA of the ORF1a gene quantified using (RT)-ddPCR in stool samples. The concentration was quantified in gene copies per dry weight of stool. The limit of blank (LOB), determined as the upper 95% confidence limit of the negative extraction control, ranged from 11.2 to 1,550 gc/mg-dry weight. The reported number is either the measured concentration of SARS-CoV-2 ORF1a or the LOB if the concentration wasn't detectable.

- Biomarker: SARS-CoV-2
- Specimen: stool
- Units: gc/dry gram
- Gene target: ORF1a
- Participants: 48
- Negative samples: 186
- Positive samples (not quantifiable): 0
- Quantifiable samples: 191
- Limit of quantification: unknown
- Limit of detection: unknown
- Limit of blank: 1,550.00

Work Plan

Phase 1	*Phase 2*	*Phase 3*
Objectives: Build a Shedding Hub Team; Develop the Basic Structure and Workflow	**Objectives**: Expand to Additional Pathogens and Biomarkers; Invite Contribution of Data and Models	**Objectives**: Develop analytical tools (e.g., dashboard, packages) and promote usage and contribution
Data Sources: Open Access Data, Published in Literature **Priority**: Quantifiable Longitudinal Measurements	**Data Sources**: Limited Access Data **Priority**: Quantifiable Longitudinal Measurements	**Priority**: Semi-quantifiable (CT values) or Non-quantifiable (Presence/Absence) Measurements
Models: Developed within the Shedding Hub Team	**Models**: Published Models by Experts	**Models**: Contributed Models by Research Community
Prioritized Biomakers: Pathogens of Interest for WWS	**Prioritized Biomakers**: Additional Pathogens and Biomarkers	**Prioritized Biomakers**: Additional Pathogens and Biomarkers
May 2024 – Sep 2025	*Oct 2025 – Sep 2026*	*Oct 2026 – Sep 2027*

Use Cases

- To support wastewater-based epidemiology applications estimating disease incidence for various infectious diseases
- To better understand of sensitivity of different disease diagnostic methods (nasopharyngeal swab vs. rectal swab)
- To support decision making for disease control and prevention policys (e.g., how long the quanrantine period should be?)
- To support wastewater monitoring for drug use

Acknowledgements

Shedding Hub Team

Yuke Wang

Till Hoffmann

Weifei Xiao

Youwei Hu

Zirui Chen

Haisu Zhang

CIDMATH Team