Documentation

Wastewater Surveillance AMELAG

Robert Koch Institute, & **Federal Environment Agency**

Contributors
Unit 32[1]

[1] Robert Koch Institute

Cite
Robert Koch Institute, & Federal Environment Agency. (2025). Wastewater Surveillance AMELAG [Data set]. Zenodo. https://doi.org/10.5281/zenodo.17590804

Abstract
The dataset "Wastewater Surveillance AMELAG" of the Robert Koch Institute and the Federal Environment Agency provides data from the monitoring of infectious pathogens in wastewater. Data on SARS-CoV-2 viral load has been collected since February 2022 in a nationwide network of wastewater treatment plants, laboratories and local authorities. Since then, the data has been supplemented by the viral load of other respiratory viruses (Influenza A/B, RSV). In addition to individual values from the wastewater treatment plants, the data set also contains population-weighted, aggregated time series. In addition, analysis scripts are provided as context materials.

Table of Content

- Information on the dataset and context of origin
- Content and structure of the dataset
- Guidelines for reuse of the data

--- die deutsche Version finden Sie hier ---

Information on the dataset and context of origin

In AMELAG (“Abwassermonitoring für die epidemiologische Lagebewertung”, German for wastewater monitoring for epidemiological situation assessment), running from 22.11.2022 to 31.12.2025, local authorities, wastewater treatment plants (WWTP) and laboratories are working together to take, analyze and evaluate wastewater samples. The project aims at testing wastewater samples for selected pathogens and to establish it as an additional indicator for the epidemiological situation assessment at state and federal level. Further aims of the project include further development of structures and processes for a nationwide wastewater surveillance network, to develop concepts for continuity and to research the possibilities for monitoring other pathogens in wastewater. Currently, wastewater samples from selected treatment plants are being tested for SARS-CoV-2, influenza viruses and eespiratory syncytial viruses (RSV).
Wastewater surveillance is a technique for detecting pathogens in wastewater to better control health protection measures. Wastewater surveillance has a range of applications. Wastewater data, however, underlie several limiations. For example, they do not allow for an accurate assessment of disease severity or the burden on the healthcare system. In epidemiological assessments, the data should be combined with other indicators, e.g. from syndromic surveillance.

Administrative and organizational information

AMELAG is a project funded by the Federal Ministry of Health (BMG) and is being conducted in cooperation with the Federal Ministry for the Environment, Nature Conservation, Nuclear Safety and Consumer Protection (BMUV).
The project is being carried out jointly by the Robert Koch Institute (RKI) and the Federal Environment Agency (UBA). Further information on AMELAG can be found on the project website.
The participating WWTPs are responsible for taking samples, which are analyzed by the participating laboratories. In addition to commercial laboratories, state laboratories and the Federal Environment Agency, the Central Medical Service of the German Armed Forces also carries out part of the analysis.

Some of the WWTPs and laboratories are also involved in wastewater surveillance projects in the federal states (Baden-Württemberg, Bavaria, Berlin, Brandenburg, Hamburg, Hesse, Rhineland-Palatinate, Saxony-Anhalt).
Other WWTPs and laboratories are part of the following research projects:

- WBEready - A research consortium consisting of Emschergenossenschaft and LippeverbandEGLV, Research Institute for Water Management and Climate Future at RWTH Aachen FiW, University Hospital Frankfurt, Goethe University Frankfurt am Main, University Medicine Essen (Institute for Artificial Intelligence, Institute for Urban Public Health), RWTH Aachen, Institute for Urban Water Management.
- Establishment of methods for the detection of viruses in wastewater to assess the infection situation in the population (University of Dresden)
- Development of a state-wide wastewater surveillance system in Thuringia using mobility data and artificial intelligence (research consortium of the University of Weimar, University of Jena, University of Hamburg, Hamm-Lippstadt University of Applied Sciences, SMA Development GmbH, KOWUG Kommunale Wasser- und Umwelttechnik GmbH, Analytik Jena GmbH)
- Establishment of a multiplex PCR from wastewater and for detection and characterization of RSV in the context of SARS-CoV-2 wastewater monitoring (AMELAG) (University of Bonn and Düsseldorf).

The company ENDA was commissioned with data management. The data collected are stored and processed in a database (PiA-Monitor).

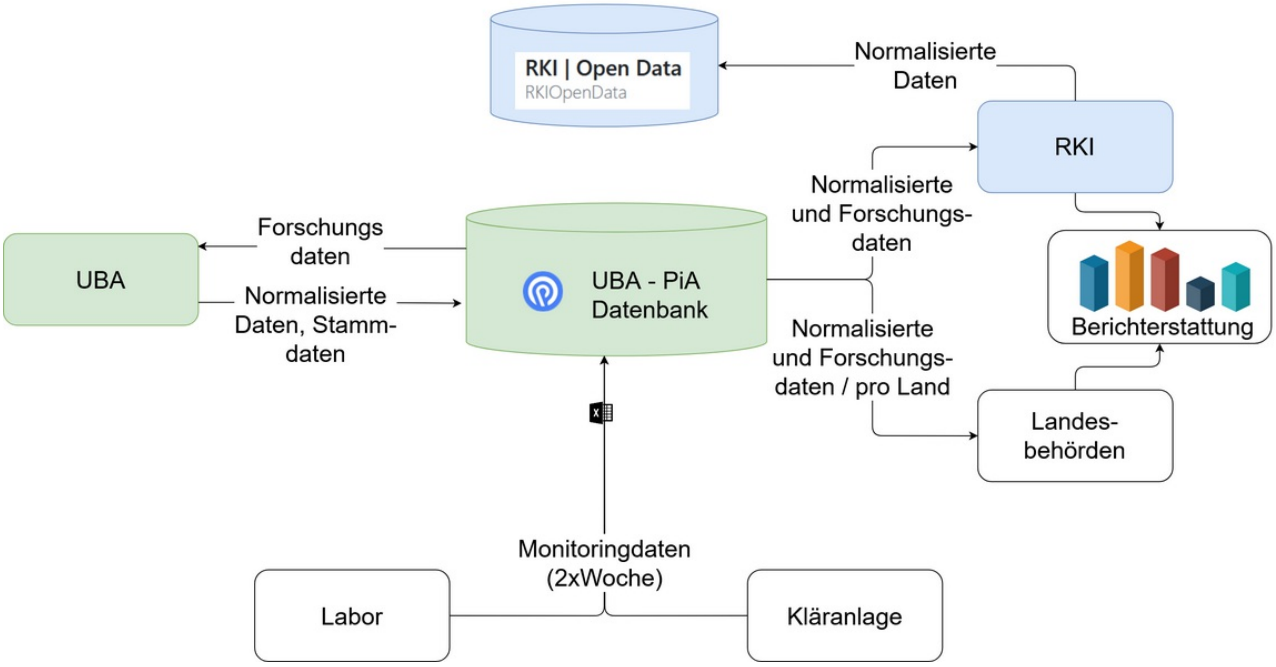
The data are processed, edited and published by the Department MF 4 | Subject and Research Data Management. Questions about data management and the publication infrastructure can be directed to the Open Data team of the MF4 department at OpenData@rki.de.

Data collection

In AMELAG, technical guidelines were developed based on the handouts for sampling and laboratory analysis created as part of the ESI-CorA project. The raw data of the SARS-CoV-2 samples analyzed in the ESI-CorA project are reused in AMELAG and included in the evaluated data.
Raw wastewater samples are generally collected twice a week at each participating WWTP, along with essential parameters such as volume flow, pH value, and temperature. These parameters are necessary for normalization and quality assurance. Where possible, the raw sewage samples should be taken after the grit chamber of the WWTP. A 24-hour composite sample is collected using an automatic sampler. The 24-hour samples are usually taken from Mondays to Tuesdays, and from Wednesdays to Thursdays. As a rule, one liter of the sample is filled into sample bottles and sent to the analysis laboratory.
In the laboratory, the viral nucleic acid is concentrated, extracted and the viral gene sequences are quantified by digital PCR (dPCR) or quantitative real-time PCR (qRT-PCR). For SARS-CoV-2, at least two representative gene fragments (preferably N1, N2, E, ORF or RdRp) are determined, for the Influenza virus only one gene fragment (M1 for Influenza A Virus and M1, NS1, NS2 or HA for Influenza B Virus), for RSV also only one gene fragment (N for RSV A and RSV B, M or N for RSV A/B).

> Robert Koch Institute, Department 32 (2024): "ESI-CorA: SARS-CoV-2 wastewater surveillance" [Dataset]. Zenodo. DOI: 10.5281/zenodo.10781653

Data flow



At the UBA, metadata on the WWTPs and the laboratories as well as the regularly collected monitoring data are centrally stored and processed further in a web application, the PiA-Monitor (Pathogens in Wastewater). The monitoring data to be collected regularly from the WWTP and the data of the laboratories are merged and imported into the database by the data providers via the web application. The UBA, the RKI and the federal states can access the data within the scope of their respective rights.

Plausibility check and further processing of the data

A plausibility check is run on the data as they are imported. The formats, completeness of the information (mandatory fields), value ranges of the monitoring data, plausibility of the dates and compliance with stored metadata are checked. Only data records that successfully pass the quality check are imported into the database. For SARS-CoV-2, the geometric mean of the viral load (gene copies/L) is then determined from the two or more measured target genes.

Normalization procedure

A varying wastewater composition, e.g. due to irregular industrial influences or heavy rainfall events, can lead to changing concentrations of the pathogens. To take these external influences into account, the measured viral load can be normalized.
In AMELAG, normalization of the SARS-CoV-2 data is performed according to flow rate. The dry weather inflow of the WWTP is the reference. The following formula was used:

$$Gene_{normalized} = Q_{KA\_current}/Q_{KA\_median} \cdot Gene_{averaged}$$

where:

- $Q_{KA\_aktuell}$: Volume flow of the wastewater treatment plant in the sampling period and
- Q_{KA_median} : Median of the volume flow of the wastewater treatment plant

Normalization is automated with the data import. The measured Influenza and RSV data are currently not normalized as the normalization does not show an imporved data quality.

Data evaluation

The data are evaluated at the RKI using R scripts. The scripts are contained in the context materials. A detailed description of the methodology is provided in the technical guidelines. The results are published in the RKI´s weekly report.
For each WWTP, the measured values for SARS-CoV-2 (normalized), Influenza A and B viruses (not normalized) and RSV A, RSV B and RSV A/B are reported in gene copies per liter (gene copies/L). In addition, the measured values of the logarithmized normalized gene copies are smoothed using a generalized additive model (GAM) and associated confidence intervals are calculated.

Aggregation of the WWTP values

The individual time series of the WWTPs are aggregated in order to depict a nationwide course for each pathogen. To do this, the mean value is first calculated over the logarithmized measured values of the individual WWTPs averaged over one week. Then, for each location and for each week, the deviation from the weekly mean value over all locations is calculated. For each WWTP-laboratory combination, the mean of these differences over all weeks is determined and then subtracted from the originally measured values. This procedure adjusts for mean differences in viral loads between different WWTP-laboratory combinations. Finally, the weighted mean (weighted according to the number of inhabitants connected to the respective WWTP) of these adjusted values is calculated for each week in which measured values are available for at least 10 locations.
As different WWTPs and laboratories are involved in the data collection, inconsistencies in the data from individual sites may arise, potentially having a significant impact on the values aggregated across all sites. As soon as such inconsistencies are detected, these values are excluded from the aggregated curves (`amelag_aggregierte_kurve.tsv`) until the underlying causes have been fully clarified. The values remain in the data for the individual sites (`amelag_aggregierte_kurve.tsv`).

Notes on data evaluation

Some things to take into account when evaluating the data:

- Different target genes were measured at the different sites
  - SARS-CoV-2: a combination of preferably N1, N2, E, ORF oder RdRp
  - Influenza A-Virus: M1
  - Influenza B-Virus: M1, NS1, NS2, HA
  - RSV A: N
  - RSV B: N
  - RSV A/B: M, N
- Some cities are have more than one sewage treatment plant or more than one inflow.
- For values below the limit of quantification (LOQ), half of the LOQ is used as the value (0.5 * LOQ).

Limitations

Wastewater data do not allow conclusions to be drawn about disease severity or the burden on the healthcare system. At present, it is not possible to draw precise conclusions about incidence/prevalence or underreporting from wastewater data. When assessing a sitaution epidemiologically, the data should always be considered in combination with other indicators, such as those from syndromic surveillance. Absolute viral loads cannot be compared directly to the number of infected persons, especially over longer periods of time, as, for example, the amount of virus excreted per infected person can differ between different virus variants.
The values determined are influenced by a variety of factors (e.g. changes in the wastewater supply, heavy rainfall events, or tourist events), which can only be partially compensated for by normalization. The time delay from sampling to transmission and further publication by the RKI can take up to two weeks.

Content and structure of the dataset

The AMELAG dataset provides data and contextual material on SARS-CoV-2 detections in wastewater. The data collected in the project are available for individual sites and as aggregated time series.

The dataset also contains:

- License file with the license to use the dataset in German and English
- Dataset documentation in German
- Metadata for automated further processing
- Context materials for data analysis

Data for individual WWTP

The file `amelag_einzelstandorte.tsv` contains the normalized SARS-CoV-2, non-normalized influenza viral load data and non-normalized RSV data for the individual sites.

> amelag_einzelstandorte.tsv

Variables and variable values

The file amelag_einzelstandorte.tsv contains the variables and their values shown in the following table. A machine-readable data schema is stored in Data Package Format in tableschema_amelag_einzelstandorte.en.json:

> tableschema_amelag_einzelstandorte.en.json

Variable	Type	Characteristic	Description
standort	string	Examples: `Aachen`, `Ratzeburg`, `Weil am Rhein`	Location of the wastewater treatment.
bundesland	string	Values: `BB`, `BE`, `BW`, `BY`, `HB`, `HE`, `HH`, …	Federal state (abbreviated) in which the wastewater treatment plant is located.
datum	date	Format: `YYYY-MM-DD` Missing values: `NA`	Date on which the 24-hour composite sample started in the wastewater treatment plant.
viruslast	number	Values: `≥0` Missing values: `NA`	Measured viral load in gene copies per liter.
viruslast_normalisiert	number	Values: `≥0` Missing values: `NA`	Flow-normalized viral load (as described in the variable "viruslast").
vorhersage	number	Values: `≥0` Missing values: `NA`	Predicted viral load (predicted using a GAM regression with adaptive smoothing and the non-normalized viral loads, optimized by a cross-validation for the log10-transformed values, transformed back to the original scale).
obere_schranke	number	Values: `≥0` Missing values: `NA`	Upper bound of the pointwise 95% confidence interval of the GAM predicted value.
untere_schranke	number	Values: `≥0` Missing values: `NA`	Lower bound of the pointwise 95% confidence interval of the GAM predicted value.
einwohner	integer	Values: `≥0` Missing values: `NA`	Number of inhabitants connected to the site's sewage treatment plant.
laborwechsel	string	Values: `ja`, `nein` Missing values: `NA`	Indicates whether change in laboratory or change in laboratory mathods occurred.
typ	string	Values: `SARS-CoV-2`, `Influenza A`, `Influenza B`, `Influenza A+B`, `RSV A`, `RSV B`, `RSV A+B`, …	Virus type.
unter_bg	string	Values: `ja`, `nein` Missing values: `NA`	Indicates if at least half of the measured genes are under the limit of quantification (`ja` =yes, `nein` =no).

Data aggregated across all WWTP

In the file `amelag_aggregated_curve.tsv` contains the time series of the SARS-CoV-2, influenza virus and RSV loads on an aggregated or nationwide level.

> amelag_aggregierte_kurve.tsv

Variables and variable characteristics

The file amelag_aggregierte_kurve.tsv contains the variables and their values shown in the following table. A machine-readable data schema is stored in Data Package Format in tableschema_amelag_aggregierte_kurve.en.json:

> tableschema_amelag_aggregierte_kurve.en.json

Variable	Type	Characteristic	Description

datum	date	Format: `YYYY-MM-DD`	Date of Wednesdays of a week. The data of the underlying individual time series are averaged within the period from the previous Thursday to the specified Wednesday.
n	integer	Values: `≥0` Missing values: `NA`	Number of locations that have transmitted at least one measured value in the period defined by "date".
anteil_bev	number	Values: `≥0` Missing values: `NA`	Proportion of the total population in Germany that is connected to the transmitting sewage treatment plants.
viruslast	number	Values: `≥0` Missing values: `NA`	Measured viral load (in gene copies per liter averaged over all sites and weighted by connected inhabitants of the wastewater treatment plants). Before averaging across the sites, all measured values of the sites in the last 7 days were transformed using the logarithm of 10 and averaged across the individual sites. The indicated viral load is the mean value transformed back to the original scale.
viruslast_normalisiert	number	Values: `≥0` Missing values: `NA`	Flow-normalized viral load (as described in the variable "viruslast").
vorhersage	number	Values: `≥0` Missing values: `NA`	Predicted viral load (predicted using a GAM regression and the non-normalized viral loads, optimized by a cross-validation for the log-transformed values, transformed back to the original scale).
obere_schranke	number	Values: `≥0`	Upper bound of the 95% confidence interval of the GAM predicted value.
untere_schranke	number	Values: `≥0`	Lower bound of the 95% confidence interval of the GAM predicted value.
typ	string	Values: `SARS-CoV-2`, `Influenza A`, `Influenza B`, `Influenza A+B`, `RSV A`, `RSV B`, `RSV A+B`, ...	Virus type.

Context materials

To reproduce the results of the AMELAG weekly report, the R scripts used to create the analysis are provided. The scripts can be found in the "Contextual materials" folder of the dataset.

> Context Materials

Metadata

To increase findability, the provided data are described with metadata. The Metadata are distributed to the relevant platforms via GitHub Actions. There is a specific metadata file for each platform; these are stored in the metadata folder:

> Metadaten/

Versioning and DOI assignment are performed via Zenodo.org. The metadata prepared for import into Zenodo are stored in the zenodo.json. Documentation of the individual metadata variables can be found at https://developers.zenodo.org/representation.

> Metadaten/zenodo.json

The zenodo.json includes the publication date and the date of the data status in the following format (example):


```
"publication_date": "2024-06-19",
"dates": [
  {
    "start": "2023-09-11T15:00:21+02:00",
    "end": "2023-09-11T15:00:21+02:00",
    "type": "Created",
    "description": "Date when the published data was created"
  }
],
```


Additionally, we describe tabular data using the Data Package Standard.
A Data Package is a structured collection of data and associated metadata that facilitates data exchange and reuse. It consists of a `datapackage.json` file that contains key information such as the included resources, their formats, and schema definitions.

The Data Package Standard is provided by the Open Knowledge Foundation and is an open format that enables a simple, machine-readable description of datasets.

The list of data included in this repository can be found in the following file:

> datapackage.json

For tabular data, we additionally define a Table Schema that describes the structure of the tables, including column names, data types, and validation rules. These schema files can be found in:

> Metadaten/schemas/

Guidelines for reuse of the data

Open data from the RKI are available on Zenodo.org, GitHub.com, OpenCoDE, and Edoc.rki.de:

- https://zenodo.org/communities/robertkochinstitut
- https://github.com/robert-koch-institut
- https://gitlab.opencode.de/robert-koch-institut
- https://edoc.rki.de/

License