Data set documentation
SARS-CoV-2 infections in Germany

Robert Koch Institute | RKI
Nordufer 20
13353 Berlin

FG 32 | Surveillance and electronic reporting and information system (DEMIS) | ÖGD contact point
Michaela Diercke (Management)

MFI | Method Development, Research Infrastructure and Information Technology
Linus Grabenhenrich (Management)

IT 4 | Software Architecture and Development
Herrmann Claus (Head)

MF 4 | Information and Research Data Management
Hannes Wuensche (Data Curation)

Cite
Robert Koch Institute (2023): SARS-CoV-2 infections in Germany, Berlin: Zenodo. DOI:10.5281/zenodo.4681153.

Information on the data set and context of origin

This data set contains comprehensive information on SARS-CoV-2 infections in Germany, which were reported to the Robert Koch Institute (RKI) by the health authorities in accordance with the Infection Protection Act (IfSG). The data includes information on the number of confirmed cases, deaths and recoveries, from which further key figures related to the COVID-19 pandemic can be derived. The dataset is updated daily and contains detailed information at county level, broken down by different age groups. The provision of the dataset is intended to help improve understanding of the COVID-19 pandemic in Germany and support reporting, research and analysis in this area.

Administrative and organizational information

The "SARS-CoV-2 infections in Germany" dataset provides the daily case numbers of positive SARS-CoV-2 infections, deaths and recoveries reported by the health authorities in Germany in accordance with the requirements of the Infection Protection Act (IfSG).

The underlying data is transmitted to the Robert Koch Institute (RKI) via the reporting system in accordance with the IfSG. Responsible for the operation of the

Reporting System is the RKI's Division 32 | Surveillance and Electronic Reporting and Information System (DEMIS) | ÖGD Contact Point.
The processing and preparation of the raw data available in the reporting system is carried out by the RKI's IT 4 | Software Architecture and Development department.
The publication of the data, data curation and quality management of the (meta-)data are carried out by the MF 4 | Information and Research Data Management department. Questions about data management and the publication infrastructure can be directed to the Open Data Team of the MF4 department at OpenData@rki.de.

Content and structure of the data set

The dataset contains epidemiological data on the course of SARS-CoV-2 infections in Germany. The dataset contains:

- Case number data with daily updated reports of SARS-CoV-2 infections
- Archive with the collection of all previous case number tables
- License file with the license to use the data set
- Data set documentation in German
- Metadata file for import into Zenodo

Data and data preparation

The case number data represent a daily updated status (00:00) of all previously reported cases of infection in Germany. This means that all SARS-CoV-2 infections transmitted to the RKI's reporting system by the health authorities via the responsible state authorities by 00:00 a.m. of the day YYYY-MM-DD are included in the data status. The data is generated completely new every day and this data status replaces the data status of the previous day.

The case number data contains the district ID as the only geoinformation. This is based on the official municipality key (AGS) of quarter 2 2020, retrieved from the portal of the Federal Statistical Office. The district ID is derived from the code of the federal state (Land), the regional district (RB) and the administrative district (LK). For a more precise representation of Berlin, the 12 city districts are broken down as separate "Landkreise". This deviates from the specifications of the AGS. The following allocation is made:

IdCounty	District	IdCounty	District
11001	Berlin Mitte	11007	Berlin Tempelhof-Schöneberg
11002	Berlin Friedrichshain-Kreuzberg	11008	Berlin Neukölln
11003	Berlin Pankow	11009	Berlin Treptow-Köpenick
11004	Berlin Charlottenburg-Wilmersdorf	11010	Berlin Marzahn-Hellersdorf
11005	Berlin Spandau	11011	Berlin Lichtenberg
11006	Berlin Steglitz-Zehlendorf	11012	Berlin Reinick-endorf

Case number data

Archive/YYYY-MM-DD_Germany_SARSCoV2_Infections.csv

The central date of the data set is the current case number data. The archive folder contains the case number data under the file name "YYYY-MM-DD_Germany_SARSCoV2_Infections.csv". In the file name, the sequence "YYYY-MM-DD" represents the creation date of the file and thus also the date of the data status it contains. "YYYY" stands for the year, "MM" for the month and "DD" for the day of creation or the data status contained.

Characteristics of the case number data In the .csv case number table, the columns differentiate the various characteristics of a case group. One unique case group is shown per row. A case group does not include individual cases. However, it is possible that only one case is contained in the case group. A case group is basically characterized by the following properties (the characteristics of these properties are shown in brackets):

- Location of the infections (IdLandkreis)

- Group of people (gender, age group)

- Time of notification of the infection (notification date)

- Start of illness (ref date, actual start of illness)

- Power of the group (number of cases, number of deaths, number of recoveries)
- Notification status (NewCase, NewDeath, NewRecovery)

A case group assumes a unique characteristic with regard to its number of cases ("NumberCase"), "AgeGroup", "Gender", its district ("Idistrict"), "NotificationDate", date of illness ("Refdate") and the information as to whether the date of illness is known ("ActualDiseaseStart"). Furthermore, the "number of deaths" or "number of recoveries" of each case group is specified, whereby only one of the two characteristics "number of deaths" or "number of recoveries" can be assumed. This means that if there are deaths or recoveries in a case group, the number of deaths or the number of recovered cases is specified in a new group. If, for example
B. both cases are in one case group, the case group is divided into two further groups, namely a group of deaths and a group of recoveries.

Example

A new case group w is registered (IdCounty, gender, age group, registration date, ref date, actual start of illness are constant). This contains a case group at the beginning:

Case group w: 5 infected, 0 deaths and 0 recovered cases

If 1 of the cases die and 2 recover, case group w splits into 3 groups: Case group

x: 2 infected, 0 deaths and 0 recovered cases
Case group y: 1 infected person, 1 death and 0 recovered cases Case group z: 2 infected persons, 0 deaths and 2 recovered cases

---

The characteristics of the reporting status indicate whether, in relation to the previous day, changes have occurred in the cases of infection, deaths and recoveries in a case group. This makes it possible to track the changes compared to the previous day. These result from new notifications of infections (including late notifications), corrections (e.g. due to erroneous notifications, but also corrections regarding district, age, gender or onset of illness) and changes in the state of health (recovered, deceased). The characteristics of the reporting status temporarily split case groups. The split is temporary, as it only shows the changes from the day of publication to the previous day. New cases form a separate case group for the day of the new notification. As a case is only newly reported, newly recovered or newly deceased or corrected on one day, the temporary splitting of the case group on the day of the new notification of the notification status is followed by a merging of the groups on the following day. A more detailed explanation of this process is provided in the following section.

Characteristic values The case number data contains the characteristics and their values shown in the following table:

Feature	Characteristic	Explanation
IdCounty	1001 to 16077	Identification number of the district based on the Official Municipality key (AGS) plus the 12 Districts of Berlin (11001 to 11012); Territorial status: 30.06.2020 (2nd quarter)
Sex	W, M, unknown	Gender of the case group: female (W), male (M) and (unknown)
Age groupA00-A04	, A05-A14, A15-A34, A35-A59, A60-A79, A80+, unknown	Age range of cases included in the group, stratified by 0-4 years, 5-14 years, 15-34 years, 35-59 years, 60-79 years, 80+ years and unknown
Reporting date	YYYY-MM-DD	Date when the case has become known to the public health department. YYYY corresponds to the year, MM to the month and DD to the day.
Refdate	YYYY-MM-DD	Date of onset of illness. If this is not known, the date of notification.
Actual onset of illness	0, 1	1: Refdate is the Start of illness 0: Refdate is the date of notification

Feature	Characteristic	Explanation
NumberCase	Whole number	Number of reported Cases in the corresponding Case group For NewFall = -1, the Number negative: It is a Correction of the Case group indicating, how many infections have been reported a lot
Numberofdeaths	Whole number	Number of reported Deaths in the corresponding Case group For NewDeath = -1, is the number is negative: It is a Correction of the Case group indicating, how many deaths have been reported a lot
NumberGeneses	Whole number	Number of recovered Cases in the corresponding Case group For NeuGenesen = -1, is the number is negative: It is a Correction of the Case group indicating, How many recovered cases too much has been reported are

Characteristic	Characteristic	Explanation
NewCase, NewDeath, NewGenesis	0, 1, -1	0 : Cases in the group are contained in the publication for the current day and in the publication for the previous day. This means that these cases have been known for more than one day. 1 : Cases in the group are included in the current publication for the first time. This means that they are newly transmitted or newly assessed cases for the publication date. -1: Cases in the group are included in the publication of the previous day, but are removed from the Case number data removed. This means that cases are removed from the current day. Such a case group can arise, for example, from incorrect reports, which are displayed as a correction.

Characteristic	Characteristic	Explanation
NewDeath, NewGenese		-9Cases in the group are not reported as recovered ("New recovered") or deceased ("New death") in either the publication for the current day or the publication for the previous day. This means that no information is known about the course of the infection for the cases in the group. This is often the case, for example, when a case group has just been reported as infected.

The temporary splitting of the case groups by the characteristics of the reporting status is illustrated in the following example. Temporary groups are indicated by a '. New notifications become clear when looking at the values of the characteristics:

Example

If a new case group is registered on day TT (IdCounty, gender, age group, notification date, refdate, actual onset of illness are constant), it assumes the notification status NewCase = [1]. If no recoveries or deaths are known, but are reported in the case group, NewDeath and NewRecoveries = [-9]:

The cases in case group w' are new in the dataset of day TT (new case [1]), the cases in the group are not death or recovery cases (new death [-9], new recovery [-9]).

> Case group w':
> Infected [4], deaths [0] and recovered cases [0]
> NewCase [1], NewDeath [-9], NewGenesis [-9]

On the next day, DD+1, the cases from case group w' are no longer new. Their notification status therefore changes from [1] to [0]. The temporary case group w'

(NewCase [1]) becomes the continuous case group w (NewCase [0]):

> Case group w: Infected [4], deaths [0] and recovered cases [0] New case [0], new death [-9], newly recovered [-9]

On day TT+1, an additional, new case is registered in case group w. As this is a new case, it again forms a temporary, separate group w':

> Case group w':
> Infected [1], deaths [0] and recovered cases [0]
> NewCase [1], NewDeath [-9], NewGenesis [-9]

On the next day, DD+2, the cases in case group w'(DD+1) are no longer new, their notification status changes as it did the day before for case group w'(DD). By changing the reporting status to w'(TT+1), w' merges into w. The number of infected persons in both case groups is added together.

> Case group w: Infected [5], deaths [0] and recovered cases [0]
> New case [0], new death [-9], newly recovered [-9]

Reports of deaths or recoveries are similar to new infection reports. These form temporary case groups y' and z' which later become permanent case groups y and z:

Day TT+3
>Case group w:
>Infected [4], deaths [0] and recovered cases [0]
>NewCase [0], NewDeath [-9], NewRecovery [-9]

> Case group y':
> Infected [1], deaths [1] and recovered cases [0]
> NewCase [0], NewDeath [1], NewGenesis [-9]

Day TT+4
>Case group w:
>Infected [2], deaths [0] and recovered cases [0]
>NewCase [0], NewDeath [-9], NewRecovery [-9]

> Case group y:
> Infected [1], deaths [1] and recovered cases [0]
> NewCase [0], NewDeath [0], NewGenesis [-9]

> Case group z':
> Infected [2], deaths [0] and recovered cases [2]
> NewCase [0], NewDeath [-9], NewGenesis [1]

Note on convalescents

Based on the detailed information on a case of illness provided to the RKI by the health authorities, a duration of illness is calculated for each case.

is estimated. For cases in which only symptoms indicating a mild course of the disease are reported, the duration of the disease is assumed to be 14 days. For hospitalized cases or cases with symptoms that indicate a severe course (e.g. pneumonia), the duration of the illness is assumed to be 28 days. An estimated date of recovery is calculated for each case based on the onset of the illness or, if this is not known, on the date of notification. Since significantly longer courses of illness are also possible in individual cases, or the information used here is not transmitted to the RKI for all cases, the data calculated in this way are only rough estimates for the number of recovered persons and should therefore only be used taking these limitations into account.

Formatting the data The emergency recording surveillance data is contained in the data set as a comma-separated .csv file. The character set used in the .csv file is UTF-8. The individual values are separated by a comma ",". Dates are formatted in the ISO 8601 standard.

- Character set: UTF-8
- Date format: ISO 8601
- .csv separator: Comma ","
- compressed in `.xz` format

Metadata

To increase accessibility, the data provided is described with metadata. Metadata is distributed to the corresponding platforms via GitHub Actions. A specific metadata file exists for each platform; these are stored in the metadata folder:

Metadata/

Versioning and DOI assignment is done via Zenodo.org. The metadata provided for import into Zenodo is stored in zenodo.json. The documentation of the individual metadata variables can be found at https://developers.zenodo.org/#representation.

Metadata/zenodo.json

Notes on the subsequent use of the data

Open research data of the RKI are provided on GitHub.com, Zenodo.org and Edoc.rki.de:

- https://github.com/robert-koch-institut

- https://zenodo.org/communities/robertkochinstitut

- https://edoc.rki.de

License

The data provided in the data set is freely available, subject to the naming of the source. This means that any person has the right to process and modify the data, to create derivatives of the dataset and to use it for commercial and non-commercial purposes. Further information on the license can be found in the LICENSE file of the dataset.