Datensatzdokumentation

SARS-CoV-2 Infektionen in Deutschland

Robert Koch-Institut | RKI
Nordufer 20
13353 Berlin

FG 32 | Surveillance und elektronisches Melde- und Informationssystem (DEMIS) | ÖGD Kontaktstelle
Michaela Diercke (Leitung)

MFI | Methodenentwicklung, Forschungsinfrastruktur und Informationstechnologie
Linus Grabenhenrich (Leitung)

IT 4 | Softwarearchitektur und -entwicklung
Herrmann Claus (Leitung)

MF 4 | Informations- und Forschungsdatenmanagement
Hannes Wuensche (Datenkuration)

Zitieren
Robert Koch-Institut (2023): SARS-CoV-2 Infektionen in Deutschland, Berlin: Zenodo. DOI:10.5281/zenodo.8336615.

Informationen zum Datensatz und Entstehungskontext

Der vorliegende Datansatz enthält umfassende Informationen zu SARS-CoV-2-Infektionen in Deutschland, die gemäß dem Infektionsschutzgesetze (IfSG) von den Gesundheitsämtern an das Robert Koch-Institut (RKI) gemeldet wurden. Die Daten umfassen Informationen zur Anzahl der bestätigten Fälle, Todesfälle und Genesungen, aus denen sich weitere Kennzahlen im Zusammenhang mit der COVID-19-Pandemie ableiten lassen. Der Datensatz wird täglich aktualisiert und enthält detaillierte Informationen auf Landkreisebene, die nach verschiedenen Altersgruppen aufgeschlüsselt sind. Die Bereitstellung des Datensatzes soll dazu beitragen, das Verständnis der COVID-19-Pandemie in Deutschland zu verbessern und die Berichterstattung, Forschung und Analyse in diesem Bereich zu unterstützen.

Administrative und organisatorische Angaben

Im Datensatz "SARS-CoV-2 Infektionen in Deutschland" werden die tagesaktuellen Fallzahlen, der nach den Vorgaben des Infektionsschutzgesetzes - IfSG - von den Gesundheitsämtern in Deutschland gemeldeten positiven SARS-CoV-2 Infektionen, Todes- und Genesungsfälle

bereitgestellt.

Die zugrundeliegenden Daten werden an das Robert Koch-Institut (RKI) über das Meldesystem gemäß IfSG übermittelt. Zuständig für den Betrieb des Meldesystems ist das Fachgebiet 32 | Surveillance und elektronisches Melde- und Informationssystem (DEMIS) | ÖGD Kontaktstelle des RKI.
Die Verarbeitung und Aufbereitung der im Meldesystem vorliegenden Rohdaten erfolgt durch das IT 4 | Softwarearchitektur und -entwicklung des RKI.
Die Veröffentlichung der Daten, die Datenkuration sowie das Qualitätsmanagement der (Meta-)Daten erfolgen durch das Fachgebiet MF 4 | Informations- und Forschungsdatenmanagement.
Fragen zum Datenmanagement und zur Publikationsinfrastruktur können an das Open Data Team des Fachgebiets MF4 unter OpenData@rki.de gerichtet werden.

Inhalt und Aufbau des Datensatzes

Der Datensatz enthält epidemiologische Daten über den Verlauf der SARS-CoV-2 Infektionen in Deutschland. Im Datensatz enthalten sind:

- Fallzahlendaten mit tagesaktuellen Meldungen von SARS-CoV-2 Infektionen
- Archiv mit der Sammlung aller bisherigen Fallzahlentabellen
- Lizenz Datei mit der Nutzungslizenz des Datensatzes
- Datensatzdokumentation in deutscher Sprache
- Metadaten Datei zum Import in Zenodo

Daten und Datenaufbereitung

Die Fallzahlendaten bilden einen tagesaktuellen Stand (00:00 Uhr) aller bisherig gemeldeten Infektionsfälle in Deutschland ab. Das bedeutet, dass alle, bis 00:00 Uhr des Tages JJJJ-MM-TT, von den Gesundheitsämtern, über die zuständigen Landesbehörden, an das Meldesystem des RKIs übermittelten SARS-CoV-2 Infektionen im Datenstand enthalten sind. Die Daten werden täglich vollständig neu erzeugt und dieser Datenstand ersetzt den Datenstand des Vortages.

Die Fallzahlendaten enthalten als einzige Geoinformation die Landkreis ID. Diese richtet sich nach dem Amtlichen Gemeindeschlüssel (AGS) des Quartal 2 2020, abgerufen im Portal des Statistischen Bundesamtes. Die Landkreis ID ergibt sich aus der Kennzahl des Bundeslandes (Land), des Regionalbezirks (RB) und des Landkreises (LK).
Für eine genauere Darstellung Berlins, werden die 12 Stadtbezirke als eigene "Landkreise" aufgegliedert. Hier wird von den Vorgaben des AGS abgewichen. Folgende Zuordnung wird getroffen:

IdLandkreis	Bezirk	IdLandkreis	Bezirk
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 Reinickendorf

Fallzahlendaten

> Aktuell_Deutschland_SARSCoV2_Infektionen.csv

Zentrales Datum des Datensatzes sind die aktuellen Fallzahlendaten. Diese sind im Hauptverzeichnis unter "Aktuell_Deutschland_SARSCoV2_Infektionen.csv" abrufbar und werden täglich überschrieben.

Merkmale der Fallzahlen Daten

In der .csv Fallzahlentabelle differenzieren die Spalten die verschiedenen Merkmale einer Fallgruppe. Pro Zeile ist eine eineindeutige Fallgruppe abgebildet. Eine Fallgruppe umfasst keine Einzelfälle. Jedoch ist es möglich, dass in der Fallgruppe nur ein Fall enthalten ist. Eine Fallgruppe wird grundlegend durch folgende Eigenschaften charakterisiert (in den Klammern finden sich die Merkmale dieser Eigenschaften):

- Ort der Infektionen (IdLandkreis)
- Personengruppe (Geschlecht, Altersgruppe)
- Meldezeitpunkt der Infektion (Meldedatum)
- Erkrankungsbeginn (Refdatum, IstErkrankungsbeginn)
- Mächtigkeit der Gruppe (AnzahlFall, AnzahlTodesfall, AnzahlGenesen)
- Meldestatus (NeuerFall, NeuerTodesfall, NeuGenesen)

Eine Fallgruppe nimmt eine eineindeutige Ausprägung hinsichtlich ihrer Anzahl von Fällen ("AnzahlFall"), "Altersgruppe", "Geschlecht", ihres Landkreises ("IdLandkreis"), "Meldedatum"s, Erkrankungsdatums ("Refdatum") und der Informationen ob das Erkrankungsdatum bekannt ist "IstErkrankungsbeginn", an.
Weiterhin wird die "AnzahlTodesfall" oder "AnzahlGenesen" jeder Fallgruppe angegeben, wobei

nur eines der beiden Merkmale "AnzahlTodesfall" oder "AnzahlGenesen" angenommen werden kann. Das heißt, sofern es in einer Fallgruppe Todesfälle oder Genesene gibt, werden die Anzahl der Todesfälle oder die Anzahl der genesenen Fälle in einer neuen Gruppe angegeben. Treten z. B. beide Fälle in einer Fallgruppe auf, teilt sich die Fallgruppe in zwei weitere Gruppen auf, und zwar in eine Gruppe der Todesfälle und eine Gruppe der Genesenen.

Beispiel

Es wird eine neue Fallgruppe w registriert (IdLandkreis, Geschlecht, Altersgruppe, Meldedatum, Refdatum, IstErkrankungsbeginn sind konstant). Diese enthält eine Fallgruppe zu Beginn:

> Fallgruppe w: 5 Infizierte, 0 Todesfälle und 0 genesene Fälle

Sterben 1 und genesen 2 der Fälle so spaltet sich die Fallgruppe w in 3 Gruppen:

> Fallgruppe x: 2 Infizierte, 0 Todesfälle und 0 genesene Fälle
> Fallgruppe y: 1 Infizierte, 1 Todesfälle und 0 genesene Fälle
> Fallgruppe z: 2 Infizierte, 0 Todesfälle und 2 genesene Fälle

---

Die Merkmale des Meldestatus geben an, ob, bezogen auf den Vortag, in einer Fallgruppe Veränderungen bei den Infektionsfällen, Todesfällen und Genesenen entstanden sind. Das ermöglicht die Veränderungen zum Vortag nachzuvollziehen. Diese entstehen durch Neumeldungen von Infektionen (inklusive Nachmeldungen), Korrekturen (z. B. durch irrtümliche Meldungen, aber auch Korrekturen bzgl. Landkreis, Alter, Geschlecht oder Erkrankungsbeginn) und Veränderung des Gesundheitszustands (genesen, verstorben).
Die Ausprägungen des Meldestatus spalten Fallgruppen temporär auf. Die Aufspaltung erfolgt temporär, da sie nur die Veränderungen vom Publikationstag zum Vortag abbilden. Neue Fälle bilden für den Tag der Neumeldung eine eigene Fallgruppe. Da ein Fall nur an einem Tag neu gemeldet, neu genesen oder neu verstorben oder korrigiert wird, folgt auf die temporäre Aufspaltung der Fallgruppe am Tag der Neumeldung des Meldestatus, eine Zusammenlegung der Gruppen am Folgetag. Eine genauere Erläuterung zu diesem Prozess wird im folgenden Abschnitt gegeben.

Merkmalsausprägungen

Die Fallzahlendaten enthalten die in der folgenden Tabelle abgebildeten Merkmale und deren Ausprägungen:

Merkmal	Ausprägung	Erläuterung
		Identifikationsnummer des Landkreises basierend auf dem Amtlichen

IdLandkreis	1001 bis 16077	Gemeindeschlüssel (AGS) zuzüglich der 12 Bezirke Berlins (11001 bis 11012); Gebietsstand: 30.06.2020 (2. Quartal)
Geschlecht	W, M, unbekannt	Geschlecht der Fallgruppe: weiblich (W), männlich (M) und (unbekannt)
Altersgruppe	A00-A04, A05-A14, A15-A34, A35-A59, A60-A79, A80+, unbekannt	Altersspanne der in der Gruppe enthaltenen Fälle, stratifiziert nach 0-4 Jahren, 5-14 Jahren, 15-34 Jahren, 35-59 Jahren, 60-79 Jahren, 80+ Jahren sowie unbekannt
Meldedatum	JJJJ-MM-TT	Datum, wann der Fall dem Gesundheitsamt bekannt geworden ist. JJJJ entspricht der Jahreszahl, MM dem Monat und TT dem Tag.
Refdatum	JJJJ-MM-TT	Datum des Erkrankungsbeginns. Wenn das nicht bekannt ist, das Meldedatum.
IstErkrankungsbeginn	0, 1	1: Refdatum ist der Erkrankungsbeginn 0: Refdatum ist das Meldedatum
AnzahlFall	Ganze Zahl	Anzahl der gemeldeten Fälle in der entsprechenden Fallgruppe Für NeuerFall = -1, ist die Anzahl negativ: Es handelt sich um eine Korrektur der Fallgruppe, die angibt, wie viele Infektionen zu viel gemeldet worden sind
AnzahlTodesfall	Ganze Zahl	Anzahl der gemeldeten Todesfälle in der entsprechenden Fallgruppe Für NeuerTodesfall = -1, ist die Anzahl negativ: Es handelt sich um eine Korrektur der Fallgruppe, die angibt, wie viele Todesfälle zu viel gemeldet worden sind
AnzahlGenesen	Ganze Zahl	Anzahl der genesenen Fälle in der entsprechenden Fallgruppe Für NeuGenesen = -1, ist die Anzahl negativ: Es handelt sich um eine Korrektur der Fallgruppe, die angibt, wie viele genesene Fälle zu viel gemeldet worden sind
		0 : Fälle der Gruppe sind in der Publikation

NeuerFall, NeuerTodesfall, NeuGenesen	0, 1, -1	für den aktuellen Tag und in der für den Vortag enthalten. Das bedeutet diese Fälle sind seit mehr als einem Tag bekannt. 1 : Fälle der Gruppe sind erstmals in der aktuellen Publikation enthalten. Das heißt, es sind für den Publikationstag neu übermittelte oder entsprechend neu bewertete Fälle. -1: Fälle der Gruppe sind in der Publikation des Vortags enthalten, werden jedoch nach dem aktuellen Tag aus den Fallzahlendaten entfernt. Das heißt, es sind Fälle die ab dem aktuellen Tag wegfallen. Eine solche Fallgruppe kann beispielsweise durch fälschliche Meldungen entstehen, die so als Korrektur angezeigt werden.
NeuerTodesfall, NeuGenesen	-9	Fälle in der Gruppe sind weder in der Publikation für den aktuellen Tag, noch in der Publikation des Vortags, als genesen ("NeuGenesen") oder verstorben ("NeuerTodesfall") gemeldet. Das bedeutet, dass zu den Fällen in der Gruppe keine Information über den Gesundheitsverlauf der Infektion bekannt ist. Das ist zum Beispiel häufig der Fall, wenn eine Fallgruppe gerade erst als infiziert gemeldet worden ist.

Die temporäre Aufspaltung der Fallgruppen durch die Merkmale des Meldestatus wird im folgenden Beispiel verdeutlicht. Temporäre Gruppen sind durch ein ' gekennzeichnet. Neumeldungen wird bei Betrachtung der Ausprägungen der Merkmale deutlich:

Beispiel

Es wird eine neue Fallgruppe am Tag TT registriert (IdLandkreis, Geschlecht, Altersgruppe, Meldedatum, Refdatum, IstErkrankungsbeginn sind konstant), so nimmt sie den Meldestatus NeuerFall = [1] an. Sind noch keine Genesenen oder Todesfälle bekannt, sind in der Fallgruppe gemeldet, sind NeuerTodesfall und NeuGenesen = [-9]:

Die Fälle der Fallgruppe w' sind im Datensatz von Tag TT neu enthalten (NeuerFall [1]), die Fälle der Gruppe sind keine Todes- oder Genesungsfälle (NeuerTodesfall [-9], NeuGenesen [-9]).

> Fallgruppe w':
> Infizierte [4], Todesfälle [0] und genesene Fälle [0]
> NeuerFall [1], NeuerTodesfall [-9], NeuGenesen [-9]

Am nächten Tag, TT+1 sind die Fälle aus Fallgruppe w' nicht mehr neu. Ihr Meldestatus ändert sich daher von [1] auf [0]. Die temporäre Fallgruppe w' (NeuerFall [1]) wird zur stetigen Fallgruppe w (NeuerFall [0]):

> Fallgruppe w:
> Infizierte [4], Todesfälle [0] und genesene Fälle [0]
> NeuerFall [0], NeuerTodesfall [-9], NeuGenesen [-9]

An Tag TT+1 wird ein zusätzlicher, neuer Fall in der Fallgruppe w registriert. Da es sich um einen neuen Fall handelt, bildet er wieder eine temporäre, eigene Gruppe w':

> Fallgruppe w':
> Infizierte [1], Todesfälle [0] und genesene Fälle [0]
> NeuerFall [1], NeuerTodesfall [-9], NeuGenesen [-9]

Am nächsten Tag, TT+2 sind auch die Fälle der Fallgruppe w'(TT+1) nicht mehr neu, ihr Meldestatus ändert sich wie am Tag zuvor bei Fallgruppe w'(TT). Durch die Änderung des Meldestatus in w'(TT+1), geht w' in w auf. Die Anzahl der Infizierten beider Fallgruppen wird addiert.

> Fallgruppe w:
> Infizierte [5], Todesfälle [0] und genesene Fälle [0]
> NeuerFall [0], NeuerTodesfall [-9], NeuGenesen [-9]

Ähnlich wie mit neuen Infektionsmeldungen verhält es sich mit Meldungen von Todes- oder Gesundungsfällen. Diese bilden temporäre Fallgruppen y' und z' welche später in stetige Fallgruppen y und z übergehen:

Tag TT+3

> Fallgruppe w:
> Infizierte [4], Todesfälle [0] und genesene Fälle [0]
> NeuerFall [0], NeuerTodesfall [-9], NeuGenesen [-9]

> Fallgruppe y':
> Infizierte [1], Todesfälle [1] und genesene Fälle [0]
> NeuerFall [0], NeuerTodesfall [1], NeuGenesen [-9]

Tag TT+4

> Fallgruppe w:
> Infizierte [2], Todesfälle [0] und genesene Fälle [0]
> NeuerFall [0], NeuerTodesfall [-9], NeuGenesen [-9]

> Fallgruppe y:
> Infizierte [1], Todesfälle [1] und genesene Fälle [0]
> NeuerFall [0], NeuerTodesfall [0], NeuGenesen [-9]
>
> Fallgruppe z':
> Infizierte [2], Todesfälle [0] und genesene Fälle [2]
> NeuerFall [0], NeuerTodesfall [-9], NeuGenesen [1]

Hinweis zu Genesenen

Anhand der dem RKI von den Gesundheitsämtern übermittelten Detailinformationen zu einem Erkrankungsfall wird für jeden Fall eine Dauer der Erkrankung geschätzt. Für Fälle, bei denen nur Symptome angegeben sind, die auf einen leichten Erkrankungsverlauf schließen lassen, wird eine Dauer der Erkrankung von 14 Tagen angenommen. Bei hospitalisierten Fällen oder Fällen mit Symptomen, die auf einen schweren Verlauf hindeuten (z. B. Pneumonie) wird eine Dauer der Erkrankung von 28 Tagen angenommen. Ausgehend vom Beginn der Erkrankung, bzw. wenn dieser nicht bekannt ist, vom Meldedatum ergibt sich ein geschätztes Datum der Genesung für jeden Fall. Da im Einzelfall auch deutlich längere Erkrankungsverläufe möglich sind, bzw. die hier genutzten Informationen nicht bei allen Fällen dem RKI übermittelt werden, sind die so berechneten Daten nur grobe Schätzungen für die Anzahl der Genesenen und sollten daher auch nur unter Berücksichtigung dieser Limitationen verwendet werden.

Formatierung der Fallzahlen Daten

Die Fallzahlen Daten sind im Datensatz als kommaseparierte .csv Datei enthalten. Der verwendete Zeichensatz der .csv Datei ist UTF-8. Trennzeichen der einzelnen Werte ist ein Komma ",". Datumsangaben sind im ISO8601 Standard formatiert.

- Zeichensatz: UTF-8
- Datumsformat: ISO8601
- .csv Trennzeichen: Komma ","

Metadaten

Zur Erhöhung der Auffindbarkeit sind die bereitgestellten Daten mit Metadaten beschrieben. Über GitHub Actions werden Metadaten an die entsprechenden Plattformen verteilt. Für jede Plattform existiert eine spezifische Metadatendatei, diese sind im Metadatenordner hinterlegt:

> Metadaten/

Versionierung und DOI-Vergabe erfolgt über Zenodo.org. Die für den Import in Zenodo bereitgestellten Metadaten sind in der zenodo.json hinterlegt. Die Dokumentation der einzelnen Metadatenvariablen ist unter https://developers.zenodo.org/#representation nachlesbar.

> Metadaten/zenodo.json

Hinweise zur Nachnutzung der Daten

Offene Forschungsdaten des RKI werden auf GitHub.com, Zenodo.org und Edoc.rki.de bereitgestellt:

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

Lizenz

Der Datensatz "SARS-CoV-2 Infektionen in Deutschland" ist lizenziert unter der Creative Commons Namensnennung 4.0 International Public License | CC-BY 4.0 International.

Die im Datensatz bereitgestellten Daten sind, unter Bedingung der Namensnennung der Quelle, frei verfügbar. Das bedeutet, jede Person hat das Recht die Daten zu verarbeiten und zu verändern, Derivate des Datensatzes zu erstellen und sie für kommerzielle und nicht kommerzielle Zwecke zu nutzen. Weitere Informationen zur Lizenz finden sich in der LICENSE bzw. LIZENZ Datei des Datensatzes.