Datensatzdokumentation

Corona-Datenspende - Teildatensatz Vitaldaten

Robert Koch-Institut | RKI
Nordufer 20
13353 Berlin

Marc Wiedermann, Robert W. Bruckmann & Dirk Brockmann
P4 | Epidemiologische Modellierung von Infektionskrankheiten | Robert Koch-Institut
Research on Complex Systems | Institut für theoretische Biologie | Humboldt-Universität zu Berlin

Beitragende:
MF 4 | Fach- und Forschungsdatenmanagement
Hannes Wuensche (Datenkuration)

Zitieren
Wiedermann M., Bruckmann, R. W., Brockmann D. (2023). Corona-Datenspende - Teildatensatz Vitaldaten. Zenodo. DOI: 10.5281/zenodo.8229284.

Informationen zum Datensatz und Entstehungskontext

Die Daten von Fitnessarmbändern und Smartwatches, sogenannten Wearables, können Hinweise auf Symptome einer COVID-19 Erkrankung liefern. Mit Hilfe der Corona-Datenspende-App (CDA) konnten Bürger:innen dem Robert Koch-Institut solche Daten für wissenschaftliche Zwecke zur Verfügung stellen. Zusammen mit Informationen aus anderen Quellen, z.B. offiziellen Meldedaten zu Fallzahlen, helfen diese Daten den Wissenschaftler:innen, die Ausbreitung des Coronavirus besser zu erfassen und zu verstehen. Die Datenerhebung im Rahmen der CDA wurde zum Jahreswechsel 2022/2023 planmäßig eingestellt.

Seit ihrer ersten verfügbaren Version (April 2020) erhob die CDA sogenannte Vitaldaten der Nutzer:innen. Dies beinhaltet insbesondere den täglichen Ruhepuls, die körperlichen Aktivität und das Schlafverhalten.

Die in diesem Repository bereit gestellten Datenpunkten beinhalten räumlich und zeitlich aggregierte Informationen zum mittleren Ruhepuls, der mittleren täglichen Schrittzahl und der mittleren Schlafdauer pro Tag und Landkreis & kreisfreier Stadt, Regierungsbezirk, Bundesland sowie dem täglichen deutschlandweiten Mittelwert. Diese Zuordnung orientiert sich an den Definitionen der europäischen Gebietseinheiten (*NUTS*) von *NUTS3* (Landkreise) zu *NUTS0* (Nationalstaat).

Eine visualle und interaktive Aufbereitung der Daten findet sich bereits im Vitaldaten-Explorer, welcher durch das Team der CDA bereit gestellt wurde.

Die hier bereitgestellten Datenpunkte dienen der Weiternutzung in der Wissenschaft und der interessierten Öffentlichkeit. Sie decken den vollständigen Erhebungszeitraum der CDA vom April 2020 bis Dezember 2022 ab. Da es sich bei den bereit gestellten Daten um räumliche Mittelwerte handelt sind keine Rückschlüsse auf Einzelpersonen möglich.

Weitere Information zur CDA finden sich auf corona-datenspende.de/science.

Projektbeteiligte und Rollenbesetzung

Herausgeber der CDA ist das Robert Koch-Institut, eine deutsche Bundesbehörde im Geschäftsbereich des Bundesministeriums für Gesundheit. Die Projektleitung und -koordination liegt bei der Projektgruppe P4 | Modellierung von Infektionskrankheiten. Die App wurde in Zusammenarbeit mit Thryve (mHealth Pioneers GmbH) entwickelt, einem auf Digital Health spezialisierten Unternehmen. Dieses Unternehmen ist der technologische Dienstleister. Fragen bezüglich der Corona-Datenspende können an corona-datenspende@rki.de gerichtet werden.

Die Veröffentlichung der Daten sowie das Qualitätsmanagement der (Meta-)Daten erfolgen durch das Fachgebiet MF 4 | Fach- und Forschungsdatenmanagement. Fragen zum Datenmanagement und zur Publikationsinfrastruktur können an das Open Data Team des Fachgebiets MF4 unter OpenData@rki.de gerichtet werden.

Durch sorgfältige Auswahl und regelmäßige Kontrolle stellt das Robert Koch-Institut sicher, dass der Dienstleister alle organisatorischen und technischen Maßnahmen trifft, die zum Schutz der Daten erforderlich sind. Alle Maßnahmen stehen im Einklang mit dem geltenden Datenschutzrecht.

Datenerhebung & Technische Voraussetzungen

Bis zum 31.12.2022 konnten sich Nutzer:innen die CDA in einem der gängigen App-Stores (Google Play, Apple AppStore, Huawai AppGallery) herunterladen. Neben einem Smartphone mit installierter CDA wurde für die Teilnahme am sogenannten Fiebermonitor ein Fitnessarmband oder eine Smartwatch benötigt. Die CDA unterstützte in der letzten Version Geräte von Apple, Samsung, Fitbit, Garmin, Amazfit, Oura, Polar und Withings.

Bei einer Neuanmeldung wurde als Teil der grundlegenden demografischen Variablen die Postleitzahl erfasst, welche sich eindeutig einer NUTS3-Region (und somit auch den übergeordneten Regionen NUTS2 bis NUTS0) zuordnen lässt. In späteren Versionen der CDA wurden bei Neuanmeldungen lediglich die ersten drei Stellen der Postleitzahl abgefragt. Hiervon betroffene Teilnehmer:innen sind im vorliegenden Datensatz nicht berücksichtigt, die sich keine eindeutige Zuordnung zu NUTS3 vornehmen lässt.

Alle Nutzer:innen der CDA nahmen automatisch am Teilprojekt Fiebermonitor teil, in dessen Kontext die hier bereitgestellen Vitaldaten erhoben und ausgewertet wurden. Optional war eine Teilnahme an Befragungsstudien möglich. Die hier veröffentlichten Daten umfassen keine Befragungsdaten.

Die Teilnahme an der CDA war freiwillig und unentgeltlich. Alle Teilnehmer:innen erklärten sich mit der anonymisierten Verwendung ihrer Daten für wissenschaftliche Zwecke einverstanden.

Teilnehmer:innen

Ein- und Ausschlusskriterien

Eine Teilnahme war allen Personen ab 16 Jahren möglich, die bis zum Jahreswechsel 2022/2023 Zugang zu einen deutschen App-Store hatten. Eine Veröffentlichung der CDA in App Stores außerhalb von Deutschland war nicht möglich.

Aus der Aggregation entfernt wurden:

- Personen mit Postleitzahlkürzel, das nicht trennscharf zu einer NUTS3-Gebietseinheit zugeordnet werden kann. Hierzu zählen inbesondere Teilnehmer:innen für die nur die ersten drei Stellen der Postleitzahl vorliegen (siehe oben)
- Schlafdaten von Nutzer:innen einer *Apple Watch* aufgrund von Unregelmäßigkeiten in der Messmethodik
- unplausible Vitalwerte auf Individualebene, insbesondere Datenpunkte mit
    - mehr als 50.000 Schritte pro Tag
    - mehr als 24 Stunden Schlafdauer
    - täglicher Ruhepuls unter 30 und über 150 Schlägen pro Minute

Grundgesamtheit

Insgesamt hatten mehr als eine Million Menschen die CDA installiert. Mehr als 500,000 Menschen haben schlussendlich mindestens ein Wearable verbunden und somit mindestens einen Datenpunkt übermittelt. An den Befragungsstudien nahmen regelmäßig bis zu 30,000 Menschen teil.

Der hier bereitgestellt Datensatz aggregiert die Wearable-Messungen aller Nutzer:innen über die gesamte Projektlaufzeit (991 Tage, 15.04.2020 bis 31.12.2022), und stellt tägliche Mittelwerte für Ruhepuls, Schlafdauer und Schrittanzahl über alle Gebietseinheiten der Bundesrepublik Deutschland zur Verfügung. Im Durchschnitt ergibt sich ein Tagesmittelwert in den Landkreisen bzw. den kreisfreien Städten aus 507 Personen für die Schrittzählung, 158 Personen für Schlafaufzeichnungen, und 310 Personen für Ruhepulsmessungen. In jeden bereitgestellten Mittelwert gehen dabei mindestens drei individuelle Datenpunkte ein.

Insgesamt werden die Vitaldaten von 493487 Teilnehmer:innen berücksichtigt. Davon haben 199.033 Teilnehmer:innen freiwillige Angaben zum eigenen Geschlecht gemacht. Hierbei gaben 75220 (37.79%) Personen an weiblich zu sein, 123563 (62.08%) waren männlich und 250 (0.13%) divers.

Aufbau und Inhalt des Datensatzes

Zentraler Inhalt des Datensatzes sind Aggregationen der innerhalb der CDA erhobenen Vitaldaten. Im Datensatz insgesamt enthalten sind:

- Räumlich aggregierte tägliche Vitaldaten (451.896 Zeilen, 9 Variablen)
- Lizenz-Dateien mit der Nutzungslizenz des Datensatzes in Deutsch und Englisch
- Datensatzdokumentation in deutscher Sprache
- Metadaten zur Bereitstellung

Die 451.896 Zeilen des Datensatzen Schlüssen sich auf in 991 x 401 Datenpunkte pro NUTS3-Gebietsebene, 991 x 38 Datenpunkte pro NUTS2-Gebietsebene, 991 x 16 Datenpunkte pro NUTS1-Gebietsebene, sowie 991 Datenpunkte für die NUTS0-Gebietsebene.

Vitaldaten

Formatierung der Daten

Die Daten der Studie sind im Datensatz als tabulatorgetrennte `.tsv` Datei enthalten. Der verwendete Zeichensatz der `.tsv` Datei ist UTF-8. Trennzeichen der einzelnen Werte ist das Tabulatorzeichen. Datumsangaben entsprechen dem Standard ISO-8601. Für jede Kombination aus NUTS-Region und Datum enthält die `.tsv` Datei eine Zeile mit den gemittelten Vitalparametern und den dazugehörigen Standardfehlern.

- Zeichensatz: UTF-8
- Datumsformat: ISO 8601
- Enthaltenes Dateiformat: `.tsv`
- Trennzeichen: Tab `\t`

Variablen und Variablenausprägungen

Insgesamt werden die arithmetisch gemittelten Tageswerte der drei Indikatoren Ruhepuls, Schlafdauer und Schrittzahl zur Verfügung gestellt (siehe oben). Für jeden Mittelwert stellt der Datensatz zusätzlich einen Präzisionsschätzer in Form des Standardfehlers bereit. Der Standardfehler berechnet sich aus der Standardabweichung des Mittelwerts und der Anzahl der in den Mittelwert einfließenden Datenpunkte. So sind Aussagen über die Genauigkeit der jeweiligen Werte möglich. Diese Daten sind für jeden Tag im Projektzeitraum und jede NUTS-Gebietseinheit in Deutschland verfügbar. Eine Zuordnung aller NUTS-Codes zu den entsprechenden Landkreisen, Regierungsbezirken und Bundesländern findet sich unter anderem in der deutschsprachigen Wikipedia. Zu beachten ist, dass im vorliegenden Datensatz die Zuordnung *NUTS 2016* vom 18. Januar 2018 verwendet wurde, welche 401 Landkreise und kreisfreie Städte definiert.

Varname	Format	Ausprägungen	Kurzbeschreibung
date	Datum	`2020-04-15` bis `2022-12-31`	Datumsvariable nach ISO-8601 Standard
			Die Nomenclature des Unités territoriales statistiques (NUTS) ist ein System zur geografischen Gebietseinteilung in der Europäischen Union. Sie ist in der Bundesrepublik Deutschland wie folgt definiert:

nuts_level	Text	`NUTS0`, `NUTS1`, `NUTS2`, `NUTS3`	- NUTS0: Bundesebene - NUTS1: 16 Bundesländer - NUTS2: 38 Regionen 19 Regierungsbezirke</li 10 ehemalige Regierungsbezirke</li 9 Länder ohne weitere Unterteilung) - NUTS3 401 Landkreise & kreisfreie Städte
nuts_code	Text	2- bis 5-stelliger, alphanumerischer Gebietscode der entsprechenden NUTS-Ebenen (z.B. `DE111`)	Jede Gebietseinheit besitzt eine eindeutige ID, und kann - sofern vorhanden - der entsprechenden Obereinheit zugeordnet werden. Beispiel: - DE11D (5-stellig, NUTS3): Ostalbkreis - DE11 (4-stellig, NUTS2): Stuttgart - DE1 (3-stellig, NUTS1): Baden-Württemberg - DE (2-stellig, NUTS0): Bundesrepublik Deutschland
heartrate_mean	Dezimalzahl	positiv reelle Zahlen (auf 2 Nachkommastellen	tagesgemittelter Ruhepuls

		gerundet)	
heartrate_standard_error	Dezimalzahl	positiv reelle Zahlen (auf 2 Nachkommastellen gerundet)	Parameter zur Einschätzung der Streuung und Präzisio in den Ruhepuls-Tagesmittelwerten.
steps_mean	ganzzahlig	positive natürliche Zahlen (auf 2 Nachkommastellen gerundet)	tagesgemittelte Schrittzahl
steps_standard_error	ganzzahlig	positive natürliche Zahlen (auf 2 Nachkommastellen gerundet)	Parameter zur Einschätzung der Streuung und Präzisio in den Schrittzahl-Tagesmittelwerten.
sleep_mean	Dezimalzahl	positiv reelle Zahlen (auf 2 Nachkommastellen gerundet)	tagesgemittelte Schlafdauer
sleep_standard_error	Dezimalzahl	positiv reelle Zahlen (auf 2 Nachkommastellen gerundet)	Parameter zur Einschätzung der Streuung und Präzisio in den Schlafdauer-Tagesmittelwerten.

Kontextmaterialien

Umfassende weitere Informationen zur Corona-Datenspende sowie aktuelle Auswertungsergebnisse finden sich auf https://corona-datenspende.de/science/.

Die hier bereitgestellten Datenpunkte können über den Vitaldaten-Explorer der Corona-Datenspende visualisiert und exploriert werden.

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