HOTSPOT Data Challenge

Beschreibung der Ausgabedaten

Stand: 25. August 2026

1. Worum es geht

Im Projekt Data4Health.Ruhr werden personenbezogene Gesundheitsdaten aus verschiedenen Quellen (Krankenhaus-Informationssystem, Gesundheitsamt Bochum) zu **räumlich aggregierten Kennzahlen** verdichtet und sollen anschließend als Karten- und Analysegrundlage weitergegeben werden. Die hier bereitgestellten CSV-Dateien sind genau dieser Aggregations-Output — das, was das System nach außen gibt.

Wichtig: Die Eingangsdaten sind künstlich erzeugte Musterdaten. Diese wurden aber durch die **unveränderte Produktions-Pipeline** geschickt, d. h. Struktur, Aggregationslogik und Anonymisierungsgate entsprechen exakt dem Echtbetrieb.

2. Wie die Daten entstehen (Verarbeitungskette)

1. Rohdatensatz: eine Zeile = eine Person bzw. ein Fall, mit Adresse, Geschlecht, Alter und Befunden.
2. Die Adresse wird geokodiert (Photon) und ergibt Längen-/Breitengrad.
3. Zuordnung zu einer 1-km-Zelle des BKG-Gitters (ETRS89-LAEA).
4. Ableitung der Zeitperiode (Jahr bzw. Jahr + Quartal).
5. **Deduplizierung** auf eine Zeile pro Person pro Periode: mehrere Klinikbesuche im selben Quartal zählen als eine Person; Merkmals-Flags werden per logischem ODER über alle Fälle der Person zusammengefasst.
6. Aggregation entlang von drei Achsen: gesamt / nach Geschlecht / nach Altersgruppe.
7. Inner-Join gegen drei Zensus-2021-Tabellen (k-Anonymitäts-Gate und Bevölkerungsnenner).
8. CSV-Export.

3. Dateistruktur

Vier Datenbestände in den Ordnern `aggregated_output_<name>`, je Zeitauflösung immer **drei** Dateien:

<name>	Inhalt	Zeitauflösung	Zeilen	Zellen
`kkukb`	Stationäre Klinikfälle, 7 chronische Erkrankungen	quartalsweise	27.000–38.000	∼3.800
`sterbefaelle`	Sterbefälle (natürlich / ungeklärt)	jährlich *und* quartalsweise	750–4.000	∼140
`zahnaerztlich`	Zahnärztlicher Dienst (Gebissstatus, Kariesrisiko)	jährlich	760–1.400	∼143
`schuleingang`	Schuleingangsuntersuchung (Impfungen, BMI, Befunde)	jährlich	740–1.350	∼149

Die drei Dateien pro Auflösung sind **drei Randverteilungen derselben Personenmenge**:

- `filter1_total` — alles zusammengefasst
- `filter2_sex` — zusätzlich nach Geschlecht aufgespalten (`M` / `F`)
- `filter3_age` — zusätzlich nach Altersgruppe aufgespalten (`0-15`, `16-64`, `65+`)

Formal: `;` als Trenner, Komma als Dezimalzeichen, UTF-8 mit BOM.

4. Was eine Zeile bedeutet

Eine Zeile = eine Kombination aus Ort × Zeit × (optional) demografischem Merkmal.

Beispiel aus `aggregated_output_kkukb/..._filter3_age.csv`:

```
1kmN2717E4295;4295000;2717000;2022;2;65+;0;0;0;0;1;0;0;1;718
```

Lesart: *In der 1-km²-Zelle mit dem Südwest-Eckpunkt (4295000, 2717000) gab es im 2. Quartal 2022 genau **eine** Person der Altersgruppe 65+ mit stationärem Aufenthalt; bei dieser Person war eine Hypertonie dokumentiert. In dieser Zelle leben laut Zensus 2021 insgesamt 718 Personen der Altersgruppe 65+.*

Zeilen, in denen ein Wert 0 wäre, werden nicht ausgegeben — die Datei ist ein „long format“ mit ausschließlich belegten Kombinationen.

5. Was eine Spalte bedeutet

Block	Spalten	Bedeutung
Ort	`grid_id`	ID der 1-km-Gitterzelle, Format `1kmN<northing/1000>E<easting/1000>`
	`easting`, `northing`	Koordinaten der Zellenecke in Metern (EPSG:3035) — exakt und unverrauscht
Zeit	`year`	Kalenderjahr
	`quarter`	Quartal 1–4 (nur bei quartalsweiser Auflösung)
Demografie	`sex`	nur in `filter2`: `M` oder `F`
	`age_group`	nur in `filter3`: `0-15`, `16-64`, `65+`
Indikatoren	`<merkmal>_cases`	**Anzahl unterschiedlicher Personen** in dieser Zelle / Periode / Gruppe, auf die das Merkmal zutrifft. Trotz Suffix `_cases` sind es Personen-, keine Fallzahlen.
Summe	`total_cases`	Anzahl unterschiedlicher Personen in dieser Zelle / Periode / Gruppe insgesamt
Nenner	`population`	Bevölkerung der Zelle (bzw. der Geschlechts- oder Altersgruppe darin) laut **Zensus 2021**; dient zur Ratenbildung

Die Indikatorspalten je Datensatz

- **KKUKB (7 Spalten):** `cerebral_infarction`, `dementia`, `depression`, `diabetes_type_2`, `hypertension`, `myocardial_infarction`, `obesity` — abgeleitet aus ICD-10-Präfixen (E11, E66, I10, I63, I21,

F32/F33, F00–F03/G30). Die Flags sind **nicht disjunkt**: eine Person kann in mehreren Spalten zählen.

- **Sterbefälle (2):** `natuerlich`, `ungeklaert`
- **Zahnärztlich (4):** `behandlungsbeduerftig`, `saniert`, `gesund`, `kariesrisiko`
- **Schuleingang (∼100):** je Untersuchungsmerkmal eine Spalte *pro Ausprägung*, benannt `<merkmal>__<auspraegung>_cases`, z. B. `impf_masern__ja_cases` / `impf_masern__nein_cases`, `bmi_perz_wertung__adipositas_cases`, `bef53_sprachstoerung__auffaellig_abklaerung_erforderlich_cases`. Diese Spaltengruppen sind pro Merkmal **disjunkt und summieren sich** (bis auf fehlende Angaben) auf `total_cases`.

6. Ziel der Challenge

Aufgabe: Brecht die Anonymität dieser Ausgabedaten.

Zeigt einen nachvollziehbaren Weg auf, wie aus diesen aggregierten CSV-Dateien Rückschlüsse auf einzelne Personen möglich sind — oder wie Ihr dem sehr nahe kommt. Gefragt ist nicht nur eine theoretische Möglichkeit, sondern ein **konkret beschriebenes Vorgehen**, idealerweise mit Rechenweg, Beispielzeile aus den Daten und Angabe der benötigten Zusatzinformationen.

Bewertet wird:

1. **Angriffsweg** — Welche Schritte, welche Annahmen, welches Vorwissen ist nötig?
2. **Zusatzdaten** — Welche externen Quellen setzt Ihr ein? (öffentlich verfügbar? kommerziell? persönliche Ortskenntnis? andere Behördendaten?)
3. **Ergebnis** — Was genau erfahrt Ihr über eine identifizierte Person, und wie sicher ist die Zuordnung?
4. **Skalierbarkeit** — Funktioniert das für einen Einzelfall oder systematisch für viele Personen?
5. **Gegenmaßnahme** — Welche Änderung an der Pipeline würde genau diesen Angriff verhindern, und was kostet das an analytischem Nutzen?

7. Einreichung der Ergebnisse

Eure Ergebnisse sendet Ihr als PDF mit maximal 10 Seiten oder Slides per E-Mail an info@data-science.ruhr
Die Deadline für die Einreichung ist Sonntag, 13. September 2026, 23:59 Uhr.