**Dokumentation der Datenanalyse**

**Projektakronym:** NUM-CON-MON

**Ansprechpartner[[1]](#footnote-1)**: Marie Gebhardt (TMF e.V.), Skriptersteller war freundlicherweise Alexander Kiel (Universität Leipzig), das Ursprungsprojekt wurde von Dr. Thomas Datzmann eingereicht

### Dokumentation zum Analyseprogramm

#### Ausführungsumgebung

###### Container-Image Verwaltung/Betriebssysteme/virtuelle Maschinen

* Linux Komandozeile mit Bash
* blazectl [v0.17.0, <https://github.com/samply/blazectl>]
* jq [v1.7.1, <https://jqlang.org/>]

###### Programmiersprachen und abhängige Bibliotheken

* CQL, [v1.5, <https://cql.hl7.org/>]
* jq [v1.7.1, <https://jqlang.org/>]

###### Zu verwendende Skripte:

Repository: [https://github.com/medizininformatik-initiative/NUM\_CON\_MON](https://github.com/medizininformatik-initiative/NUM_CON_MON/tree/main)

* Ausführung der Skripte ist im Readme erläutert: <https://github.com/medizininformatik-initiative/NUM_CON_MON/blob/main/README.md>
* Berechnung der aggregierten Statistiken inkl. Deidentifizierung (evaluate-measure.sh)
* Versand des Reports (send-report.sh)
* Informationen zum Deployment und Versand der Ergebnisse via DSF an das FDPG (DEPLOYMENT.md)

#### Datenextraktion

entfällt, da hier auf den FHIR-Daten direkt eine FHIR Measure per CQL berechnet wird

###### Ablauf der verteilten Analyse mithilfe von Blaze (hier Berechnung der aggregierten Statistiken)

Das Skript evaluate-measure.sh ermöglicht über den Aufruf von num-con-mon.cql:

1. die Erstellung von Kohorte 1: Patienten und Patientinnen mit Broad Consent
2. die Erstellung von Kohorte 2: Kohorte 1 mit Einwilligung in Nutzung der Krankenkassendaten für wissenschaftliche Zwecke
3. Ermittlung der IK-Nummer
4. Ermittlung der Abteilungskontakte aus dem Attribut Kontaktebene mit einem zeitlichen Abstand von maximal 2 Monaten zum Fallereignis
5. Ermittlung der Fachabteilungsschlüssel für die unter 4. ermittelten Fälle, sofern verfügbar
6. Ermittlung des Geschlechts der Patienten und Patientinnen (m, w, d, unbestimmt)
7. Ermittlung der Altersklasse der Patienten und Patientinnen (<18, 18-64, >=65)

Das Ergebnis wird mithilfe von de-identify.sh so aufbereitet, dass die Daten anonym ausgeleitet werden können. Wie die Anonymisierung erreicht wird, findet sich im folgenden Kapitel „Anonymisierungsschritte“.

###### Beschreibung der Eingabe

Das Skript wird direkt auf dem FHIR-Server des Datenintegrationszentrums ausgeführt, der auch an die Machbarkeitstools des FDPG angeschlossen ist. Dies ist nicht im Widerspruch zum MII-Datenschutzkonzept, da es sich um eine Schaufensteranfrage handelt, die nach den Prinzipien der Machbarkeitsanfragen funktioniert.

###### Beschreibung der Ausgabedaten

*Tabelle 1: Beschreibung der Ausgabedateien der Datenanalyse*

|  |  |  |  |
| --- | --- | --- | --- |
| **Dateiname** | | **Format**: | **Soll ausgeleitet werden** |
| **1** | ik-number.csv | CSV | Nein (informativ) |
| **2** | department-key.csv | CSV | Nein (informativ) |
| **3** | gender.csv | CSV | Nein (informativ) |
| **4** | age-class.csv | CSV | Nein (informativ) |
| **5** | report-de-identified.json  Gesamtergebnis für die Statistik | JSON (FHIR MeasureReport) | ja |

###### Hinweise für die Umsetzung der Berechnung ohne CQL

Eine ELM Version des CQL Scripts steht unter num-con-mon-elm.json zur Verfügung. Bitte auf Aktualität achten, da diese Datei manuell erstellt wird.

Eine Berechnung per FHIR-Search ist nicht vorgesehen.

**Anonymisierungsschritte**

Im Skript de-identify.sh wird ein Grenzwert/Threshold (Default=10) definiert für die Größe der Populationsgruppen einzelner Stratifier. Alle Strata für die der Grenzwert unterschritten wird, werden in ein einzelnes zusammengefasst, wobei die Werte kommagetrennt zusammengefasst werden und die Anzahlen entsprechend summiert werden. Falls der Grenzwert dann immer noch unterschritten wird, wird diese Anzahl maskiert.

Die genaue Herangehensweise für die Strata kann den zugehörigen „.jq“ Skripten entnommen werden.

1. **Rechenressourcen**

Mit der Ausführung der Datenanalyse gehen die folgenden technischen Anforderungen einher. Die notwendigen technischen Ressourcen sind ausgehend von in der Entwicklung der Skripte verwendeten Testdaten geschätzt und können in Bezug auf die beim jeweiligen Partner verfügbaren Datenvolumen differieren.

Grundsätzlich werden die Ressourcen des FDPG FHIR Servers benutzt, da das CQL Script direkt auf dem Server ausgeführt wird. Der Linux Host auf dem die Bash Skripte ausgeführt werden, benötigt lediglich minimale Ressourcen. Die Deidentifizierung per jq findet bereits auf aggregierten Daten statt.

*Tabelle 2: Bei Ausführung verwendete Ressourcen*

|  |  |
| --- | --- |
| Anzahl CPU Cores | *1* |
| Anzahl GPUs: | *0* |
| RAM in GB: | *1* |
| Peripherspeicher in GB: | *0,1* |
| Geschätzte Ausführungszeit | *Sekunden bis Minuten* |

1. Hier bitte Antragstellende Person oder verantwortlichen Wissenschaftler eintragen. Die Person sollte als Ansprechpartner für Fragen der DIZ zur Verfügung stehen. [↑](#footnote-ref-1)