**Datensatzdokumentation** Entwurf, Stand 22.11.2022

**Impf-Hotline in Deutschland**

**Informationen zum Datensatz und Entstehungskontext**

Die COVID-19-Impfung gilt als Wendepunkt in der Kontrolle der COVID-19-Pandemie und erfährt daher ein hohes Maß an öffentlicher Aufmerksamkeit. Ziel der am 27.12.2020 beginnenden Impfkampagne war es, möglichst schnell eine hohe Impfquote zu erreichen. Dabei gab es insbesondere in den ersten Monaten nach dem Starttermin eine hohe Nachfrage nach COVID-19-Impfungen, die zunächst ausschließlich in Impfzentren vorgenommen wurden. Zudem gab es in der Bevölkerung ein großes Aufkommen an Fragen zur COVID-10-Impfung, da es sich um neue Impfstoffe und auch um eine neuartige Vorgehensweise bei einer Impfkampagne handelte.

Um die Termine zu koordinieren und die mit der Impfung und der COVID-19-Pandemie aufkommenden Fragen zu beantworten, wurde in Deutschland eine bundesweit einheitliche Telefon-Hotline eingerichtet. Diese wird von verschiedenen Telefonstudios betrieben und ist unter der bundesweiten Patientenservice-Rufnummer 116 117 erreichbar. Die im Rahmen des Telefonbetriebs anfallenden Daten werden protokolliert und gespeichert, dabei handelt es sich auch

**Rollenbesetzung**

Projektleiter mar,an,con GmbH: Prof. Dr. Peter Lorscheid. Bei der mar,an,con GmbH werden ACD- und Sabio-Daten tagesaktuell zusammengefügt, bereinigt und aufbereitet und anschließend auf GitHub zur Verfügung gestellt.

*Weitere Rollen noch unklar.*

**Zusammensetzung der Datenquellen**

Bei den ACD-Daten (ACD=Automatic Call Distribution) handelt es sich um Daten, die automatisch während des Telefonstudiobetriebs anfallen. Jeder Anruf erhält einen Identifikationscode; zu einem Identifikationscode werden i.d.R. mehrere Datensätze erstellt, die die verschiedenen Stufen des Anrufs repräsentieren. Als Stufen werden hierbei unterschieden:

* Eintritt in ACD-Plattform
* Initiales Gespräch
* Versuch Weitervermittlung

Für jede Gesprächsstufe werden dabei die Details festgehalten, insbesondere

* Telefonnummer des Anrufenden
* die Telefonnummer, über die die Verbindung gesucht wurde (neben der 116117 gibt es einige weitere Nummern, etwa die Arztrufzentrale NRW)
* Organisation und Name des Telefonagenten (nur für die Gesprächsschritte, bei denen mit einem Agenten verbunden wurde)
* die Zeitpunkt des Verbindungsaufbaus in dieser Anrufstufe
* die Gesamtdauer der Anrufstufe
* die Dauer zur Herstellung der Verbindung, aufgeteilt in die Wartezeit bis zum Klingeln, die Klingelzeit bis zur Onlineschaltung des Anrufs und die SRN-Introzeit,
* die Online-Dauer der Anrufstufe sowie weitere Zeiten wie Haltezeit, Rückfragenzeit;
* der Aufleger, d.h. diejenige Person, die diese Stufe des Gesprächs beendet hat (entweder der Agent oder der Anrufer).

Angaben zu den Gesprächsinhalten sind in den ACD-Daten nicht enthalten. Diese entstammen den Daten, die durch die Sabio Serviceware zur Verfügung gestellt werden. Hierbei werden die Gesprächsinhalte verschlagwortet. Dabei könne einem Gespräch mehrere Anfrage-Inhalte zugeordnet werden, wodurch mehrere Datensätze in dem Sabio-Datensatz entstehen. Neben der Telefonhotline werden in den Sabio-Daten auch Besuche der FAQ-Webseite erfasst.

Insgesamt besteht der Sabio-Datensatz aus folgenden Angaben:

* Angaben zum Benutzer, dies ist i.d.R. der Telefonagent, inkl. Login-E-Mail des Benutzers und der Zuordnung des Benutzers zu einer Benutzergruppe;
* die Verschlagwortung des Gesprächsinhalts („Titel-Ast-Nachricht-Dokument“);
* die Treffermenge *(was ist das?*);
* die Dauer der Anfrage in Sekunden;
* Datum und Uhrzeit der Anfrage.

**Aufbereitung der übermittelten Daten**

Ziel der Aufbereitung ist es, für jede Anruf-Id einen Datensatz zur Verfügung zu stellen, in dem alle relevanten Informationen enthalten sind. Dazu müssen zunächst die ACD-Datensätze zu den unterschiedlichen Stufen eines Anrufs zusammengefasst werden; anschließend alle zu dem jeweiligen Anruf gehörenden Sabio-Daten angefügt und im hinsichtlich der Gesprächsinhalte ausgewertet werden.

Die in den einzelnen Datensätzen zu einer Anruf-Id des ACD-Datensatzes enthaltenen Zeitinformationen werden aggregiert, mit der aggregierten Gesamtlänge wird zu dem Verbindungszeitpunkt, der den Beginn des Anrufs repräsentiert, auch der Endzeitpunkt des Anrufs berechnet. Aus den Stufeninformationen wird ermittelt, welche Phasen des Gesprächs durchlaufen werden, wobei nun 5 ACD-Phasen unterschieden werden. Dies wird durch Dummy-Variablen gekennzeichnet:

* Link only: Es wird nur eine Verbindung aufgebaut, ein Eintritt in das ACD-System findet aber noch nicht statt. Wenn diese Dummy-Variable auf „1“ steht, sind die vier übrigen „0“.
* ACD-Admission: Es findet ein Eintritt in das ACD-System statt. Diese Dummy-Variable ist immer „1“, wenn die Link only-Variable „0“ ist.
* Try call: Es findet ein Anrufversuch statt, aber die Verbindung wird beendet, bevor der Kontakt zu einem Agenten zustande kommt.
* Call: Es kommt der Kontakt zu einem Agenten statt. Diese Variable steht immer auf „1“, wenn ACD-Admission „1“ ist und Try call „0“ ist.
* Try forwarding: Diese Variable drückt aus, ob der Agent den Versuch einer Weiterverbindung unternommen hat. Sie kann „1“ sein, wenn Call ebenfalls „1“ ist.

Zudem wird die Telefonnummer des Anrufenden analysiert. Dabei wird zunächst unterschieden, ob es sich um eine Mobil- oder Festnetznummer handelt. Die Festnetznummer werden zudem über die Telefonvorwahl regional zugeordnet. Aus der Vorwahl wird die Stadt/Gemeinde und der Kreis anhand des 5-stelligen Kreis-Gemeinde-Schlüssels (KGS) ermittelt. Das Bundesland ergibt sich dann aus den ersten beiden Stellen des KGS.

Zu beachten ist allerdings, dass die regionale Zuordnung zu Städten/Gemeinden aus der Telefonvorwahl nicht überall eindeutig ist. So wird etwa die Vorwahl von Hamburg („040“) nicht nur in der Hansestadt Hamburg, sondern auch in 18 weiteren Gemeinden im Hamburger Umland verwendet (z.B. Neu Wulmstorf, Norderstedt und Reinbek). Ein Anrufer aus diesen Gemeinden, die in Schleswig-Holstein und Niedersachsen liegen, würde also auch dem falschen Kreis und in diesem Fall sogar dem falschen Bundesland zugeordnet. Umgekehrt hat die zu Hamburg gehörende Nordseeinsel Neuwerk die Vorwahl „04721“ von Cuxhaven (Niedersachsen), in diesem Fall stimmt also die Zuordnung eines Neuwerker Anrufers zu Kreis und Bundesland ebenfalls nicht. Trotz der vorgenannten Einschränkungen sind die regionalen Zuordnungen hinreichend genau, d.h. die meisten Anrufer mit Vorwahl „040“ rufen aus Hamburg an und nur sehr wenige Anrufe mit einer anderen Vorwahl als „040“ kommen dennoch aus Hamburg.

Schwerwiegender ist, dass die Mobilfunknummern nicht regional zugeordnet werden können. Dabei ist zu beachten, dass der Anteil der Haushalte, die keine Festnetznummer haben bzw. die beim Telefonieren die Mobilnummer bevorzugen, regional unterschiedlich und in urbanen Regionen tendenziell höher als auf dem Land ist. Daher ist die regionale Verteilung der Festnetznummern mit Vorsicht zu interpretieren. Valide interpretierbar ist aber der zeitliche Vergleich der regionalen Anruferverteilung sein, sodass aus den Daten Verschiebungen von regionalen Anruferschwerpunkten evident werden.

Für die inhaltliche Analyse der Holine-Daten müssen die ACD- und die Sabio-Daten verknüpft werden. Da die Anruf-Id in den Sabio-Daten nicht vorliegt, erfolgt dies über Datum und Zeitpunkt des Anrufes und den Namen des Telefonagenten. Konkret wird ein Sabio-Datensatz einem aggregierten ACD-Datensatz zugeordnet, wenn Datum und Telefonagent übereinstimmen und wenn die Uhrzeit des Sabio-Datensatzes zwischen der Anfangs- und Endzeit des aggregierten ACD-Datensatzes liegt. Im Datensatz wird angegeben, wie viele Sabio-Datensätze der Anruf-Id zugeordnet wurden.

Thematisch wird aufgrund der zugeordneten Sabio-Datensätze für einen ACD-Datensatz festgehalten, welche Themen besprochen wurden. Dabei werden auf Basis der Verschlagwortung („Titel-Ast-Thema-Nachrucht“) folgende Themen mit 1-0-Variablen betrachtet:

* Thema Test: Sabio-Datensätze, in deren Verschlagwortung die Zeichenfolge „test“ vorkommt, also insbesondere mit den Wörtern „Testen“, „Testzentrum“, „PCR-Test“ usw.
* Thema Impfen: Sabio-Datensätze, in deren Verschlagwortung die Zeichenfolge „impf“ vorkommt, also insbesondere mit den Wörtern „Impfen“, „Impfzentrum“, „Impfung“, „Impfnachweis“ oder „Impfstoff“.
* Thema Reise: Sabio-Datensätze, in deren Verschlagwortung die Zeichenfolge „reise“ vorkommt, also insbesondere mit den Wörtern „Reiseregeln“, „Ausreise“, „Einreise“, „Verreisen“.
* Thema Quarantäne: Sabio-Datensätze, in deren Verschlagwortung die Zeichenfolge „quarantäne“ vorkommt, also inbesondere mit den Wörtern „Quarantäneregeln“, „Quarantänedauer“ oder „Corona-Quarantäne“.
* Thema App: Sabio-Datensätze, in deren Verschlagwortung die Zeichenfolge „app“ vorkommt, also insbesondere die Wörter „Corona-App“ und „Warnapp“.
* Thema Inzident: Sabio-Datensätze, in deren Verschlagwortung die Zeichenfolge „inzidenz“ auftritt, also insbesondere die Wörter „Inzidenzwert“, „Inzidenzen“ oder „7-Tage-Inzidenz“.
* Thema Bundesland: Sabio-Datensätze, in deren Verschlagwortung die Zeichenfolge „bundesland“ oder eines der Bundesländer wie etwa „Bayern“, „Berlin“, „Hessen“ usw. auftritt. Damit wird lediglich festgehalten, ob länderspezifische Themen bei einem Anruf eine Rolle gespielt haben, aber nicht, um welches Bundesland es dabei konkret ging.

Zu beachten ist, dass auf diese Wiese nicht jede Verschlagwortung korrekt zugeordnet wird, also etwa das Wort „Ausland“ nicht zum Thema Reise. Die weitaus meisten Verschlagwortungen werden allerdings korrekt erkannt. Eine Verschlagwortung kann dabei auch mehreren Themen zugeordnet werden.

Insgesamt wird nur abgebildet, ob ein Thema bei einem Anruf eine Rolle spielte oder nicht, aber nicht die Intensität, mit der das Thema besprochen wurde. Kommt ein Thema in mehreren Sabio-Datensätzen vor, die einer Anruf-Id zugeordnet wurde, so bleibt dies außer Acht, die entsprechende Themenvariable wird auf „1“ gesetzt und bildet nicht etwa die Zahl der Datensätze ab, in denen das Thema vorkam.

Konnten einem ACD-Datensatz keine Sabio-Datensätze zugeordnet werden, so steht die Zählvariable für die zugeordneten Sabio-Datensätze auf „0“ und die Themenvariablen sind alle fehlend („NA“).

**Inhalt des publizierten Datensatzes**

Die Daten sind auf täglicher Basis auf Kreisebene bzw. Mobilfunk-Nummernkreis aggregiert.

Der publizierte Datensatz enthält die Telefonhotline-Daten eines Tages in Form einer csv-Datei. Hierin sind folgende Datenfelder enthalten:

* Datum der Telefonanrufe im Format tt/mm/yyyy
* Anrufquelle: hiermit ist der Kreis, in dem der Anrufer wohnt, oder der Mobilfunk-Nummernkreis gemeint.

Die Einträge sind für Festnetzanrufe wie folgt aufgebaut: „FN-kgs-*xxxxx*“. Dabei steht die 5-stellige Zahl „*xxxxx*“ für den 5-stelligen Kreisgemeindeschlüssel (d.h. auf Kreisebene, die beiden ersten Stellen davon sind „01“-„16“ und beziehen sich auf das Bundesland. In wenigen Einzelfällen kann es vorkommen, dass kein Kreis zugeordnet werden kann, dann fehlt die Ziffernfolge „*xxxxx*“.

Die Einträge für Mobilfunkanrufe sind wie folgt aufgebaut: „MF-*Anbieter*“; dabei steht „*Anbieter*“ für „O2-Telefonica“, „Telekom“ oder „Vodafone“. Auch hier kann es Einzelfälle geben, wo ein Nummernkreis nicht zugeordnet werden kann, dann fehlt die Angabe „*Anbieter*“.

* Kreis: In diesem Feld steht der Kreis im Klartext. Für Mobilfunk-Anrufe ist dieses Feld leer.
* Anzahl\_Anrufe: Anzahl der in der Hotline vorhandenen Anrufe.
* Anzahl\_Verbindung\_zu\_Agent: Anzahl der Anrufe, bei denen eine Verbindung zu einem Agenten hergestellt wurde.
* Thema\_*Themenbereich*: Diese Felder geben an, bei wie vielen Anrufen auf Basis der oben beschriebenen Kodierung über einen der Themenbereiche „Testen“, „Reise“, „Impfen“, „Quarantäne“, „Bundesländer“, „App“ und „Inzidenz“ gesprochen wurde.