Master Biomedical Informatics and Data Science

Hausarbeit
Modul Forschungsdatenmanagement
WS 2020/21

FAIRE Publikation eines Beispieldatensatzes im Projekt TREAT COVID-19

Referentin: Prof. Dr.-Ing. Dagmar Waltemath

Vorgelegt von Dr. med. Oliver Hölsken
Attilastr. 90
12247 Berlin

ORCID: 0000-0001-6086-9275

Abgabetermin: 07.04.2021

Inhaltsverzeichnis

Tabellenverzeichnis

Abbildungsverzeichnis

1 Einleitung

1.1 Klinische Fragestellung

Es soll die Wirksamkeit verschiedener Medikamente auf die Behandlung von COVID-19 Patient:innen anhand eines synthetischen Datensatzes analysiert werden. Dabei soll insbesondere der Einfluss von Geschlecht und Ethnie auf die Wirksamkeit der Präparate untersucht werden.

1.2 Datensatz

Der zu bearbeitende Datensatz stammt von einem synthetischen Patienten-Datensatz-Tool namens Synthea™ ab. Er wurde durch die Dozent:innen des Kurses Forschungsdatenmanagement modifziert. Der Datensatz ist über den FAIRDOMHub abrufbar. Es wurden vier verschiende Datensätze in einem zusammgefügt:

Datensatz	Beschreibung
Patients	Demographische Daten
Observations	Messwerte (Vital- und Laborparameter)
Conditions	Diagnosen und Beschreibungen des Patientenstatus
Medications	Medikation

Tabelle 1: Datensatz-Beschreibung

Die Daten sind ursprünglich im FHIR HL7® Standard für elektronische Gesundheitsdaten abgelegt worden. Dabei sind die Zeileneinträge (wahrscheinlich) Werte von einzelnen Patienten und die Spalten Messwerte und Erhebungen sowie demographische Parameter wie der Wohnort. Es werden auch Diagnosen und Medikamente angegeben, auch unter verwendung standardisierter Codes (s. *HL7 Version 3 Standard: Core Principles and Properties* und *Synthea Dokumentation* für weiterführende Informationen).

2 Hauptteil

2.1 FAIR-Kriterien

Die FAIR Data Prinzipien beschreiben verschiedene Qualitätsdimensionen eines Datensatzes.FAIR ist ein Akronym für Findable, Accessible, Interoperable und Reproducible. Die FAIR Data Prinzipien werden sehr gut in *dieser Übersicht* zusammgenfasst

2.1.1 F: Findable, Auffindbarkeit

Das Kriterium F1 beschreibt die Zuweisung eines persistent Identifier (PID), z.B. ein Digital Object Identifier (DOI). Dies ermöglicht die einfache Auffindbarkeit und Zitierbarkeit. Der ursprüngliche Datensatz ist online (`https://synthea.mitre.org/downloads`) abrufbar und Teil einer Veröffentlichung, siehe Walonoski et al. (2020). Die DOI lautet: `https://doi.org/10.1016/j.ibmed.2020.100007`. Anzumerken ist, dass dem im Rahmen des Kurses modifizierten Datensatz keine DOI zugewiesen ist. Der Datensatz wurde daher als ***Harvard Dataverse*** hochgeladen, sodass eine eindeutige DOI zugewiesen wurde. Diese lautet: 10.7910/DVN/6IJZDG und ist (nach einem Login auf der Internseite), hier abrufbar. Auf das Harvard Dataverse wird näher im Rahmen des Abschnitts **Datenpublikation** eingegangen.

2.1.2 A: Accessible, Zugänglichkeit

Das Kriterium A1 beschreibt die Abrufbarkeit der (Meta)Daten über ein standardisiertes Kommunikationsprotokoll. Dies ist im Rahmen des HTTPS für den ursprünglichen Datensatz erfüllt, und zwar offen und kostenlos (A1.1). Der modifizierte Datensatz, der im Kurs verwendet werden soll, kann nur nach einer Freigabe durch die Projektleiterin Prof. Dr.-Ing. Dagmar Waltemath eingesehen werden (A1.2). Der Datensatz wurde auf das Harvard Dataverse gestellt und kann hier direkt über den folgenden URL Code ohne Zugangsbeschränkung heruntergeladen werden:

```
https://dataverse.harvard.edu/api/access/datafile/4496490
```

2.1.3 I: Interoperable, Interoperabilität

Das Kriterium I3 beschreibt die Referenzierbarkeit des Datensatzes zu einer Metadadensatz-Datei. Dies ist bei dem vorliegenden Datetensatz noch nicht erfolgt, somit fällt es schwer, nachzuvollziehen, was genau z.B. CODES in den einzelnen Spalten bedeutet. Es wurde eine umfangreicher Metadatensatz angefertigt, der Bestandteil des Harvard Dataverses ist und unter *Sektion 2.3. Metadaten* genauer beschrieben wird.

Zur Erhöhung der Maschinenlesbarkeit wurden folgende Änderungen im Programm Microsoft Excel vorgenommen (Programm Microsoft® Excel® für Microsoft MSO (16.0.13801.20288) 64-Bit):

1. Die Datei wurde im *.csv* Format abgespeichert.
2. Alle Spalten wurden eindeutig benannt, da es Duplikate gab. Z.B. Gab es sowohl bei den *OBSERVATIONS* als auch bei den *CONDITIONS* die Spalte *DESCRIPTION*. Daher wurden den Spaltennamen die Kürzel *OBS*, *CON* und *MED* vorangestellt.
3. Die Säule *PATIENT_ID* wurde gelöscht, da sie nicht relevant für den Datensatz ist (nach Hinweis der Dozent:innen).
4. Die Einträge in den spalten MED_CODE, MED_DISPENSES und MED_REASONCODE wurden automatisch in das Format float64 (also eine Kommazahl) durch die Python-Funktion $pd.read_csv$ überführt.

2.1.4 R: Reproducible, Wiederverwendbarkeit

Das Kriterium R1.2. beschreibt detaillierte Provenienz-Informationen eines Datensatzes. Der vorliegende Datensatz enthält keine Informationen über die Herkunft (Urheberschaft) oder die (maschinelle) Generierung. Über das Metadaten-Fenster des Harvard Dataverse wurde der Reiter *Provenance* bearbeitet. Hier wurden beispielsweise die Software (Synthea™, v.2.7.0), die Orginalpublikation und das zugehörige Github Repositorium der Entwickler:innen referenziert.

Der neue Datensatz wurde als Version *V1* im FAIRDOMHub hochgeladen.

2.2 Datenqualität

Zur Bewertung der Datenquatlität wurde die *3x3 Data-Quality-Assessment Matrix* der Oregon Heatlh Science University (OHSU), Department of Medical Informatics and Clinical Epidemiology, verwendet (***Website***). Bei dieser Matrix werden die drei "Data-quality constructs": *Complete*, *Correct* und *Current* gegenüber den drei Data dimensions: *Patients*, *variables* und *time* evaluiert (nach Weiskopf et al. (2017)), s. Abbildung 1. Innerhalb der Zellen stehen *operationalized constructs*.

	A: COMPLETE	B: CORRECT	C: CURRENT
1: PATIENTS	**1A** There are sufficient data points for each patient.	**1B** The distribution of values is plausible across patients.	**1C** All data were recorded during the timeframe of interest.
2: VARIABLES	**2A** There are sufficient data points for each variable.	**2B** There is concordance between variables.	**2C** Variables were recorded in the desired order.
3: TIME	**3A** There are sufficient data points for each time.	**3B** The progression of data over time is plausible.	**3C** Data were recorded with the desired regularity over time.

Abbildung 1: DQA-Matrix nach Weiskopf et al.
 `https://doi.org/10.5334/egems.218`

2.2.1 DQA Scope Fragebogen

Als Tool zu Beurteilung der Datenqualität, werden eine Reihe von *Ja/Nein*-Fragen zu den Daten-Dimensionen und zum Study-Design beantwortet (*Scope Identification Questions*). Das Ziel des Fragebogens ist es, die operationalized constructs zu erfassen und dazugehörige Empfehlungen für den erhobenen/zu erhebenden Datensatz zu erhalten. In diesem Abschnitt werden die Fragen und die gegebenen Antworten aufgeführt (Tabelle 2).

Frage	Antwort
Phase 1 - Identify Scope	
Does your study involve more than one patient?	Ja
Does your study involve more than one variable?	Ja
Does your study require information from more than one point in time for each patient?	Nein
Phase 2 - Verify Responses	
Does your study involve looking at more than one variable for multiple patients at one point in time?	Ja
Phase 3 - Assess Needs	
Do you have a specific time frame(s) of interest? E.g., a range of dates or a specific season?	Ja
Do you require or expect that variables be recorded in a certain order? E.g., diagnosis after laboratory result?	Nein

Tabelle 2: Scope Identification Questions zum Datensatz

2.2.2 Ergebnis

Das Ergebnis der Fragen ist in Abbildung 2 dokumentiert. Die für den Datensatz relevanten *Scope Identification Questions* erscheinen gelb markiert. Die geplante Studie wäre laut klinischer Fragestellung vom Design her im besten Fall eine Randomisierte kontrollierte Studie (RCT). Hierbei werden zwei zufällig ausgewählte Gruppen gebildet, wobei eine das zu untersuchende Präparat und die andere ein Placebo erhält. Es wird genau definiert, welche Endpunkte analysiert werden (Anteil der Erkrankten bzw. Verstorbenen in beiden Gruppen). Daraus kann die Wirksamkeit berechnet werden.

Bei dem Datensatz handelt es sich um eine Erhebung der Gesundheitsdaten mehrerer Patienten, von denen ein Teil SARS-CoV-2 positiv ist. Es werden mehrere Variablen erhoben (Geschlecht, Alter, Ethnie, Medikation). Ein molekulardiagnostisches Testergebnis (SARS-CoV-2 RNA) liegt nicht bei allen Patienten vor.

Eine Gruppenzuweisung ist nicht erkennbar (Zu testendes Medikament erhalten oder nicht). Es handelt sich somit wahrscheinlich um eine Deskriptive Studie, keine Interventionsstudie. Messdaten werden nicht wiederholt aufgenommen, verschiedene Messdaten stehen aber ggf. in einem zeitlichen und kausalen Zusammenhang. Der zeitliche Zusammenhang und der zeitliche Rahmen, in denen die Messwerte erhoben worden sind, müssen überprüft werden.

	A: Complete	B: Correct	C: Current
1: Patients	Are there sufficient data points for each patient?	Is the distribution of values across patients plausible?	Were all data recorded during the timeframe of interest?
2: Variables	Are there sufficient data points for each variable?	Is there concordance between variables?	Were variables recorded in the desired order?
3: Time	Are there sufficient data points for each time?	Is the progression of data over time plausible?	Were data recorded with the desired regularity over time?

Abbildung 2: Ergebnis des DQA Fragebogens
3x3 DQA Version 1.0, developed and written by Nicole Weiskopf and Chunhua Weng, is licensed under a Creative Commons Attribution-ShareAlike 4.0 International License. `https://apps.ohsu.edu/medical-informatics-clinical-epidemiology/data-quality-assessment/`

2.2.3 Berechnung des Qualitätsscores

Die Qualitätsscores für die beidenoperationalized constructs 2A und 2B wurden berechnet.

2A beschreibt die Fragestellung, ob ausreichend Datenpunkte für relevante Variabeln verfügbar sind. Dies wurde für die Beispiele SARS-CoV-2 Nachweis *SARS-CoV-2 RNA Pnl Resp NAA+probe* und die Verteilung des Merkmals Ethnie (*RACE*) exemplarisch durchgeführt (s. Jupyter-Notebook *210407_Implementierung_TREAT COVID-19.ipynb*, Abschnitt *Abfragen zur Datenqualität* für den Quellcode)

1. Bei 6 von 495 (1,2%) Patient:innen wurde eine SARS-CoV-2-PCR durchgeführt, von diesen waren wiederum 100% positiv. Für die anderen 490 Patienten liegen keine PCR Ergebnisse vor.

2. Die folgende Verteilung des Merkmals *RACE* finden sich in Tabelle 3 und Abbildung 3.

Bei den Werten *w* und *Alien* liegen sehr wahrscheinlich inkorrekte Daten vor (s. Verbesserungsvorschläge).

2B beschreibt die Fragestellung, ob es eine Konkordanz zwischen Variabeln gibt. Der Datensatz wurde exemplarisch auf korrekte bzw. inkorrekte Kombination zweier Variabeln untersucht (s. Jupyter-Notebook für den Quellcode)

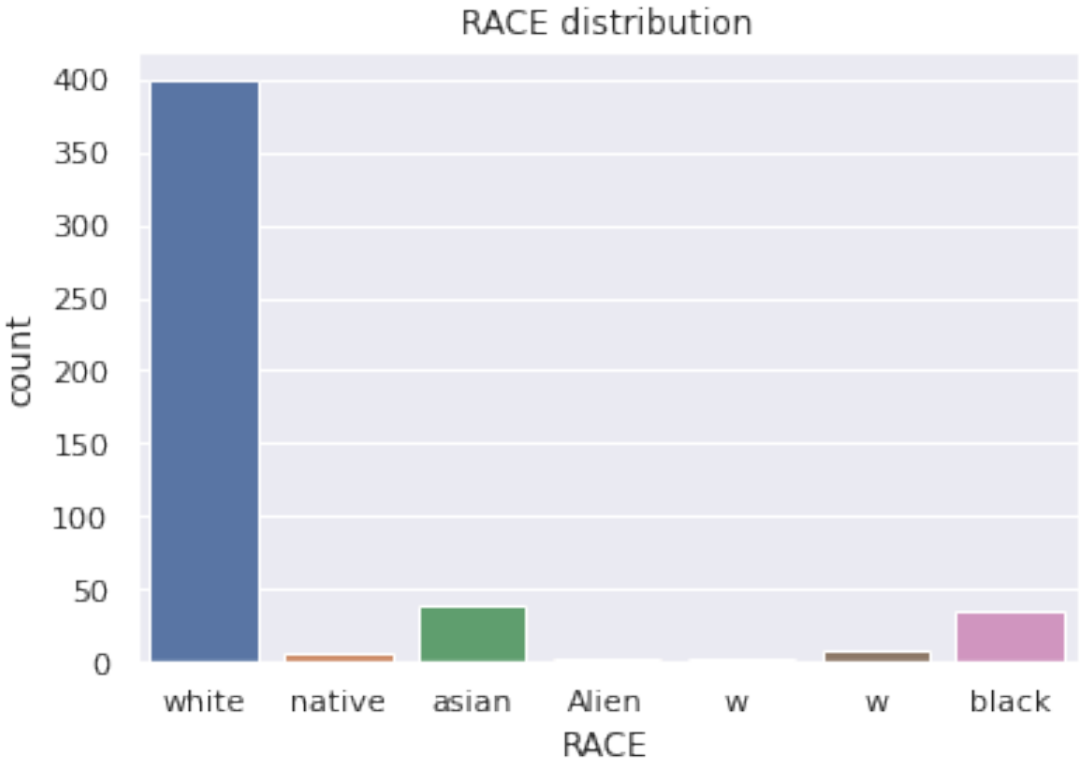

Abbildung 3: Verteilung des Merkmals Ethnie bei allen Patient:innen

RACE	Anzahl n	Prozent (%)
white	399	79,8
asian	39	7,9
w	12	2,4
native	7	1,4
Alien	3	0,6

Tabelle 3: RACE distribution im Datensatz, n = 495

1. Das Todesdatum muss nach dem Startdatum einer Medikation liegen (sonst wird eine Toter behandelt). Die Python-basierte Analyse des Datensatzes ergab, dass insgesamt 70 Patienten verstorben sind. Bei 28 dieser Patienten lag das Sterbedatum *vor* dem Beginn der Medikation, somit handelt es sich hierbei sehr wahrscheinlich um eine inkorrekte Kombination.

2. Die Condition *Normal pregnancy* muss mit dem Geschlecht *F* einhergehen. Die Python-basierte Analyse des Datensatzes ergab, dass insgesamt 50 mal das Merkmal *Normal pregnancy* vorkam. Allerdings hatten 22 das Merkmal *M* bei der Variable *SEX*, was 44% der Werte entspricht (s. Abbhildung 3).

Somit handelt es sich hierbei sehr wahrscheinlich um eine inkorrekte Kombination.

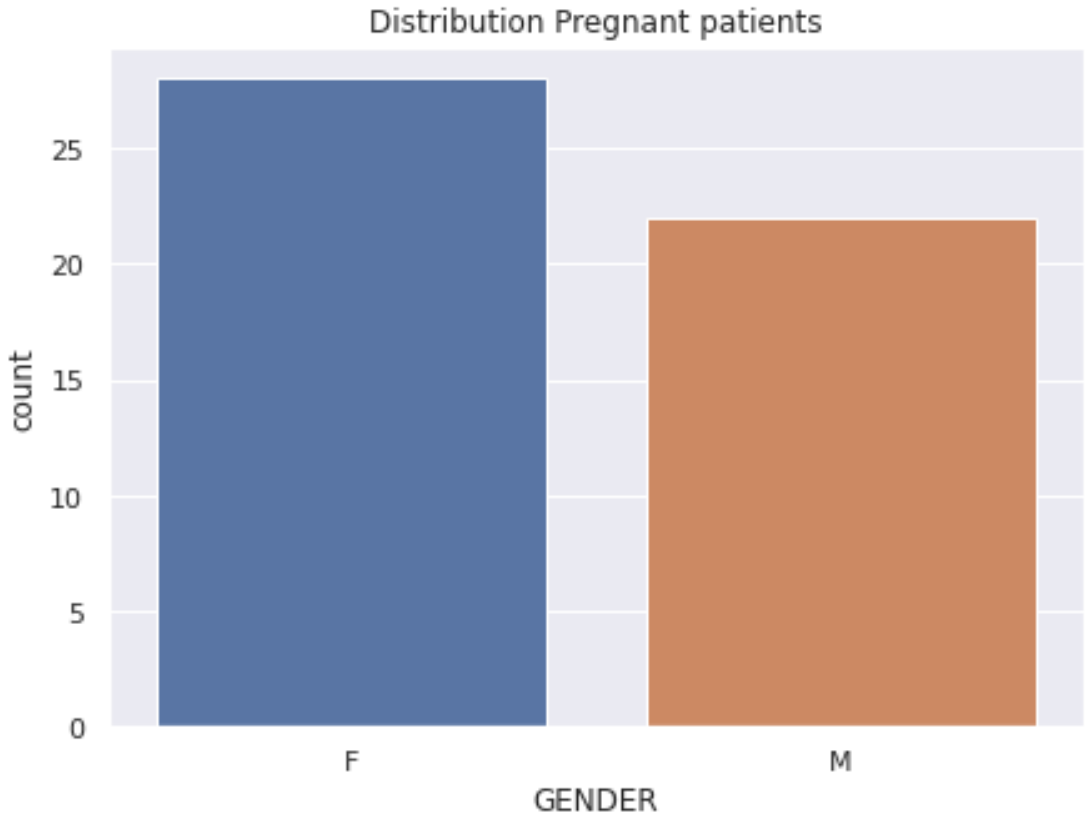

Abbildung 4: Verteilung des Merkmals Geschlecht bei schwangeren Patienten

2.2.4 Verbesserungsvorschläge

1. Entfernung von Messdaten mit irregulären Werten

   Beispiel: Im Merkmal Ethnie/*RACE* wurden Einträge mit den Merkmalen *Alien* und *w* entfernt.

2. Entfernung von Messdaten mit irregulären Merkmalskombinationen

   Einträge mit der Kombinaton des Wertes *Normal Pregnancy* in der Spalte CON_DESCRIPTION mit dem Wert *M* für das Merkmal *SEX* wurden aus dem Datensatz entfernt.

3. Verstorbene Patienten

(a) Verstorbene Patient:innen, die Medikamente nach dem Todesdatum erhalten haben, wurden aus dem Datensatz entfernt. Dafür wurden die Datumsformate der Spalten *DEATHDATE* und *MEDSTART* in das maschinenlesbare Format YYYY-MM-DD überführt.

(b) Außerdem wurden verstorbene Patient:innen aus dem Datensatz entfernt, die vor dem Jahr 2019 verstorben sind ($n = 19$) (bei der Annahme, dass der Studienzeitraum nicht älter als 2 Jahre ist).

4. Fehlender SARS-CoV-2 Nachweis (keine Implementierung)

   Es ist schwierig, den Datensatz hier zu verbessern. Es können natürlich keine zusätzlichen Testdaten eingetragen werden. Es muss allerdings in der Methodik klar darauf hingewiesen werden, dass nur für einen kleinen Teil der Patient:innen im Datensatz ein SARS-CoV-2-Nachweis vorliegt. Dies ist wichtig für die Nachnutzbarkeit.

Der neue Datensatz wurde als Version *V2* im FAIRDOMHub hochgeladen.

2.3 Metadaten

Folgende 4 Metadaten wurden auf verschiedenen Ebenen hinzugefügt

Als Metadaten-Standard wurde der Investigation Study Assay Tabular (ISA-Tab) `https://isa-specs.readthedocs.io/en/latest/isatab.html` gewählt, dieser ist im FAIRDOMHub implementiert ist. ISA ist eine Abkürzung für **I**nvestigation (Kontext des Projektes), **S**tudy (Eine (klinische) Studie) und **A**ssay (Messwert) und hat das Ziel, eine umfassende und standardisierte Form der Metadaten-Erfassung zu ermöglichen/fördern. Dieser Metadaten-Standard ist weit verbreitet in der Biomedizin.

Einer der großen Vorteile am ISA-Tab Metadaten-Standard ist, dass neben der der GUI-Seite FAIRDOMHub auch einige Software tools diesen Standard erzeugen bzw. lesen können, so z.B. Python. Eine Übersicht ist hier verlinkt.

Ein grober Überblick über die verschiedenen Organisationsebenen erfolgt in der unten stehenden Aufführung. Nähere Informaitonen sind in der ISA-Struktur des FAIRDOMHubs abgespeichert.

1. **Investigation**: COVID-19 and Synthetic Health Records

(a) **Study**: Daten und Metadaten zur Hausarbeit

i. **Assay**:

A. Datensätze in verschiedenen Versionierung

B. 210405_Metadatenschema_Datensatz.v1.xlsx

C. 210407_Implementierung_TREAT COVID-19.ipynb

Folgende weitere 4 Metadaten(typen) wurden zum Datensatz hinzugefügt:

1. Die Merkmale mit *DATE* wurde dem ISO 8601 Standard angepasst (von DD.MM.YYYY zu YYYY-MM-DD).

   (a) Erhöhung der Maschinenlesbarkeit

   (b) Kein Datenverlust bei Transformation zwischen Dateitypen

2. Eine deskriptives Metadatenschema inkl. Benennung der Spaltenüberschriften und eine Versionierung des Datensatzes wurde angelegt (*s. 210405_MetadatenschemaDatensatz.v1.xlsx*)

   (a) Die Nachvollziehbarkeit, um welche Datentypen es sich handelt, wird erhöht. File-level Meta-Daten erhöhen die Nachnutzbarkeit und somit die Wahrscheinlichkeit, publiziert zu werden.

   (b) Die Reproduzierbarkeit wird erhöht, durch eine umfassende Beschreibung der Veränderungen zwischen den einzelnern Versionen.

3. Eine Beschreibung über die Data Provenance wurde auf der Ebene der FAIRDOMHub - Study **Daten und Metadaten zur Hausarbeit** Ebene hinzugefügt

   (a) Die Herkunft inklusive die Urherberschaft werden genauer beschrieben, um zum einen legale Streitigkeiten zu vermeiden aber auch um die Sichtbarkeit der Autoren der Erstpublikation zu erhöhen. Die Verlinkung zum Github Repositorium ist eine hilfreiche Quelle für Anwender und Nutzer des Synthea™ Tools.

      i. Copyright und Lizenzrechte liegen bei der MITRE Kooperation

      ii. Verlinkung zum Originaldatensatz

      iii. Verlinkung zum Originalpaper und Assoziaiton zum Daten-Item auf FAIDROMhub (Submitter Prof. Dr.-Ing. Dagmar Waltemath)

      iv. Verlinkung zum Synthea™ Github-Repository

4. Folgende MeSH Terms wurden auf der Ebene des Datasets als Tags hinzugefügt. Dies erhöht die Auffindbarkeit der Daten.
   (a) SARS-CoV-2
   (b) COVID-19
   (c) COVID-19 Nucleic Acid Testing
5. Der neue Datensatz wurde als Version *V3* im FAIRDOMHub hochgeladen.
6. Das deskriptive Metadatenschema wurde als Datei *210405_MetadatenschemaDatensatz.v1.xlsx* hochgeladen.

2.4 Datenaufbereitung

Data cleaning oder Data cleansing ist ein Überbegriff für Methoden, die zur Identifikation und Behebung von fehlerhaften Daten verwendet werden. Nach Müller (2003) lautet die Definition:

„*Data cleaning is the process of detecting and correcting (or removing) corrupt or inaccurate records from a record set, table, or database and refers to identifying incomplete, incorrect, inaccurate or irrelevant parts of the data and then replacing, modifying, or deleting the dirty or coarse data.*“

Im Rahmen der Hausarbeit wurde ein Data cleaning Skript in Python erstellt und auf den Datensatz angewendet (Abschnitt **Datenaufarbeitung** im Jupyter Notebook *210407_Hausarbeit_COVID-19.ipynb*).

Folgende Schwachstellen konnten identifiziert und bereinigt werden.

2.4.1 Fehlende Daten

In einer graphischen Darstellungsweise Heatmap (Abbildung 4) kann man sich einen Überblick über die Vollständigkeit des Datensatzes verschaffen.

Dies lässt sich auch numerisch darstellen:

1. Fehlende Daten (%)
   (a) BIRTHDATE - 1%
   (b) DEATHDATE - 97%
   (c) MARITAL - 35%

...

(d) CON_STOP - 38%

...

(e) MED_REASONDESCRIPTION - 16%

Es fällt auf, dass die Vollständigkeit der verschiedenen Variabeln des Datensatzes sehr unterschiedlich ist. Dies ist prinzipiell kein Problem, weil z.B. Das *DEATHDATE* nicht bei allen Patienten vorkommen muss oder bei dne *OBSERVATIONS* auch qualitative Angaben erlaubt sind und somit die *OBS_UNITS* Spalte durchaus keinen Eintrag haben kann.

Andere fehlende Werte sind hingegen potentiell problematisch, z.B. kann durch fehlende Einträge bei *MED_STOP* keine Behandlungsdauer berechnet werden. Datenpunkte von Patienten:Innen mit fehlenden kritischen Einträgen können jederzeit aus dem Datensatz herausgenommen werden.

Es fiel auf, dass drei Zeilen des Datensatzes komplett ohne Inhalt waren. Diese wurden gelöscht.

2.4.2 Duplikate

Duplizierte Werte stellen ebenfalls nicht immer ein Problem vor, da z.B. Geburtsdaten und Messwerte mehrfach vorkommen können. Allerdings sind Zeilen, bei denen vielen oder gar alle Einträgen gleich sind, problematisch und sollten gelöscht werden.

Es fiel auf, dass zwei Zeilen des Datensatzes doppelt vorhanden waren, also exakt die gleichen Einträge in allen Spalten hatten. Jeweils eine der duplizieren Zeilen wurde entfernt.

2.4.3 Irreguläre Daten und Extremwerte

Folgende Änderungen wurden vorgenommen:

1. Spalte *BIRTHPLACE*: Austausch von irregulären Schreibweisen, z.B. wurde *Hanoi H† fê¥ng VN* in zwei Zeilen durch *Hanoi Hanoi VN* ersetzt.

2. Mittels Python-Funktion *.describe* wurden die Spalten mit numerischen Werten auf Extremwerte untersucht. Dies zeigte, dass in der Spalte *MED_DISPENSES*,

also die Verordnungsanzahl, einen sehr hohen Wert (*583*) gab. Das arithmetische Mittel lag bei 18.3. Der hohe Wert wurde allerdings nicht aus dem Datensatz gelöscht, ist aber wahrscheinlich inkorrekter.

3. Drei Einträge in der Spalte CON_CODE hatten sehr lange SNOMED CT Codes. Eine Überprüfung ergab, dass diese Werte nicht mit der CON_DESCRIPTION übereinstimmte. Der korrekte Code wurde daraufhin eingefügt.

2.4.4 Datenformate

1. Ähnlich wie in de Abschnitten Datenqualität und Metadaten für das Datumsformat beschrieben, müssen die Formate von Daten bestimmten Formen und Standards entsprechen, dies erhöht die Maschinenlesbarkeit und ermöglicht die leichtere Interoperabilität. Alle betreffenden Spalten wurdem dem ISO 8601 Standard angepasst (YYYY-MM-DD) (bereits in V3).

2. Ein wichtiger Aspekt für die Maschinenlesbarkeit von Python ist die Groß- und Kleinschreibung. Diese wurde am Beispiel des *BIRTHPLACE* einheitlich auf Kleinschreibung angepasst. Außerdem wurden Leerzeichen in den Zeilen gelöscht.

3. Die Spalten MARITAL, RACE und GENDER enthalten kategorische Werte, sind aber als Python *object* abgespeichert. Diese können in kategorische Werte über die *.astype* Funktion in Python zu kategorischen Werten überführt werden. Dies erhöht die Geschwindigkeit von Rechenschritten und erleichtert die Daten-Manipulation für weitere Analyseschritte.

4. Die Spalten CON_CODE und MED_CODE enthalten Codes für bestimmte Ontologien. Diese Werte wurden im als Python *object int64* abgespeichert. Dies erleichtert die weitere Analyse mittels Python (ist aber nicht unbedingt erforderlich :)

5. Die Spalten CON_START und CON_STOP wurden in ein Python *datetime* Objekt überführt.

Der neue Datensatz wurde als Version *V4* im FAIRDOMHub hochgeladen.

2.5 Datenpublikation

Die Datenpublikation ist ein zentraler Prozess im Forschungs-Daten-Lebenszyklus. Sie ist von großer Bedeutung für die Auffindbarkeit, Archivierung und Nachnutzung der Daten.

2.5.1 Auswahl des Publikationsservers/Repositorium

Der Datensatz wurde auf folgenden Repositorien/Servern (teil)-veröffentlicht:

1. **Harvard Dataverse**
    (a) Ist ein weit verbreiteter Publikationsserver, er umfasst mehrere wissenschaftliche Domänen und fasst diese als *Subject* zusammen, z.B. *Medicine, Health, Life Sciences*
    (b) Harvard Dataverse hat darüber hinaus eine *COVID-19 Data Collection*, diese würde sich sehr gut zur Veröffentlichung des TREAT-COVID-19 Datasets eignen (Abbildung 6).
    (c) ermöglicht die Publikation von zitierfähigen Datensätzen mit der Erstellung einer DOI (s. Abschnitt 2.3 FAIR-Kriterien)
    (d) ermöglicht die **Langzeitarchivierung** (i.d.R. min. 10 Jahre)
    (e) erfasst umfangreiche **Metadaten** und erlaubt Verlinkungen zu Publikationen und GitHub-Repositorien
    (f) **Metadaten** werden den Dublin Core Kriterien entsprechend angelegt.
    (g) Für Dateien stehen jedem/jeder Forscher:in 2.5 GB Speicher und für Datensätze 1 TB Speicher zur Verfügung, es reicht die Registrierung per e-Mail.
    (h) Die Hierarchie-Ebene ist wie folgt:
        i. Dataverse (Sammlung von Datasets)
            A. Dataset 1
                A. Descriptive Metadata
                B. Data files
                    A. Data file-Metafile
                B. Documentation
                C. Code

2. **GitHub-Repository**

   (a) Vorteil ist die **Versionskontrolle**, da Änderungen sehr gut verfolgt werden können.

   (b) Ein weitere Vorteil ist die **Kollaborationsmöglichkeiten**. Man kann relativ leicht Änderungsvorschläge erhalten und implementieren.

   (c) GitHub ist sehr gut geeignet, um **Implementierungs-Tools** zu teilen und weiterzuentwickeln.

   (d) GitHub erkennt eine Vielzahl von Programmiersprachen und es erleichtert die Generierung von Packages.

   (e) Der Datensatz wurde im **Repository TREAT-COVID-19** abgelegt.

3. **Google Drive**

   (a) Dient als Backup-Speicher des Datensatzes und ist mit dem GitHub Repository verlinkt.

   (b) Zugang nach Anfrage durch Autor

   (c) Wird zur Verfügung gestellt über die Graduate School Rhein Neckar im Rahmen des Masters Biomedical Informatics and Data Science.

2.5.2 Zusammenfassung der Bedingung für Veröffentlichung

Der folgende **Zitationsstandard** wird angewandt. 5 sind menschenlesbar, wobei DOI und UNF maschinenlesbar sind.

1. author name(s)
2. year (date published in the Dataverse repository)
3. title
4. global persistent identifier: DOI or Handle
5. publisher (repository that published the dataset)
6. version number
7. universal numerical fingerprint (UNF): for tabular data

Universal numerical fingerprints (UNF) dienen der eindeutigen Signatur des semantischen Kontext eines digitalen Objektes, ohne den Inhalt zu übertragen. Der UNF Algorithmus ermöglicht das Format-unabhängige Speichern eines Dokumentes, was sehr bedeutend ist für die Interoperabilität (siehe auch Altman and King (2007)). Die UNF wird zusammen mit der DOI vergeben und beginnt mit *UNF* gefolgt von einer Ziffernfolge, z.B. UNF:3:DaYlT6QSX9r0D50ye+tXpA==.

2.5.3 Wahl der Lizenz

Vorweg ist anzumerken, dass eine Publikation unseres Datensatzes möglich ist, da der Urheber (MITRE Cooperation) definiert hat:

The data is free from cost, privacy, and security restrictions. It can be used without restriction for a variety of secondary uses in academia, research, industry, and government

Somit ist die Nachnutzung möglich, es muss nur die Zitation angegeben werden. Dies wurde in der ISA-Struktur auf FAIRDOMHub und in den Metadaten auf Harvard Dataverse getan.

Ein Nachteil des Harvard Dataverse ist, dass die Publikation von Datensätze standardmäßig die weltweite Befreiung von Urheberrechten vorsieht (**Creative Commons Zero v1.0 Universal**). Allerdings wird eindrücklich darauf hingewiesen, dass eine Namensnennung und Zitation im Rahmen der Guten wissenschaftlichen Praxis erfolgen soll bei der Weiterverwendung des Datensatzes.

Prinzipiell steht einer Veröffentlichug des Datensatzes under CC0-1.0 Lizenz nichts entgegen, da es sich um synthetische Patientendaten handelt, die keine sensiblen Daten enthalten. Für den Fall, dass sensible klinische oder personenbezogene Daten veröffentlicht werden sollen, sehen die Regularien des Harvard Dataverses vor, dass eine Veröffentlichung unter CC0 ausgeschlossen werden kann. Dann müssen spezielle Data Usage Agreement aufgesetzt werden.

Die Veröffentlichung des Datensatzes über das FAIRDOMHub wäre etwas restriktiver möglich, z.B. **CC BY 4.0 International Share Alike**, sodass eine Zitation des Autors und die Anzeige von Änderungen am Werk erfolgen müssen. Außerdem darf das Werk nur unter den gleichen Bedingungen weitergeben werden.

2.5.4 Wahl des Formats

Bei der Vorgabe eines Datei-Typs sind sowohl Harvard Dataverse als auch FAIRDOMHub relativ frei, so können z.B. auch .xlsx Dateien hochgeladen werden. Das Harvard Dataverse empfiehlt das Hochladen von Datafiles as .csv und ggf. die Überführung in ein bestimmtes Tabellenformant (s. hier. Das FAIRDOMHub bietet bestimmte assay-spezifische (z.B. 3' or Whole Gene Expression array) aber auch Master Templates an, diese sind modifizierbar:

This is the general format for all SEEK templates. Each spreadsheet has multiple sheets, defining experimental metadata, organims/samples, data and optional sheets for instrument descriptions and results. This template can be refined and/or extended to cater more specifically for different data types. `https://docs.seek4science.org/help/templates.html`

Die Verwendung von diesem Master Template hat allerdings den Nachteil, dass nur eine .xls Datei heruntergeladen werden kann. Außerdem ist es empfehlenswert, Metadaten und Datensatz getrennr voneinander aufzubewahren.

Literaturverzeichnis

Altman, M. and King, G. (2007). A proposed standard for the scholarly citation of quantitative data. *D-Lib Magazine*.

Müller, H. und Freytag, J. (2003). Problems, methods, and challenges in comprehensive data cleansing. Humboldt-Universität Berlin [Online; Abgerufen am 06.04.2021].

Walonoski, J., Klaus, S., Granger, E., Hall, D., Gregorowicz, A., Neyarapally, G., A, W., and Eastman, J. (2020). Synthea™ novel coronavirus (covid-19) model and synthetic data set. *Intelligence-Based Medicine*, 1(100007).

Weiskopf, N., Bakken, S., Hripcsak, G., and Weng, C. (2017). A data quality assessment guideline for electronic health record data reuse. *EGEMS*, 5(14).

Eigenständigkeitserklärung

Ich versichere, dass die vorstehende Arbeit von mir selbstständig ohne unerlaubte fremde Hilfe und ohne Benutzung anderer als der angegebenen Hilfsmittel angefertigt wurde, und dass ich alle Stellen, die wörtlich oder sinngemäß aus veröffentlichten oder unveröffentlichten Schriften entnommen sind, als solche gekennzeichnet habe.

Berlin, den 7. April 2021

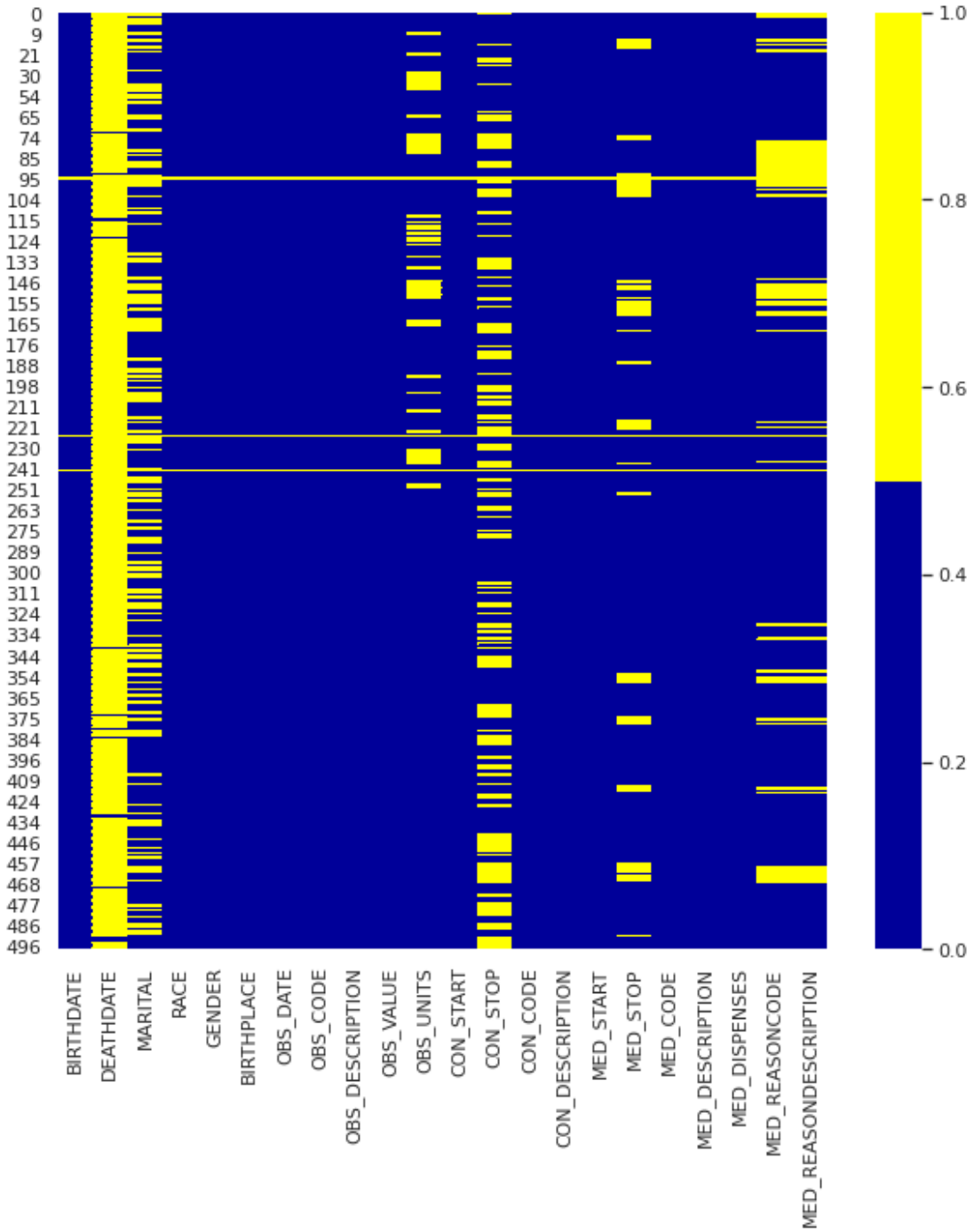

Abbildung 5: Heatmap der fehlenden Datenpunkt

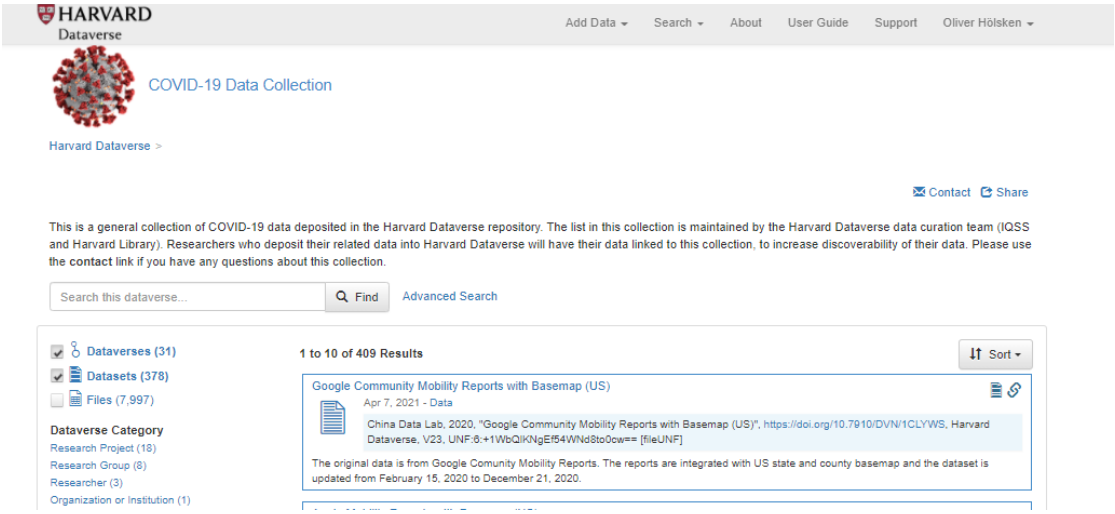

Abbildung 6: Screenshot der Harvard Dataverse COVID-19 Data Collection. `https://dataverse.harvard.edu/dataverse/covid19`