

FAIRE Publikation eines Beispieldatensatzes


Felix P. Hans ¹

¹Universitäts-Notfallzentrum, Freiburg, Germany

Hausarbeit Modul
Forschungsdatenmanagement *

Masterstudiengang
Biomedizinische Informatik und Data Science
April 2021

DOI:10.5281/zenodo.4667703

*Dozentin: Dagmar Waltemath, Professor of Medical Informatics 

Inhaltsverzeichnis

1	Abstract	3
2	Die FAIR-Prinzipien	4
3	Aufgabe 1 - Evaluation der FAIR-Kriterien	6
3.1	Aufgabenstellung	6
3.2	Antwort	6
3.2.1	Datensatzbeschreibung	7
3.2.2	Aufbereitung des Datensatzes	8
4	Aufgabe 2 - Datenqualität	12
4.1	Aufgabenstellung	12
4.2	Antwort - 3x3 DQA	12
4.3	Antwort - quality metrics	15
5	Aufgabe 3 - Meta-Daten	17
5.1	Antwort - Meta-Daten-Ergänzung	17
6	Aufgabe 4 - Datenaufbereitung	19
6.1	Antwort - Cleansing	19
7	Aufgabe 5 - Datenpublikation	20
7.1	Antwort - Publikation	20

1 Abstract

Der Umgang mit Forschungsdaten unterliegt im Lichte der zunehmenden Digitalisierung einem rasanten Wandel. Die Analyse hochdimensionaler Daten ist zunehmend auch in der medizinischen Forschung niederschwellig verfügbar und erfreut sich zunehmender Verbreitung. Mit steigender Komplexität der Daten wird zunehmend Wert auf ein transparentes Forschungsdatenmanagement gelegt. Dieses beinhaltet neben konkret definierten Governancestrukturen die sogenannten **FAIR**-Kriterien [1]. FAIRer Umgang mit Forschungsdaten ist wie folgt charakterisiert:

- **F**indable: Auffindbarkeit von Daten durch bestmögliche Beschreibung und Ablage in einem öffentlichen Repository.
- **A**ccessible: Höchstmögliche Zugänglichkeit zu Daten durch stringente Planung bei der Datenerhebung (z.B. Einwilligungen, Austauschprotokolle).
- **I**nteroperable: Benutzung offener Formate, gemeinsamer Metadatenstandards und konsistentem Vokabular für ein Höchstmaß an Interoperabilität.
- **R**eusable: Lizenzierung und Veröffentlichung von Daten und Metadaten zur bestmöglichen Weiternutzung.

Diese Hausarbeit befasst sich beispielhaft mit der FAIRen Aufbereitung eines Datensatzes und der entsprechenden Dokumentation.

2 Die FAIR-Prinzipien

Die FAIR-Prinzipien wurden formuliert, um den größtmöglichen Erkenntnisgewinn aus Forschungsdaten zu ziehen. Diese Erkenntnisse sollen menschlichen und maschinellen Verarbeitungsprozessen gleichermaßen zugänglich sein. Die interdisziplinäre Nutzung von Daten und Metadaten aus multiplen Quellen soll neue Erkenntnisse zu Tage fördern und nach Kräften wiederverwendbar und offen sein. Wesentliche Gründe für ein FAIRes Forschungsdatenmanagement sind

- die maximale Kraftentfaltung der erhobenen Forschungsdaten
- die Sichtbarkeit der Forschungsdaten zu verbessern und die Zitier-Fähigkeit zu erhöhen
- die Reproduzierbarkeit der eigenen Ergebnisse und die Verlässlichkeit der eigenen Forschung zu verbessern
- die Angleichung an internationale Forschungsstandards und -Ansätze
- die Steigerung der eigenen Sichtbarkeit zum vereinfachten Austausch und Zusammenarbeit mit anderen forschenden, Politik und Industrie
- die Ermöglichung neue, gemeinsame Forschungsfragestellungen überhaupt erst zu adressieren
- die Evolution der Forschung weiter voran zu treiben.

Die vier Dimensionen des FAIRen Datenmanagements gliedern sich wie folgt (modifiziert nach [2,3]):

1. Findable - Auffindbar

Die Auffindbarkeit von Forschungsdaten beinhaltet

- (a) die Verwendung eindeutiger und dauerhafter Identifikatoren (z.b. DOI oder Handle)
- (b) die Erstellung und Pflege bestmöglicher und voll-umfänglicher Metadaten zur Beschreibung und dem Verständnis des Datensatzes
- (c) die Registrierung und Indizierung dieser Metadaten in einem durchsuchbaren Verzeichnis
- (d) die Sicherstellung des Zugangs zu Daten über Repositorien oder Archive (lokal oder öffentlich)

2. Accessible - Zugänglich

Die Zugänglichkeit von Forschungsdaten beinhaltet

- (a) die Standardisierung von Daten- Austausch- und Zugriffsprotokollen und die Planung von höchstmöglicher Offenheit (*open, free, transparent data*) der Forschungsdaten. Auch sensible Daten können unter Lizenz nach Transformation zur Weiternutzung verfügbar gemacht werden. Auch die Nicht-Veröffentlichung von Daten kann in öffentlichen Metadaten-Repositorien transparent gestaltet werden.
- (b) den offenen, freien und universell implementierbaren Charakter dieser Austausch- und Zugriffsprotokolle
- (c) wo notwendig, die Authentifizierung und Rechteverwaltung innerhalb der Austausch- und Zugriffsprotokolle
- (d) die dauerhafte Verfügbarkeit der Metadaten auch bei nicht mehr vorhandenen Forschungsdaten

3. Interoperable - Interoperabel Die Interoperabilität von Forschungsdaten beinhaltet

- (a) eine formale, zugängliche, gemeinsam genutzte und breit anwendbare Sprache für (Meta)-Daten
 - (b) die Anwendung der FAIR-Prinzipien auch auf dieses Vokabular von (Meta)-Daten
 - (c) die Referenzierung von (Meta)-Daten aufeinander und die Nachvollziehbarkeit deren Abstammung
4. **Reusable - Wiederverwendbar** Die Wiederverwendbarkeit von Forschungsdaten beinhaltet
- (a) die detaillierte Beschreibung von (Meta)-Daten mit präzisen und relevanten Attributen zur sicheren Ermittlung des korrekten Kontextes für eine mögliche Wiederverwendung der Daten.
 - (b) die klare menschen- und maschinenverständliche Angabe von Nutzungslizenzen für (Meta)-Daten
 - (c) die Angabe von Provenienz-Informationen zu Entstehungszeitpunkt, -Kontext und erfolgten Transformationsschritten
 - (d) die Einhaltung fachgebietsrelevanten Community Standards

3 Aufgabe 1 - Evaluation der FAIR-Kriterien

3.1 Aufgabenstellung

Wählen Sie jeweils 1 FAIR-Kriterium aus jeder Gruppe (F1-R1.3) aus und argumentieren Sie, ob dieses Kriterium erfüllt ist oder nicht. Verbessern Sie den Datensatz bezüglich der gewählten Kriterien. Laden Sie den Datensatz mit Appendix V1 in den FAIROMHub hoch.

3.2 Antwort

Die vier ausgewählten Evaluations-Operatoren (nach [3]) sind:

1. F2. Daten werden mit umfangreichen Metadaten beschrieben.

Jeder Datensatz sollte mit ausführlichen Metadaten beschrieben werden: Diese Metadaten dokumentieren u.a., wie die Daten generiert wurden, wer sie erhoben / bearbeitet / publiziert hat und unter welchen Bedingungen (Lizenz) sie verwendet werden dürfen. Metadaten liefern somit den notwendigen Kontext für die richtige Interpretation von Forschungsdaten. Diese Informationen müssen ebenfalls maschinenlesbar sein.

2. A1.1 Das Protokoll ist offen, frei und universell implementierbar.

Um die Datennachnutzung zu maximieren und den Datenabruf zu erleichtern, sollte das Protokoll frei (kostenfrei) und offen (open source) sein und somit global implementierbar sein. Jede Nutzerin und jeder Nutzer mit einem Computer und einer Internetverbindung kann mindestens auf die Metadaten zugreifen. Das Repositorium nutzt ein offenes (kein proprietäres oder kommerzielles) Kommunikationsprotokoll (Beispiele: HTTP(S), FTP, SMTP).

3. I1. (Meta)-Daten nutzen eine formale, zugängliche, gemeinsam genutzte und breit anwendbare Sprache für die Wissensrepräsentation.

Menschen tauschen Informationen u.a. durch die Verwendung von gängigen Sprachen aus. Dies gilt ebenfalls für Computer. Daher sollten Daten auch für Maschinen in einer verständlichen Darstellung verfügbar sein. Wenn (Meta)Daten durchsucht werden, müssen die Computersysteme entscheiden können, ob der Inhalt der Datensätze vergleichbar ist. Für die Erstellung und Anwendung solcher Metadaten werden kontrollierte Vokabulare / Ontologien / Thesauri und ein klar definiertes Framework benötigt, z.B. im Sinne des Semantic Web.

4. R1.2. (Meta)Daten enthalten detaillierte Provenienz-Informationen.

Provenienz-Informationen geben an, wie die Daten generiert wurden, in welchen Kontext sie wiederverwendet werden können und wie zuverlässig sie sind. Damit sind sie für die Wiederverwendung notwendig und bilden ein wichtiges Kriterium bei der Validierung von Daten in wissenschaftlichen Datenbanken.

3.2.1 Datensatzbeschreibung

Der zur Verfügung gestellte Datensatz besteht aus den folgenden Spalten (-Überschriften):

1. Patients
 - (a) BIRTHDATE
 - (b) DEATHDATE
 - (c) MARITAL
 - (d) RACE
 - (e) GENDER
 - (f) BIRTHPLACE
2. Observations
 - (a) DATE
 - (b) PATIENT
 - (c) CODE
 - (d) DESCRIPTION
 - (e) VALUE
 - (f) UNITS
3. Conditions
 - (a) START
 - (b) STOP
 - (c) CODE
 - (d) DESCRIPTION
4. Medications
 - (a) START
 - (b) STOP
 - (c) CODE
 - (d) DESCRIPTION
 - (e) DISPENSES
 - (f) REASONCODE
 - (g) REASONDESCRIPTION

Die Tabelle liegt im *.xlsx-Format vor und wurde mutmaßlich in Microsoft Excel erstellt. Die Tabelle ist mit verschiedenen starken Strichen unterteilt, so dass naheliegend ist, dass diese Tabelle aus verschiedenen Einzeltabellen in einer grafischen Benutzeroberfläche zusammengefügt wurde. Es liegen keinerlei Metadaten zu den vorliegenden Forschungsdaten vor. Auf den ersten blick ist weder zu erkennen wer die Daten wann erhoben oder aggregiert hat, noch mit welchen Datentypen die einzelnen Felder standardmäßig beschickt werden sollen. In den Einträgen mit codesöder Einheitenliegen keine Angaben zum Referenzsystem vor. Es gibt prima vista viele leere Felder und unplausible Daten (z.B. Patients DEATHDATE '07.05.2016' mit Observation DATE '30.01.2020'). Zunächst musste die

Vermutung geprüft werden, ob die zusammengeführten Einzeltabellen überhaupt als zusammenhängende Tupel gewertet werden können, oder ob einzelne Faktentabellen daraus generiert werden müssen. Da aber bei der Zerlegung in einzelne Tabellen kein klares Mapping der Inhalte möglich wäre, wurde diese Option verworfen und der Datensatz als 'schmutzig' akzeptiert.

F2. Daten werden mit umfangreichen Metadaten beschrieben?

Da keine Metadaten vorliegen ist der Lehr-Datensatz unter dem F-Kriterium der Metadaten-Verfügbarkeit mangelhaft und damit unFAIR.

A1.1 Das Protokoll ist offen, frei und universell implementierbar?

Das Excel-Format ist proprietär, kann jedoch mittels Open-Source-Software wie z.b. OpenOffice kostenfrei manipuliert und konvertiert werden. Der Zugang zu den Ursprünglichen **Synthea-Quelldaten** ist über https frei möglich [4]. Die zu Lehrzwecken aggregierte Version liegt jedoch nur für Teilnehmende des Masterkurses im **FAIRDOMHub**. Schon hier zeigt sich die Schwierigkeit einen öffentlich nicht einsehbaren Datensatz in einem Dokument wie diesem hier zu analysieren und zu beschreiben. Eine Überprüfung ist für externen Leser daher zunächst nicht möglich und damit unFAIR.

I1. (Meta)-Daten nutzen eine formale, zugängliche, gemeinsam genutzte und breit anwendbare Sprache für die Wissensrepräsentation?

Die Metadaten im aggregierten Übungsdatensatz lassen sich nur über *reverse engineering* und über *best guesses* erstellen und beschreiben. Die Interoperabilität ist im vorliegenden Fall unFAIR.

R1.2. (Meta)Daten enthalten detaillierte Provenienz-Informationen?

Die Provenienz kann hier nur über das Layout und die Sichtung der Quelldaten bei **Synthea** erfolgen. Da somit nur Mutmaßungen über die erfolgten Prozessschritte zur Entstehung des Datensatzes möglich sind ist die Wiederverwendbarkeit unFAIR.

3.2.2 Aufbereitung des Datensatzes

Es erfolgte zunächst die Konvertierung in das CSV-Format mit OpenOffice. Zur vereinfachten Bearbeitung wurde die Spaltenüberschriften angepasst, damit bei der Manipulation keine mehrdeutigen Spaltennamen vorliegen (*unique constraint*). Hierzu erfolgte die händische Erweiterung der o.g. Spalten mit dem Präfix der Überschrift, so dass folgende Spaltennamen entstanden:

1. pt_BIRTHDATE
2. pt_DEATHDATE
3. pt_MARITAL
4. pt_RACE
5. pt_GENDER
6. pt_BIRTHPLACE
7. obs_DATE
8. obs_PATIENT

9. obs_CODE
10. obs_DESCRIPTION
11. obs_VALUE
12. obs_UNITS
13. cond_START
14. cond_STOP
15. cond_CODE
16. cond_DESCRIPTION
17. med_START
18. med_STOP
19. med_CODE
20. med_DESCRIPTION
21. med_DISPENSES
22. med_REASONCODE
23. med_REASONDESCRIPTION

Es erfolgte die Ablage der Datei im [FAIRDOMHub](#) ('Dataset_Covid_FPH.csv'). Die anschließenden Analysen und Manipulationen erfolgen in einem Jupyter Notebook.

Code 1: Einlesen des offenen CSV-Formates als dataframe

```
import numpy as np
import pandas as pd
DF = pd.read_csv('Dataset_Covid_FPH.csv', sep=";")
```

Aus den Ausgaben der Dataframe-Informationen und der Beurteilung der Daten auf Kohärenz kann eine erste Metadatentabelle angelegt werden. Der Datensatz enthält aktuell maximal 495 Einträge, die im Folgenden weiter überprüft werden.

Für die o.s. Deklinationen der FAIR-Kriterien ergeben sich nach der ersten Überarbeitung nun folgende Änderungen:

F2. Daten werden mit umfangreichen Metadaten beschrieben?

Eine erste Version der Metadaten ist erstellt, die Datentypen sind bekannt und die Anzahl der vorhandenen Datensets.

A1.1 Das Protokoll ist offen, frei und universell implementierbar?

Die Daten wurden in ein CSV-Format umgewandelt und in ein Jupyter-Notebook eingelesen. Die Forschungsdaten sind somit nun frei zugänglich und Plattform-unabhängig

I1. (Meta)-Daten nutzen eine formale, zugängliche, gemeinsam genutzte und breit anwendbare Sprache für die Wissensrepräsentation?

Datentypen, NULL-Werte und Ontologien wurden orientierend geprüft und bedürfen weiterem Cleansing. Auf Grund der initial fehlenden Dokumentation der Metadaten werden aber zum jetzigen

Stand nicht alle Felder durch Metadaten aufgewertet werden können.

R1.2. (Meta)Daten enthalten detaillierte Provenienz-Informationen?

Die Dokumentation der Provenienz beginnt mit diesem Abschnitt. Vorherige Bearbeitungsschritte können nicht geklärt werden, da hierzu keine Informationen vorhanden sind. Über die Versionierung des Datensatzes im FAIRDOMHub sowie der Versionierung und Ablage des Metadatensatzes können Transformationschritte nun eindeutig nachvollzogen werden.

Tabelle 1: Metadaten Version 1. Zuweisung beschreibender und attribuerter Merkmale zur Etablierung eines Provenienzprozesses und einer ersten Datenbeschreibung

Nr.	Name	Non-NULL	Datentyp	Beschreibung
0	pt_BIRTHDATE	495	object	Geburtsdatum, unformatiert
1	pt_DEATHDATE	70	object	Todesdatum, unformatiert
2	pt_MARITAL	337	object	Familienstand, Zeitpunkt unklar
3	pt_RACE	495	object	Ethnie
4	pt_GENDER	495	object	Geschlecht, unformatiert
5	pt_BIRTHPLACE	495	object	Geburtsort, unformatiert
6	obs_DATE	495	object	Datum einer Untersuchung
7	obs_PATIENT	495	object	Patienten-ID
8	obs_CODE	495	object	Code einer Untersuchung, mutmaßlich LOINC
9	obs_DESCRIPTION	495	object	Beschreibung des obs_code
10	obs_VALUE	495	object	Messwert der Untersuchung, Test, Datum, numerisch
11	obs_UNITS	431	object	Einheit zu obs_VALUE
12	cond_START	495	object	Startdatum einer Erkrankung, unformatiert
13	cond_STOP	309	object	Enddatum einer Erkrankung, unformatiert
14	cond_CODE	495	float64	Code einer Erkrankung, System unbekannt
15	cond_DESCRIPTION	495	object	Beschreibung des cond_CODE
16	med_START	495	object	Startdatum einer Medikation, unformatiert
17	med_STOP	429	object	Enddatum einer Medikation, unformatiert
18	med_CODE	494	object	Code einer Medikation, System unklar
19	med_DESCRIPTION	495	object	Name einer Medikation mit Stärke und Darreichungsform
20	med_DISPENSES	495	object	unklar, ggf. Anzahl der erhaltenen VE
21	med_REASONCODE	418	float64	Code für Medikamenten-Indikation
22	med_REASONDESCRIPTION	418	object	Beschreibung des med_REASONCODE
23	Unnamed: 23	4	object	unklar

4 Aufgabe 2 - Datenqualität

4.1 Aufgabenstellung

Ermitteln Sie anhand der 3x3 Matrix, welche Aspekte von Datenqualität Sie für den Datensatz berechnen müssen (Felder in der Matrix). Sie können das online-Bewertungstool der Oregon Health & Science University (OHSU) nutzen. Berechnen Sie für 2 der relevanten Matrix-Felder den Qualitäts-score gemäß der Empfehlung von Weiskopf et al. [5]. Diskutieren Sie, wie der Score verbessert werden könnte. Setzen Sie die Verbesserung im Datensatz um. Falls dies nicht möglich ist, skizzieren Sie die Schritte zu einer Verbesserung, die Sie vornehmen können. Laden Sie den Datensatz (verändert oder unverändert) mit Appendix V2 in den FAIROMHub hoch.

4.2 Antwort - 3x3 DQA

Es erfolgte Analyse der zu beachtenden Datenqualitäts-Dimensionen (data quality assessment, DQA) mittels des Bewertungstools der OHSU. Dieses Werkzeug überprüft die drei wichtigsten Dimensionen klinischer Daten auf Vollständigkeit, Korrektheit und Aktualität. Die Analyse der zutreffenden Dimensionen ist in Abbildung 1 dargestellt.

	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 1: Bewertung der zu beachtenden Datendimensionen nach Analyse durch das 'Interactive 3x3 DQA'-Tool. Gelb hinterlegt sind die zutreffenden Dimensionen)

Die Dimensionen enthalten folgende für unseren Datensatz zutreffende (gelb) Empfehlungen (in Englischer Sprache) aus [6]:

1A Are there sufficient data for each patient?

- **Measure**

- For each patient, calculate the following:
- How many variables are present (at any time)?
- How many times have data (for any variable)?
- How many overall points of data are present (across all variables and all times)?

- **Report**

- Summary statistics and/or visualizations across all patients for counts and percentages of:
- Variables present.
- Times with recorded data.
- Overall points of data present.

1B Is the distribution of values across patients plausible (approach I)? (Approach II not shown)

- **Measure**

- For each variable where applicable, specify acceptable or expected value limit(s). These limits may be absolute (e.g. heart rate over 300 bpm) or relative (e.g. heart rate above the 95th percentile of heart rates in the dataset).
- For the above variables, how many patients have recorded data that are within the specified value limits? In other words, how many patients have data that are plausible according to the value limits?

- **Report**

- List the specified value limits for each variable. Cite evidence supporting these value limits.
- For each variable, report the counts and percentages of patients whose data are within the value limits.

2A Are there sufficient data for each variable?

- **Measure**

- For each variable:
 - * How many patients have it present (at any time)?
 - * At how many times is it present (for any patient)?
 - * How many overall points of data are present (across all patients and times)?

- **Report**

- Summary statistics for each variable for counts and percentages of the following:
 - * Patients with the variable present.
 - * Times with the variable present.
 - * Overall points of data present.

2B Is there concordance between variables?

- **Measure**

- Identify variables in your dataset where the value of one implies the value of another. E.g., a diabetes diagnosis may imply abnormal glucose values. (Note, it may not be possible or practical to develop an exhaustive list of criteria for your dataset.)
- Establish criteria based upon the above that are relevant for your dataset.

- Determine if each criterion is met by each patient.
- **Report**
 - List the criteria used, and the number and percentage of patients who meet each one, as well as the number of patients whose data violate any of the criteria.

3A Are there sufficient data for each time?

- **Measure**
 - For each time:
 - * How many patients have data recorded (for any variable)?
 - * How many variables have data recorded (for any patient)?
 - * How many overall points of data are present (across all patients and variables)?
- **Report**
 - * Summary statistics for each time for counts and percentages of the following:
 - Patients with data recorded for that time
 - Variables with data recorded for that time
 - Overall points of data present

3B Is the progression of data over time plausible?

- **Measure**
 - For each sequential variable, establish criteria for valid value changes over time (e.g., height generally does not increase by more than a certain amount each year; lab values may change, but beware of extreme outliers in a sequence; etc.)
 - Determine how many patients have data that violate each rule.
- **Report**
 - List the criteria used, and the number and percentage of patients who meet each one, as well as the number of patients whose data do not violate any of the criteria.

3C Were Data recorded with the desired regularity over time?

- **Measure**
 - For each variable for which there is a desired regularity of reporting, perform the following calculation for each patient (where n =total number of points and x =time of recording):
 - I ranges between 0 and 1. If $I=1$, then the data points are spaced with perfect regularity. Multiplying I by n gives an approximation of the number of effective data points, which may be helpful.
- **Report**
 - For each variable, report summary statistics of I for each variable across patients. Also report summary statistics for $I \times n$ if relevant.

4.3 Antwort - quality metrics

Es erfolgte die Auswahl zweier Matrixfelder zur Berechnung der Qualitätswerte. Ausgeschlossen wurden die Felder A1 und B1 da im Nachgang zur Lieferung der Beispieldaten auf eine Inkonsistenz bei der Aggregation der Patienten-Identifikationsnummern hingewiesen wurde. Da dieser Fehler ohne weitere Informationen nicht aufgearbeitet werden kann fallen die Matrixfelder zu der Qualität 'Patient' aus der Berechnung heraus.

Analyse Matrixfeld 2B - concordance between variables

Exemplarisch wurde hier die Angabe zur Messung der Sauerstoffkonzentration im arteriellen Blut herangezogen, die sich innerhalb des Datensatzes über den obs_CODE (LOINC) '2708-6', der obs_DESCRIPTION 'Oxygen saturation in Arterial blood' und der Angabe von '%' in obs_UNITS charakterisieren lässt. Zur Überprüfung der Übereinstimmungen erfolgte die Abfrage des Datensatzes wie in Code 2 dargestellt.

Code 2: DQA-2B Abfrage

```
DQA2B = DF[['obs_CODE', 'obs_DESCRIPTION', 'obs_UNITS']]
subset = DQA2B.loc[(DQA2B['obs_UNITS'] == '%') | (DQA2B['obs_CODE'] =
= '2708-6') | (DQA2B['obs_DESCRIPTION'] == 'Oxygen saturation in
Arterial blood ')]
```

Über die Analyse der Abfrage zeigte sich letztendlich, dass es keine Nicht-Übereinstimmung der Parameter gibt und sich alle Angaben zur Sättigung wie in Tabelle 2 darstellen.

Tabelle 2: Ergebnis der DQA-2B Abfrage. Die Angaben zur arteriellen Sauerstoffsättigung stimmen in allen Vorkommen überein.

obs_CODE	obs_DESCRIPTION	obs_UNITS
2708-6	Oxygen saturation in Arterial blood	%

Analyse Matrixfeld 3B - progression of data over time

Exemplarisch werde hier die Angabe zu COVID-19 Infektion wie in der Aufgabe geschildert betrachtet. Nach händischer Überprüfung des Datensatzes auf die Stichwörter 'COVID' und 'Sars' zeigten sich lediglich Einträge in der Spalte 'obs_DESCRIPTION'. Diese wurde auf entsprechende Einträge wie in Code 3 durchsucht.

Code 3: DQA-3B Abfrage

```
DQA3B = DF[['obs_DATE', 'obs_DESCRIPTION', 'obs_PATIENT']]
DQA3B.loc[DQA3B['obs_DESCRIPTION'] == 'SARS-CoV-2 RNA Pnl Resp NAA+probe ']
```

Da die Grundgesamtheit der inkludierten Patienten auf Grund des o.s. Fehlers unbekannt ist wir hier für die Analyse von 495 Patienten ausgegangen. Wie in Tabelle 3 dargestellt sind 6 Patienten mit positivem Sars-CoV-2-Nachweis im Datensatz enthalten. Die Prävalenz im Datensatz beträgt

Tabelle 3: Ergebnis der DQA-3B Abfrage. Sechs Patienten weisen einen SARS-CoV-2-Nachweis auf.

Nr.	obs_DATE	obs_DESCRIPTION	obs_PATIENT
38	2020-01-03	SARS-CoV-2 RNA Pnl Resp NAA+probe	f0f3bc8d-ef38-49ce-a2bd-dfdda982b271
82	2020-03-13	SARS-CoV-2 RNA Pnl Resp NAA+probe	067318a4-db8f-447f-8b6e-f2f61e9baaa5
126	2020-11-03	SARS-CoV-2 RNA Pnl Resp NAA+probe	ae9efba3-ddc4-43f9-a781-f72019388548
153	2020-02-03	SARS-CoV-2 RNA Pnl Resp NAA+probe	199c586f-af16-4091-9998-ee4cfc02ee7a
166	2020-02-03	SARS-CoV-2 RNA Pnl Resp NAA+probe	353016ea-a0ff-4154-85bb-1cf8b6cedf20
250	2020-02-19	SARS-CoV-2 RNA Pnl Resp NAA+probe	f58bf921-cba1-475a-b4f8-dc6fa3b8f89c

demnach 1,21%. Bei einer Durchseuchung der deutschen Bevölkerung von ca. 3,6% in 2020 scheinen die Sars-CoV-2 Nachweise im Datensatz unterrepräsentiert. Wichtig ist jedoch, dass Sars-CoV-2 erst ab dem Dezember 2019 grundsätzlich nachgewiesen wurde. Die Datumsangaben in unserem Datensatz (beginnend nicht vor dem Januar 2020) sind demnach stimmig.

Der Datensatz ist zwar unter den o.s. isolierten Betrachtungen teilweise stimmig, enthält jedoch eine Vielzahl an weiteren zu prüfenden und zu bereinigenden Feldern. Insbesondere sind die Angaben zum Familienstand nicht konsistent codiert, das Datum ist unsystematisch dokumentiert und in der Spalte 'pt_BIRTHPLACE' sind Angaben zu Stadt, Provinz und Land als String in einer Zelle angegeben. Die Beurteilung des Datensatzes nach seinem zeitlichen Abbild ist auf Grund des o.s. Fehlers nicht möglich. Weitere Fehler wie falsche Formate in der Quell-Excel-Datei und die o.s. Defizite im Bereich der Verfügbarkeit von Metadaten sind ausstehende Cleansing-Aufgaben. Es erfolgt die Ablage des Datensatzes mit überarbeitetem Datumsformat (zur Analyse 3x3 DQA erfolgt) in FAIRDOMHub unter 'Dataset_Covid_FPH_appendix_v2.csv'.

5 Aufgabe 3 - Meta-Daten

Ergänzen Sie für den Datensatz 4 verschiedene Metadaten(typen). Die Metadaten können auf Ebene des Datensatzes, des Projekts, oder auf Ebene des Datenitems angelegt werden. Dokumentieren Sie, wie Sie die Auswahl der Standards getroffen haben. Erläutern Sie, wozu genau diese Metadaten sinnvoll sind und wie sie zur besseren Nachnutzbarkeit des Datensatzes beitragen. Stellen Sie die Metadaten in einer externen Datei oder als Ergänzung des Datenfiles bereit. Laden Sie in jedem Fall Datensatz (und ggf. Metadaten-Datei) in den FAIRDOMHub hoch; verwenden Sie den Appendix V3.

5.1 Antwort - Meta-Daten-Ergänzung

Wie in Aufgabe 2 erläutert erfolgte bereits die Formatierung sämtlicher **Datumsangaben** auf den Datentyp 'datetime64[ns]'. Die Recherche nach einem Einteilungssystem für die **Medikamente** über den 'med_CODE' blieb erfolglos. Weder konnten hier Übereinstimmungen mit dem deutschen Pharmazentralnummer (PZN)-System noch mit der **Food & Drug Administration(FDA)-Taxonomie** oder dem **HLZ-FHIR-Standard** ermittelt werden. Eine Verbesserung der Metadaten an dieser Stelle war leider nicht möglich. Die Angaben zum **Familienstand** über 'pt_marital' können konventionsgemäß mit S='Single' und M='Married' ergänzt werden. Hier wäre die Umwandlung in Kategoriale Variablen sinnvoll, die auch noch weitere Angaben (z.B. 'Married-civ-spouse', 'Divorced', 'Never-married', 'Separated', 'Widowed', 'Married-spouse-absent', 'Married-AF-spouse' nach [7]). Die Angaben zur Codierung der **Untersuchungen** über 'obs_CODE' folgen zumeist der **LOINC-Terminologie**. Die entsprechende Onlineressource wird daher hierzu verlinkt.

Nach Ergänzung der nun bekannten/ermittelten Metadaten wird die Version des Metadatenatzes erstellt wie in Tabelle 4 dargestellt. Die Ergänzung der Metadaten ist nach den FAIR-Prinzipien zur verbesserten Auffindbarkeit und zur Evaluation des Daten-Kontextes wichtig. So können beispielsweise stringente LOINC-Kodierungen oder (in diesem Fall leider unbekannte) Medikamenten-Identifikatoren eine Durchsuchbarkeit Des Datensätzen erleichtern und der Vergleich bzw. die Aggregation mit anderen Forschungsdatensätzen die Weiternutzung stark vereinfachen.

Die Tabellen 1 und 4 werden als PDF im FAIRDOMHub entsprechend der Aufgabenstellung als Metadaten-Ressource abgelegt.

Tabelle 4: Metadaten Version 2. Zuweisung beschreibender und attributierter Merkmale zur Fortsetzung des Provenienzprozesses

Name	Datentyp	Beschreibung
pt_BIRTHDATE	datetime64[ns]	Geburtsdatum
pt_DEATHDATE	datetime64[ns]	Todesdatum
pt_MARITAL	object	Familienstand, unketegorisiert, M='married', S='single'
pt_RACE	object	Rasse/Ethnie, unketegorisiert, 'asian', 'black' 'w/white', 'native'
pt_GENDER	object	Geschlecht, unketegorisiert, F='female', M='male', fehlendes drittes Geschlecht
pt_BIRTHPLACE	object	Geburtsort, unketegorisiert, Enthält Stadt, Land & Staat
obs_DATE	datetime64[ns]	Datum einer Untersuchung
obs_PATIENT	object	Patienten-ID
obs_CODE	object	Code einer Untersuchung, LOINC , einzeln vorhandene Strings
obs_DESCRIPTION	object	Beschreibung des obs_code, abhängig von 'obs_Code'
obs_VALUE 5	object	Messwert der Untersuchung, zumeist fehlerhaft formatiert muss über LOINC neu erstellt werden
obs_UNITS	object	Einheit zu obs_VALUE
cond_START	datetime64[ns]	Startdatum einer Erkrankung
cond_STOP	datetime64[ns]	Enddatum einer Erkrankung
cond_CODE	float64	Code einer Erkrankung, System unklar
cond_DESCRIPTION	object	Beschreibung des 'cond_CODE'
med_START	datetime64[ns]	Startdatum einer Medikation
med_STOP	datetime64[ns]	Enddatum einer Medikation
med_CODE	object	Code einer Medikation, System unklar
med_DESCRIPTION	object	Name einer Medikation mit Stärke und Darreichungsform als Text
med_DISPENSES	object	unklar, ggf. Anzahl der erhaltenen VE
med_REASONCODE	float64	Code für Medikamenten-Indikation, System unklar
med_REASONDES[...]	object	Beschreibung des 'med_REASONCODE', als Text
Unnamed: 23	object	unklar

6 Aufgabe 4 - Datenaufbereitung

Schauen Sie sich die Qualität der Daten noch einmal bezüglich Vollständigkeit und Plausibilität an. Finden Sie mindestens 3 Schwachstellen und beheben Sie diese. Dokumentieren Sie die Data-Cleansing-Schritte. Laden Sie den bereinigten Datensatz in den FAIRDOMHub hoch; verwenden Sie den Appendix V4.

6.1 Antwort - Cleansing

Der Datensatz wurde nochmals mittels OpenOffice und dem Jupyter Notebook geprüft. Folgende Cleansing-Aufgaben wurden unternommen (siehe Code 4):

- Einträge in der Spalte 'Unnamed: 23' resultieren wahrscheinlich aus fehlerhaftem Einfügen in ein Datenfeld bei der Zusammenführung der Daten. Obwohl ggf. eine Anpassung durch Verschieben von Tupel-Teilen möglich wäre entstehen hierdurch Inplausibilitäten (z.B.STOP vor START). Alle NULL-Werte aus der Spalte 'med_START' werden daher eliminiert, da die entstehenden Tupel nicht auf Richtigkeit geprüft werden können.
- Patienten ohne Geburtsdatum ('pt_BIRTHDATE', ohne Untersuchungsdatum ('obs_DATE') und ohne Erkrankungsbeginn ('cond_START') werden aus Plausibilitätsüberlegungen ausgeschlossen.
 - Entfernung von 6 Datensätzen wegen fehlender o.s. Zeitangaben
- Constraints für Datumsangaben wurden formuliert (STOP nicht vor START, DEATH nicht vor BIRTH) und analysiert. Widersprüchliche Datensätze wurden entfernt.
 - Entfernung von 5 Datensätzen wegen widersprüchlicher Zeitangaben zur Gabe von Medikation
 - Entfernung von 0 Datensätzen wegen widersprüchlicher Zeitangaben zu Geburt und Tod
 - Entfernung von 75 Datensätzen wegen widersprüchlicher Zeitangaben zum Beginn der Erkrankung

Code 4: Ausschluss inkonsistenter Daten

```
DF2 = pd.read_csv('Dataset_Covid_FPH_appendix_v3.csv', sep=";")
nan = DF2.dropna(subset=['pt_BIRTHDATE', 'obs_DATE', 'cond_START',
                        'med_START'])
red1 = nan.drop(nan[nan['cond_STOP'] < nan['cond_START']].index)
red2 = red1.drop(red1[red1['med_STOP'] < red1['med_START']].index)
red2.drop(columns=['Unnamed: 23', 'Unnamed: 0'], axis=1, inplace=True)
red2.to_csv('Dataset_Covid_FPH_appendix_v4.csv')
```

Der Entstandene Datensatz enthält noch 409 Datensätze, die in 'Dataset_Covid_FPH_appendix_v4.csv' in das FAIRDOMHub geladen wurden.

7 Aufgabe 5 - Datenpublikation

Ihr Datensatz sollte nun bereit sein für die Publikation. Wählen Sie einen geeigneten Publikations-server/-repository und beschreiben Sie die Veröffentlichungsbedingungen. Notieren Sie, unter welcher Lizenz und in welchem Format Sie den Datensatz hochladen und begründen Sie kurz die Auswahl des Repositoriums.

7.1 Antwort - Publikation

Die Publikation erfolgt über das Repository [Zenodo](#), welches eine Professions-übergreifendes web-basiertes Portal zur Ablage verschiedener digitaler Objekte ermöglicht. Die Nutzung ist kostenfrei und erfolgt für diese Hausarbeit über die Lizenz [Attribution 4.0 International \(CC BY 4.0\)](#). Die Auswahl des Repositoriums erfolgte nach subjektiven Kriterien der Verständlichkeit der Weboberfläche, der Usability und der Barrierefreiheit des Anmeldeprozesses (z.B. im Vergleich zu [Dataverse](#)).

Mit Abschluss der Publikation ist das Dokument unter der [DOI:10.5281/zenodo.4667703](https://doi.org/10.5281/zenodo.4667703) zugänglich.

Abbildungsverzeichnis

- 1 Bewertung der zu beachtenden Datendimensionen nach Analyse durch das 'Interactive 3x3 DQA'-Tool. Gelb hinterlegt sind die zutreffenden Dimensionen) 12

Tabellenverzeichnis

- 1 Metadaten Version 1. Zuweisung beschreibender und attributierter Merkmale zur Etablierung eines Provenienzprozesses und einer ersten Datenbeschreibung 11
- 2 Ergebnis der DQA-2B Abfrage. Die Angaben zur arteriellen Sauerstoffsättigung stimmen in allen Vorkommen überein. 15
- 3 Ergebnis der DQA-3B Abfrage. Sechs Patienten weisen einen SARS-CoV-2-Nachweis auf. 16
- 4 Metadaten Version 2. Zuweisung beschreibender und attributierter Merkmale zur Fortsetzung des Provenienzprozesses 18

Literatur

- [1] Mark D. Wilkinson, Michel Dumontier, I. J. Aalbersberg, Gabrielle Appleton, Myles Axton, Arie Baak, Niklas Blomberg, Jan-Willem Boiten, Luiz Bonino Da Silva Santos, Philip E. Bourne, Jildau Bouwman, Anthony J. Brookes, Tim Clark, Mercè Crosas, Ingrid Dillo, Olivier Dumon, Scott Edmunds, Chris T. Evelo, Richard Finkers, Alejandra Gonzalez-Beltran, Alasdair J. G. Gray, Paul Groth, Carole Goble, Jeffrey S. Grethe, Jaap Heringa, Peter A. C. 't Hoen, Rob Hooft, Tobias Kuhn, Ruben Kok, Joost Kok, Scott J. Lusher, Maryann E. Martone, Albert Mons, Abel L. Packer, Bengt Persson, Philippe Rocca-Serra, Marco Roos, Rene van Schaik, Susanna-Assunta Sansone, Erik Schultes, Thierry Sengstag, Ted Slater, George Strawn, Morris A. Swertz, Mark Thompson, Johan van der Lei, Erik van Mulligen, Jan Velterop, Andra Waagmeester, Peter Wittenburg, Katherine Wolstencroft, Jun Zhao, and Barend Mons. The fair guiding principles for scientific data management and stewardship. *Scientific data*, 3:160018, 2016.
- [2] Australian Research Data Commons. Fair data: Supporting knowledge discovery and innovation.
- [3] Leibniz-Informationszentrum Technik und Naturwissenschaften Universitätsbibliothek. Die fair data prinzipien.
- [4] Jason Walonoski, Mark Kramer, Joseph Nichols, Andre Quina, Chris Moesel, Dylan Hall, Carlton Duffett, Kudakwashe Dube, Thomas Gallagher, and Scott McLachlan. Synthea: An approach, method, and software mechanism for generating synthetic patients and the synthetic electronic health care record. *Journal of the American Medical Informatics Association : JAMIA*, 25(3):230–238, 2018.
- [5] Nicole G. Weiskopf, Chunhua Weng. 3x3 dqa: Dynamic, evidence-based guidelines to enable electronic health record data quality assessment and reporting for retrospective research.

- [6] Oregon Health & Science University. Interactive 3x3 dqa.
- [7] Ronny Kohavi and Barry Becker. Uci machine learning repository: Adult data set.