Inhaltsverzeichnis

[**1**](#_heading=h.gjdgxs) **Administrative Daten 2**

[**2**](#_heading=h.30j0zll) **Verantwortlichkeiten (Kontaktdaten) 2**

[**3**](#_heading=h.1fob9te) **Datenmanagement 3**

[**3.1**](#_heading=h.3znysh7) **Datenerhebung (Daten generieren und bearbeiten) 3**

[**3.2**](#_heading=h.2et92p0) **Datenspeicherung und Datenzugriff 3**

[**3.3**](#_heading=h.tyjcwt) **Datendokumentation und Metadaten erstellen 3**

[**3.4**](#_heading=h.3dy6vkm) **Data Sharing 4**

1. Administrative Daten

| Projektbezeichnung | PF-Cost Prediction  (Patient Feature-based Cost Prediction) |
| --- | --- |
| Projekttitel | ML-Modellierung für die Prädiktion klinischer Ereignisse mit besonderem Fokus auf Versionskontrolle und Validierung |
| Projektbeschreibung mit Datensatzbeschreibung | Realistische, synthetische Patientendaten sollen ausgewertet und verwendet werden, um einen Algorithmus auf Basis von Machine Learning zu entwickeln und zu trainieren. Dieser Algorithmus soll an Hand von Patientendaten (z. B. BMI, Alter, Krankenversicherung und Anzahl der Medikamente) eine Kostenschätzung für zukünftige Gesundheitskosten ausgeben. |
| Projektlaufzeit (Beginn/Ende) | 13.01.2025-14.02.2025 |
| Forschungsförderer und –programm bzw. Eigenfinanzierung | Eigenfinanzierung: Das Projekt wurde von Studierenden im Rahmen des Studiengangs Biomedizinische Informatik und Data Science - Online-Master und Hochschulzertifikate der Hochschule Mannheim in Kooperation mit MIRACUM DIFUTURE im Modul “Datenmanagement und Archivierung im Umfeld der Forschung” unter der Leitung von Gaetan Kamdje Wabo durchgeführt. Für das Projekt bestand keine Forschungsförderung oder Finanzierung von Dritten. |
| Projektleiter ICON | - |
| Studienleiter (Principal Investigator), falls bekannt ORCID | Gaetan Kamdje Wabo  Data Scientists  Abt. für Biomedizinische Informatik  der Universitätsmedizin Mannheim |
| Zuständige Ethikkommission, Votum | - |
| Relevante Policies (falls vorhanden) | - |
| Zu beachtende Vorgaben von Dritten, optional | Es wurden im Rahmen der Aufgabenstellung mögliche Zielsetzungen vorgegeben:   * Datenvorverarbeitung: Aufbau einer strukturierten Pipeline (z. B. in Python), die Rohdaten reinigt, anreichert und Features extrahiert. * Versionierung (Data & Model): Einführung von Tools wie DVC (Data Version Control) oder Git LFS, um Daten- und Modellversionen in verschiedenen Stadien zu tracken. * Modellvalidierung: Umsetzung eines strengen Validierungs- und Testprozesses (z. B. Cross-Validation, ROC-Kurven, Metrikenreport). * CICD & Deployment: Entwurf eines vereinfachten Continuous-Integration-Konzepts (z. B. GitHub Actions) für automatisierte Trainingsläufe und optionales Deployment in einer Testumgebung   In der Projektlaufzeit wurden die Projektziele entsprechend der Beschreibung (siehe oben) spezifiziert. |
| Geplantes Publikationsdatum / Zeitraum der Datenpublikation | 13.02.2025 |
| Geplante Kosten (falls Kosten nicht durch Dritte getragen werden, z.B. für Datenmanagement (FAIRe Daten), Archivierung) | - |
| Datum der Erstellung/Änderung DMP, Version | 12.02.2025, Version 1.0 |

1. Verantwortlichkeiten (Kontaktdaten)

| Rolle | Name / Kontaktdaten | Kommentar |
| --- | --- | --- |
| Studierende | Eva Mieskes eva.mieskes@master-bids.de | Projektablauf & Datenmanagementplan (DMP) |
| Studierender | Till Bömmer  till.boemmer@master-bids.de | Work Environment: Github |
| Studierender | Florian Streuter florian.streuter@master-bids.de | ETL Pipeline & Datenaufbereitung |
| Studierender | Fabian Müller-Graf fabian.mueller-graf@master-bids.de | Datenexploration |
| Studierender | Tician Schnitzler tician.schnitzler@master-bids.de | ML-Modell & Validierung |
| Studierender | Michael Kruszewski michael.kruszewski@master-bids.de | Integration & Deployment (CI/CD) |
| Dozent | Gaetan Kamdje Wabo  gaetankamdje.wabo@master-bids.de | Modul “Datenmanagement und Archivierung im Umfeld der Forschung” |

1. Datenmanagement
   1. Datenerhebung (Daten generieren und bearbeiten)
   2. **Beschreibung der zu erhebenden Forschungsdaten**

Es werden synthetische Patientendaten Datenbank (Synthea Datensatz "Allergy", Rohdaten-Tabellen im CSV-Format) verwendet. Für die Sekundäranalyse sollen Fake-Patientendaten (“Faker” Python Paket) verwendet werden.

* 1. **Verarbeitungsschritte der Rohdaten und Datenanalyse**
* Die Allergy-Datensatz wird auf seine Datenbank-Struktur untersucht.
* Es werden ausschließlich anonymisierte bzw. synthetische Patientendaten verwendet.
* Die Inhalte werden grafisch aufgearbeitet.
* Dubletten und Null-Werte werden ermittelt und entfernt.
* Patienten Features werden auf ihre Relevanz zur Kostenvorhersage untersucht.
* Eine finale Datenbank (“synthea.db”) für die Modellentwicklung wird erstellt.
* Es soll ein ML-Modell entwickelt werden. Das passende Regressionsmodell wird hierfür gewählt.
* Das Modell wird in einer Applikation (Testversion) implementiert.
  1. **Maßnahmen zur Qualitätssicherung**
* Dubletten und Null-Werte im Trainingsdatensatz werden ermittelt und entfernt.
* Das ML-Modell wird validiert über eine primäre (Teil des Original Datensatzes “Synthea Daten”) und sekundäre (anderer Datensatz “Fake Daten”) Analyse.
  1. Datenspeicherung und Datenzugriff
  2. **Datenorganisation inklusive Versionierung**

Github (<https://github.com/MichaelKru92/Projekt-ML-Modelierung>) soll für die Versionierung und Datenorganisation verwendet werden.

Projektbezogene Daten (MD, TXT, ZIP, CSV, DB, PY, IPYNB, HTML, PKL, YML, XML) werden in Github abgelegt und eindeutig identifizierbar benannt.

* 1. **Datenerhaltung (inkl. Speicherung und Backup)**

Es werden während der Projektlaufzeit die Daten ggf. lokal im Git der jeweiligen Person bearbeitet und regelmäßig in Github bereitgestellt bzw. aktualisiert.

Das Präsentations-Dokument (Google Präsentationen) wird in einem gesicherten, geteilten Google Drive-Ordner bearbeitet und gespeichert.

* 1. **Zugriffssicherheit**

Das Projekt wird am 13.05.2025 in GitHub veröffentlicht.

* 1. Datendokumentation und Metadaten erstellen

Eine Dokumentation der Forschungsdaten und damit verbundenen Metadaten für die (Roh-) Datensätze, die Transformationen sowie der Ergebnisse wird angefertigt und gespeichert in GitHub.

* Folgende Dateien werden erstellt unter Beschreibung der benutzten Datenelemente und Formate und Metadaten:
  + Auflistung aller generierten relevanten (Roh-) Datensätze
  + Auflistung aller relevanten Methoden, Programme, log-files
  + Auflistung der Ergebnisdateien

Metadaten werden in einem XML Schema zur Verfügung gestellt.

* 1. Data Sharing
  2. **Publikation, Verfügbarmachung der Daten**

Das Projekt inklusive der synthetischen Daten wird über GitHub verfügbar gemacht.

* 1. **Langzeitarchivierung**

Wie in den Regeln zur guten wissenschaftlichen Praxis der DFG vorgesehen, werden die Forschungsdaten für mind. 10 Jahre archiviert. Eine Langzeitarchivierung ist nicht vorgesehen.  
Die zu erwartende Gesamtgröße beträgt maximal 15 MB.