

Detaillierte Ergebnisse – Teilprojekt

Datenanalyse und Datenaustausch

DATENAUSWERTUNG FÜR EINE BESSERE MEDIZINISCHE VERSORGUNG UND FORSCHUNG

FORSCHUNGSINTERESSE:

- Wie können medizinische Diagnosen genauer und Therapien effektiver werden?
- Wie kann die medizinische Forschung effizienter gestaltet werden?

VORGEHEN:

Die Wissenschaftler*innen entwickelten eine Plattform zum Austausch medizinischer Daten.

ERGEBNISSE/ERKENNTNISSE:

- Die Plattform kann medizinische Daten aus unterschiedlichen Quellen des Gesundheitswesens zusammenführen (z. B. aus ärztlichen und physiotherapeutischen Praxen, Krankenhäusern, Reha-Kliniken). Die Datenzusammenführung ermöglicht eine umfangreiche Datenauswertung. Dies ist sowohl für die medizinische Forschung als auch für die Diagnostik und Therapie hilfreich.
- Die Forschung medizinische Forschung wird schneller und effizienter, da einmal erhobene Daten mehrfach für unterschiedliche Forschungsvorhaben genutzt werden und eine immer wiederkehrende, oftmals sehr aufwendige Datenerhebung entfällt. Wissenschaftler*innen können prüfen, welche Daten schon vorliegen und welche Informationen sie darüber hinaus noch benötigen.
- Zudem wird durch die Verknüpfung der Daten ermöglicht, Muster im Krankheitsbild einer Person zu erkennen, Rückschlüsse auf Vorerkrankungen zu ziehen und den Gesundheitszustand in einen Gesamtzusammenhang einzuordnen. Infolgedessen werden Diagnosen genauer, Therapien effizienter und somit die Behandlung individueller.
- Die Wissenschaftler*innen nahmen eine Testinstanz der Datenplattform in Betrieb, um sie erproben, evaluieren und finalisieren zu können.

DATENSATZ FÜR MEDIZINISCHE AUSWERTUNG - ANWENDUNGSFALL KARDIOVASKULÄRE ERKRANKUNGEN

FORSCHUNGSINTERESSE:

Wie sind medizinische Daten unterschiedlicher Art und Qualität aufzubereiten, um sie für eine Datenaustauschplattform nutzbar zu machen?

VORGEHEN:

- Die Wissenschaftler*innen untersuchten die Frage am Beispiel kardiovaskulärer Erkrankungen (Herz-Kreislauf-Erkrankungen). Zur Untersuchung dieses Anwendungsfall stellten Projektpartner des Zukunftslabors entsprechende Datenbestände bereit.
- Diese Daten strukturierten die Wissenschaftler*innen anhand von Steckbriefen, die sowohl allgemeine Informationen (z. B. mit welchen Methoden die Daten erhoben wurden) als auch spezifische Details (z. B. ob bei dem Patient Blutdruck gemessen wurde) enthielten.

ERGEBNISSE/ERKENNTNISSE:

- Es entstand ein umfangreicher und strukturierter Datenbestand für medizinische Auswertungen.
- Anhand der Steckbriefe können sich Wissenschaftler*innen aus der medizinischen Forschung einen Überblick über die Art und Qualität der Daten verschaffen und entscheiden, ob die Informationen für ihre Forschung geeignet sind.

STANDARD OPENEHR FÜR DIE MODELLIERUNG MEDIZINISCHER DATEN

FORSCHUNGSINTERESSE:

Wie kann die Abbildung von Gesundheitsdaten in elektronischen Patient*innenakten standardisiert werden?

VORGEHEN:

- Die Wissenschaftler*innen verglichen verschiedene Standards zur Modellierung medizinischer Daten.
- Zudem erarbeiteten die Wissenschaftler*innen Empfehlungen für die Wahl eines Standards, da es hierzu bisher kaum Empfehlungen gab. Zunächst interviewten sie fünf Expert*innen aus dem Fachgebiet der Datenmodellierung und werteten die Gespräche mittels der qualitativen Inhaltsanalyse aus. Auf Basis der Expert*inneninterviews erstellten die Wissenschaftler*innen eine Online-Umfrage, um die Aussagen der Expert*innen zu validieren. An dieser Umfrage nahmen Personen teil, die Daten zumeist innerhalb wissenschaftlicher Forschungsprojekte modellieren. Dadurch identifizierten die Wissenschaftler*innen Rahmenbedingungen, die die Nutzung eines Standards beeinflussen.

ERGEBNISSE/ERKENNTNISSE:

- Die Wissenschaftler*innen entschieden sich dafür, den Standard open Electronic Health Records (openEHR) zu nutzen. Dieser Technologiestandard ist frei verfügbar, sodass alle Marktteilnehmer*innen Zugriff darauf haben, um eigene Produkte bzw. Software-Lösungen zu entwickeln. Dadurch bietet openEHR die optimale Voraussetzung für Interoperabilität. openEHR

ermöglicht die Verwaltung und Speicherung sowie den Abruf und den Austausch von Gesundheitsdaten in elektronischen Patientenakten.

- Anhand eines Beispieldatensatzes nahmen die Wissenschaftler*innen eine Modellierung in openEHR vor, um Kenntnisse im Umgang mit dem Datenstandard zu gewinnen. Für die Modellierung erstellten sie unter anderem Archetypen; das sind maschinenlesbare, abgestimmte Repräsentationen relevanter klinischer Konzepte (z. B. Blutdruck) als (Daten-)Modelle. In Templates werden diese Archetypen so miteinander verknüpft, dass verschiedene Anwendungsfälle (z. B. medizinische Dokumente) abgebildet werden können.

SERVER UND PLATTFORMWERKZEUGE

FORSCHUNGSINTERESSE:

Wie können die Daten in die Plattform integriert, verwaltet und abgerufen werden?

VORGEHEN:

- Die Wissenschaftler*innen recherchierten bereits existierende Software-Komponenten, die sie für Integration (Input), die Verwaltung und den Abruf der Daten (Output) nutzen könnten. Sie prüften die Komponenten auf ihre Einsatzfähigkeit im Rahmen der Plattform.
- Bei ihrer Recherche identifizierten sie auch Komponenten, die nicht geeignet waren, und entwickelten eigene Lösungen dafür.
- Zudem analysierten sie verschiedene Server und entschieden sich für einen geeigneten.

ERGEBNISSE/ERKENNTNISSE:

- Die Wissenschaftler konnten auf bestehende Software-Komponenten zurückgreifen, darunter die EHRbase, ein openEHR-Server zur Datenverwaltung, und die NUM-Plattform, ein Portal zum Suchen und Abrufen von Daten. Die EHRbase überträgt Daten auf einen openEHR-Server, der eine weit verbreitete Software nutzt und die Daten nach den FAIR-Prinzipien verfügbar macht: (auf-)findbar, zugänglich, interoperabel und wiederverwendbar. Dadurch können die Daten sowohl von Menschen als auch von Programmen gefunden und genutzt werden, sie sind standardisiert und mit anderen Datensätzen verknüpfbar.
- Zum Abrufen der Daten verwendeten die Wissenschaftler die NUM-Plattform, die eine Übersicht über Metadaten, Abfragen und einen Datenexplorer bietet. Über dieses webbasierte Portal können die Daten des Zukunftslabors gefunden und abgerufen werden.
- Da es an einem Open-Source-Tool zum Import von Daten fehlte, entwickelten die Wissenschaftler den openEHR_FLAT_Loader, der für den Import medizinischer Daten in die EHRbase genutzt werden kann. Diese drei Plattformwerkzeuge (openEHR-Server, NUM-Plattform und FLAT-Loader) können modular in eine Datenplattform integriert werden.

ONLINE-DOKUMENTATION ZUM AUFBAU EINER DATENPLATTFORM

FORSCHUNGSINTERESSE:

Wie können die Erkenntnisse zum Aufbau einer Datenplattform nachhaltig gesichert und Interessent*innen zugänglich gemacht werden?

VORGEHEN:

Der Aufbau von Datenplattformen, die die Zusammenführung, Verwaltung und den Abruf unterschiedlicher Daten ermöglicht, ist mit viel Kosten und Zeit verbunden. Um ihre erworbenen Erkenntnisse nachhaltig zu sichern und Interessent*innen zugänglich zu machen, erstellten die Wissenschaftler*innen eine Online-Dokumentation zum Aufbau einer solchen Plattform.

ERGEBNISSE/ERKENNTNISSE:

- Die Online-Dokumentation beschreibt, wie die Plattformwerkzeuge funktionieren und in Betrieb genommen werden können.
- Sie enthält außerdem Links zu frei verfügbaren Codes (z. B. für den FLAT-Loader).
- Die Wissenschaftler*innen evaluierten diese Online-Dokumentation im Rahmen eines Studierendenprojektes. Dadurch konnten sie testen, ob die Anleitung verständlich ist und wo es Verbesserungspotenzial gibt. Die Studierenden erprobten den Einsatz der Plattform und die Nutzung der Online-Dokumentation im Rahmen eines fiktiven Szenarios. Sie sollten sich in die Rolle von Forschenden und Plattformbetreibern versetzen und die Plattform aus Sicht dieser beiden Nutzergruppen testen und bewerten. Die Ergebnisse der Evaluation nutzten die Wissenschaftler*innen, um die Plattform und die Dokumentation zu optimieren. Eine Empfehlung der Studierenden war, z. B. Erklärungsvideos zu entwickeln und den Nutzer*innen zur Verfügung zu stellen.

PRIVATHEITSBEWAHRENDE DATENANALYSE

FORSCHUNGSINTERESSE:

Wie können medizinische Daten sicher ausgetauscht und ausgewertet werden, sodass die Privatsphäre der Patient*innen geschützt bleibt und trotzdem alle wichtigen Informationen übermittelt und genutzt werden können?

VORGEHEN:

- Die medizinischen Daten, die in der Plattform des Zukunftslabors verarbeitet werden, sind äußerst sensibel und erfordern besonderen Schutz. Daher konzentrierten sich die Wissenschaftler*innen auf privatheitsbewahrende Analysemethoden, insbesondere auf Privacy Preserving Machine Learning (ML) Modelle. ML-Verfahren sind vielversprechend für die Analyse großer Mengen an Gesundheitsdaten.
- Anhand eines Tests verglichen die Wissenschaftler*innen, ob medizinische Datenbanken, die mithilfe von ML-Modellen ausgewertet werden, anfälliger für Angriffe sind als Datenbanken, die ohne ML ausgewertet werden. Hierzu führten die Wissenschaftler*innen selbst Angriffe durch: Angriffe auf Attributinferenzen und Angriffe durch Schätzung fehlender Werte.
- Bei einem Inferenzangriff wird das ML-Modell angegriffen, das zur Auswertung der Datenbank eingesetzt wird. Hierbei verwenden die Angreifer*innen Informationen über die Art und Weise, wie das Modell trainiert wurde, um daraus neue Informationen abzuleiten. Bezogen auf die medizinische Datenbank bedeutet dies, dass die Angreifer*innen Informationen erlangen, welche Attribute - also welche Merkmale von Patient*innen - das ML-Modell analysiert. Das können allgemeine Attribute sein (z. B. Hautfarbe) oder spezifische (z. B. Alter, Geschlecht). Indem die Angreifer*innen große Mengen dieser Daten analysieren, erhalten sie Wissen über die Datenbank oder deren Inhalt - ohne auf die Datenbank selbst zuzugreifen.
- Die Schätzung fehlender Werte bezieht sich auf eine Datenbank, bei der kein Maschinelles Lernen angewendet wird. In manchen Datensätzen fehlen bestimmte Werte (z. B. zum Geschlecht der Patient*innen). In solchen Fällen versuchen die Angreifer*innen, die fehlenden Informationen auf Basis von Mittelwerten zu schätzen. Wenn z. B. alle Personen in einem bestimmten Datensatz durch das Attribut „männlich“ gekennzeichnet sind, dann wird der fehlende Wert ebenfalls auf „männlich“ geschätzt.

ERGEBNISSE/ERKENNTNISSE:

- Der Vergleich dieser beiden Techniken zeigte, dass ein Inferenzangriff auf ein ML-Modell nicht schädlicher bzw. genauso schädlich ist wie die Technik zur Schätzung fehlender Werte. Das ML-Modell stellt demnach keine größere Schwachstelle dar als andere Techniken.
- Darüber hinaus fanden die Wissenschaftler*innen im Rahmen ihrer Inferenzangriffe heraus, dass diese erfolgreicher sind, je mehr die Angreifer*innen über das trainierte ML-Modell wissen. Bei sogenannten Black-Box-Modellen sind Angriffe weniger schädlich, da hier nicht klar ist, wie das Modell zu seiner Auswertung kommt. Bei sogenannten White-Box-Modellen ist bekannt, wie sie trainiert wurden, weshalb die Angriffe hier schädlicher sein können.
- Eine Herausforderung besteht darin, dass ML-Modelle oft veröffentlicht werden, um sie weiterzuentwickeln, was Angreifer*innen hilft. Um die Daten zu schützen, die durch ML-Modelle analysiert werden, kann die Differenzielle Privatsphäre genutzt werden. Dabei werden Datensätze absichtlich mit falschen Informationen angereichert, um Angreifer*innen in die Irre zu führen, indem sie falsche Schlüsse ziehen. Beispiel: die Annahme eines gemischt-geschlechtlichen Datensatzes, obwohl der ursprüngliche Datensatz nur männliche Patient*innen enthielt.