

Virtuelle Kohorten für die Medizinforschung

Künstliche Intelligenz im Dienst des Datenschutzes

Forschende des DZNE und des Helmholtz-Zentrums für Informationssicherheit (CISPA) wollen mit Hilfe „künstlicher Intelligenz“ die Weitergabe von Genomdaten für Forschungszwecke unter den strengen Auflagen des Datenschutzes erleichtern. Ziel ist die Generierung „virtueller Kohortendaten“, die zentrale Informationen realer Probanden enthalten, jedoch keine Rückschlüsse auf den Einzelnen zulassen. Das Projekt mit dem Namen „PRO-GENE-GEN“ hat ein Gesamtvolumen von rund 360.000 Euro. Es wird in den nächsten drei Jahren von DZNE, CISPA und der Helmholtz-Gemeinschaft finanziert.

Für die Therapieforschung ist das menschliche Erbgut ein wahrer Fundus. Denn hier können Mutationen angelegt sein, die bestimmte Erkrankungen begünstigen oder gar direkt auslösen. Neben den Genomdaten im strengen Sinne ist auch das „Transkriptom“ von Interesse, da es Informationen darüber enthält, welche Abschnitte des Erbguts tatsächlich aktiv sind – ein Muster, das sich krankheitsbedingt verändern kann. Wissenschaftliche Studien mit Personenbeteiligung sind allerdings enorm aufwändig. Deshalb besteht großes Interesse daran, die Daten solcher Untersuchungen allgemein zugänglich zu machen: etwa um die Nachprüfbarkeit von Forschungsergebnissen zu gewährleisten oder auch um sie Forschungsprojekten bereitzustellen, die sich erst im Nachhinein ergeben haben. „Die Medizinforschung ist heutzutage datengetrieben. Sogenanntes Big Data gilt als ein Schlüssel für die Entwicklung personalisierter Therapien, die besser als herkömmliche Behandlungen passgenau auf jeden Einzelnen zugeschnitten sind“, erläutert Dr. Matthias Becker, Bioinformatiker am DZNE-Standort Bonn.

Schutz der Privatsphäre

Eine Weitergabe von Studiendaten sei bislang jedoch nur eingeschränkt möglich, so Becker: „Der Umgang unterliegt strengen gesetzlichen Regelungen, denn die Daten sind personenbezogen. Es gibt zwar Mechanismen zur Wahrung der Privatsphäre, doch diese sind entweder mit erheblichem Aufwand verbunden oder in der Praxis nicht umsetzbar. Speziell Genomdaten können deshalb nicht im notwendigen Maße geteilt werden, um den wissenschaftlichen Fortschritt zu sichern.“

Das Forschungsteam von DZNE und CISPA will deshalb Methoden entwickeln, die eine bessere Verbreitung solcher Informationen ermöglichen. „Auf der Grundlage echter Genomdaten möchten wir synthetische Datensätze erstellen, die Schlüsselinformationen der ursprünglichen Daten beinhalten und gleichzeitig die Privatsphäre vollumfänglich absichern. Es geht gewissermaßen um ein datenschutzkonformes Abbild. Ähnlich einer Zeugenaussage, bei der die Stimme verfremdet wird, um die Identität zu schützen“, sagt Becker. „Dadurch lassen sich große Datensätze öffentlich zugänglich machen, was für Fortschritte in der Medizin enorm wichtig ist.“

Lernfähige Algorithmen

Die Forscher setzen zunächst bei bestimmten Fragestellungen an. „Es kann zum Beispiel um das Muster der Genexpression gehen. Also darum, welche Gene bei einer bestimmten Erkrankung aktiv sind“, so Prof. Mario Fritz, Wissenschaftler am CISPA, der das Forschungsvorhaben gemeinsam mit Becker koordiniert. „Wir möchten lernfähige Algorithmen darauf trainieren, solche Muster in

Genomdaten zu erkennen. Das ist eine Herausforderung, denn schon die Daten eines einzelnen Genoms sind äußerst komplex. Und man trainiert die Algorithmen anhand der Daten von Hunderten oder gar Tausenden von Personen. Hierin liegen die Stärken des maschinellen Lernens.“

Mit Hilfe einer Methode, die sich „privatsphärenkonforme, generative Modellierung“ nennt, wollen die Forschenden zudem Computermodelle entwickeln, welche die Quintessenz solcher Datenmuster nachbilden und zugleich den Personenbezug ausblenden. „Man kann sich das vielleicht wie einen intelligenten Filter vorstellen“, so Fritz. Auf diese Weise sollen reale Daten in synthetische überführt werden. „Gehen wir etwa von den Daten von 1000 Personen aus, dann stehen am Ende eine ähnlich große Anzahl synthetischer Datensätze. Das nennen wir virtuelle Kohorte. Deren Daten können genauso wie reale Daten mit gängigen Werkzeugen der Genforschung analysiert werden ohne jedoch die Privatsphäre der realen Personen zu verletzen.“

Über das Deutsche Zentrum für Neurodegenerative Erkrankungen e.V. (DZNE)

Das DZNE erforscht sämtliche Aspekte neurodegenerativer Erkrankungen (wie beispielsweise Alzheimer, Parkinson und ALS), um neue Ansätze der Prävention, Therapie und Patientenversorgung zu entwickeln. Durch seine zehn Standorte bündelt es bundesweite Expertise innerhalb einer Forschungsorganisation. Das DZNE kooperiert eng mit Universitäten, Universitätskliniken und anderen Institutionen auf nationaler und internationaler Ebene. Das DZNE ist Mitglied der Helmholtz-Gemeinschaft.

Über das CISPA Helmholtz Zentrum für Informationssicherheit

Das CISPA in Saarbrücken ist eines der weltweit führenden Forschungseinrichtungen der Informationssicherheit welches Fragen der Datensicherheit und Privatsphäre in all ihren Facetten untersucht, mit denen unsere Gesellschaft im Zeitalter der Digitalisierung konfrontiert wird. Datensicherheit und Privatsphäre in der Medizin sowie Grundlagenforschung in der künstlichen Intelligenz und des maschinellen Lernens sind Themen von zentraler Bedeutung seit der Gründung des Zentrums. Web: cispa.saarland