https://gen-ethisches-netzwerk.de/node/4451

Gen-ethischer Informationsdienst

# Bedrohungen für den genetischen Datenschutz

## Zugriffsmöglichkeiten auf die DNA-Privatsphäre

AutorIn
Dr. Zhiyu Wan




Foto: gemeinfrei auf flickr.com (16437392623)

Aktuelle Entwicklungen im Gesundheitswesen, der Wissenschaft und dem Markt für Internet-Gentests haben zu einem dramatischen Anstieg der Menge an genomischen Daten geführt, die gesammelt, verwendet und geteilt werden.

Die Expansion der genetischen Datensammlungen wirft neue und herausfordernde Bedenken in Bezug auf die Privatsphäre aller Menschen auf – sowohl rechtlich als auch technisch. In diesem Artikel stellen wir bestehende und neu entstehende Bedrohungen für den genetischen Datenschutz vor.

## Angriffe auf die Privatsphäre

Ein Großteil der Forschung mit genetischen Daten wird in Verbindung mit Datensätzen – demografische Daten, soziale und verhaltensbezogene Gesundheitsfaktoren und Messwerte auf molekularer oder klinischer Ebene (z.B. aus elektronischen Gesundheitsakten) – durchgeführt, aus denen direkt identifizierende Informationen entfernt wurden. Es gibt jedoch eine lebhafte wissenschaftliche Debatte darüber, ob genetische Daten allein oder in Kombination mit anderen Datenformen de-identifiziert oder de-anonymisiert werden können. Im Laufe der Jahre bewiesen eine Reihe von Forscher*innen ihre Fähigkeit, Personen zu identifizieren, deren Daten ohne persönliche Identifikatoren für die Genomforschung verwendet wurden.

Die Nutzung genetischer Daten auf individueller Ebene, selbst ohne personenbezogene Daten, beinhaltet die Möglichkeit der **Re-Identifizierung**. Zum Beispiel könnten Datenempfänger*innen phänotypische Informationen aus genetischen Markern ableiten und diese zur Re-Identifizierung nutzen. In einer Studie identifizierten die Forscher*innen Personen durch die Vorhersage äußerer Merkmale wie Augen- und Hautfarbe aus Genomsequenzen. In ähnlicher Weise könnten umgekehrt genetische Eigenschaften potenziell aus phänotypischen Merkmalen (z.B. aus bestimmten Erkrankungen, äußeren Merkmalen oder 3D-Gesichtsstrukturen) abgeleitet und zu Identifizierungszwecken verwendet werden – auch wenn die tatsächliche Aussagekraft dieser Analysen umstritten ist. Darüber hinaus ist es möglich, demografische Daten, die häufig mit genetischen Daten gemeinsam verwendet werden, zur Re-Identifizierung genetischer Daten zu nutzen, wenn sie mit anderen leicht zugänglichen Datenquellen in Verbindung gebracht werden. Im Jahr 2013 wurden die Teilnehmer*innen des Personal Genome Project identifiziert, indem Sweeney et al. die Datensätze anhand demografischer Eigenschaften mit öffentlich zugänglichen Wähler*innenverzeichnissen verknüpften.[1] Im selben Jahr identifizierten Gymrek et al. bestimmte Teilnehmer*innen des 1000-Genome-Projekts, indem sie deren Nachnamen aus Short Tandem Repeats (STRs) (kurze Wiederholungssequenzen die in der Forensik zu Identifikationszwecken verwendet werden) auf dem Y-Chromosom ableiteten, die sie mit anderen demografischen Daten aus öffentlichen Quellen kombinierten.[2]

Bei Genotyp-Phänotyp-Analysen, wie z.B. genomweiten Assoziationsstudien (GWAS), veröffentlichen Forscher*innen in der Regel nur zusammenfassende Statistiken. Im Jahr 2008 zeigten Homer et al. jedoch, dass auch GWAS-Statistiken anfällig für das Rückschließen auf die **Gruppenzugehörigkeit** von individuellen Proband*innen sind.[3] Das heißt, es ist nachweisbar, dass die Daten einer bekannten Zielperson in einen solchen Datensatz einge-flossen sind und damit kann auch die Zugehörigkeit dieser Person zu einer potenziell sensiblen Gruppe abgeleitet werden. Die Effizienz dieser Strategie wurde von anderen Wissenschaftler*innen in Frage gestellt, konnte jedoch in nachfolgenden Studien durch die Nutzung von weiteren statistischen Variablen noch einmal verbessert werden. Darüber hinaus haben mit individuellen genetischen Datensätzen trainierte Machine-Learning-Modelle das Potenzial, die Genotypen und Zugehörigkeiten der Teilnehmer*innen offenzulegen.

Durch die Ähnlichkeit genetischer Daten zwischen **biologischen Verwandten** können deren Genotypen und Veranlagungen für bestimmte Erkrankungen zu einem gewissen Grad abgeleitet werden, selbst wenn deren eigene genetische Daten nie erfasst wurden. In jüngster Zeit wurden noch bessere Rekonstruktionsstrategien entwickelt, um die Genotypen und Phänotypen von Individuen aus den Daten ihrer Verwandten abzuleiten. Im April 2018 verwendete das FBI in den USA für einen ungeklärten Fall genetische Daten, um den als Golden State Killer bekannten Serienmörder zu verhaften. Die Ermittler*innen stellten das genetische Profil des damals noch unbekannten Verdächtigen in der öffentlichen Datenbank GEDmatch ein. Durch die sogenannte weitreichende Verwandtensuche (*long-range familiar search*), bei der Familienangehörige anhand von Übereinstimmung von DNA-Sequenzen identifiziert werden, fanden sie einen Cousin dritten Grades des Verdächtigen. Von diesem ausgehend konnte dann ein Stammbaum mit weiteren Familienmitgliedern rekonstruiert und anschließend der Verdächtige selbst ausfindig gemacht werden.

## Datenschutz im Kontext

Die alleinige Fokussierung genomischer Forschung berücksichtigt nicht die potenziellen Auswirkungen der zunehmenden Verfügbarkeit genetischer Daten in anderen Bereichen. Eine Vielzahl von Einzelpersonen und Organisationen sammeln, verwenden und verbreiten heute Gendaten in einem nie dagewesenen Ausmaß. Infolgedessen werden diese Daten zu einer zunehmend nützlichen Ressource für verschiedene Akteur*innen wie Arbeitgeber*innen, Versicherungen, Strafverfolgungsbehörden etc. Zahlreiche Studien deuten darauf hin, dass zumindest einige Menschen besorgt darüber sind, wohin genetische Daten über sie fließen und wie sie verwendet werden – mit möglichen unerwünschten und unerwarteten Konsequenzen. Zusätzlich zu den häufig untersuchten Ängsten vor Diskriminierung können diese Informationen auch familiäre Beziehungen beeinflussen, z.B. durch Bestätigung oder Widerlegung einer biologischen Vaterschaft, Suche nach bisher unbekannten Verwandten oder Identifizierung vormals anonymer Keimzellspender*innen. Die Bedenken hinsichtlich der möglichen Verwendung genetischer Daten und der daraus resultierenden Konsequenzen werden meist als Wunsch nach genetischer Privatsphäre formuliert. Dies kann die individuelle Bereitschaft beeinträchtigen, sich klinischen Tests zu unterziehen oder an Forschungsprojekten teilzunehmen. Eine solche Zurückhaltung aufgrund von Datenschutzbedenken wiederum, kann bestehende gesundheitliche Ungleichheiten verschärfen und wissenschaftliche Erkenntnisse verhindern. Wenn Wissenschaftler*innen also ihre Studien planen, durchführen und diskutieren, müssen sie berücksichtigen, wie genetische Daten verwendet werden und wie sich die Art der Nutzung darauf auswirkt, ob die Daten auch außerhalb des Forschungssettings kontrolliert werden können.

**Außerhalb des Forschungsbereiches**

Millionen von US-Bürger*innen haben Direct-to-Consumer-Gentests (DTC-GT) von Unternehmen erworben, die persönliche Auskünfte über eine Vielzahl von Themen, wie z.B. Gesundheit, Abstammung, Verwandtschaftsverhältnisse (z.B. Vaterschaft) sowie Lebensstil und Wohlbefinden, versprechen. Es gibt inzwischen zahlreiche Medienberichte darüber, wie Verbraucher*innen mittels dieser Daten biologische Verwandte aufspüren – mit komplexen, sowohl positiven als auch negativen Konsequenzen. Manche Menschen freuen sich, neue Familienangehörige zu gewinnen oder ihre biologische Herkunft zu erfahren, während die Ergebnisse oder unerwünschte Kontakte durch unbekannte biologische Verwandte andere beunruhigen. Es gibt jedoch praktisch keine rechtlichen Einschränkungen, wie Kund*innen diese Daten verwenden dürfen, obwohl die resultierenden rechtlichen Konsequenzen beträchtlich sein können – einschließlich Scheidungen und Bemühungen, bestehende Unterhaltszahlungen für Kinder einzustellen.

Für den Schutz genetischer Daten ist ebenfalls relevant, dass Millionen von Menschen ihre Ergebnisse von DTC-GT heruntergeladen und in Datenbanken von Drittanbieter*innen veröffentlicht haben, um die Suche nach Verwandten zu erleichtern oder gesundheitsbezogene Interpretationen zu erhalten. Diese Websites unterliegen selten einer Art von Regulierung, die über das hinausgeht, was sie in ihren Nutzungsbedingungen angeben. Zudem behalten sich solche Websites das Recht vor, ihre Praktiken zu ändern, was als Reaktion auf öffentlichen Druck auftreten kann, aber auch durch Änderungen des Geschäftsmodells. Die betroffenen Datensätze erleichtern eine forensische Nutzung und bieten wahrscheinlich das größte Risiko für die Re-Identifizierung genetischer Daten.

**Forensischer Kontext**

Die Strafverfolgung und ihr möglicher Zugriff auf genetische Informationen spielt in der öffentlichen Meinung über genetische Daten eine große Rolle. Ungelöste Fälle mit hohem Bekanntheitsgrad, die letztendlich mittels Gendaten gelöst wurden, haben ein starkes Interesse an dieser Problematik geweckt. In den USA gab es im Laufe der Jahre Bestrebungen, die von der Regierung betriebenen forensischen Datenbanken auf Bundes-, Landes- und lokaler Ebene zu erweitern. Die Strafverfolgungsbehörden können auch versuchen, die Offenlegung genetischer Informationen zu erzwingen, die sich im Besitz einer Einzelperson oder eines Unternehmens wie eines Gesundheitsdienstleisters, eines DTC-GT-Unternehmens oder Forscher*innen befinden. Darüber hinaus können Ermittler*innen auch versuchen, öffentliche Datenbanken zu nutzen oder die Dienste eines DTC-GT-Unternehmens für forensische Genealogiezwecke zu

nutzen. Bisher hat sich die Strafverfolgung in den USA weitgehend auf öffentlich zugängliche Datenbanken (z.B. GEDmatch) und private Datenbanken im Besitz von Unternehmen konzentriert, die freiwillig mit der Polizei zusammenarbeiten (z.B. FamilyTreeDNA). Gleichzeitig gab es nur begrenzt Forschung zu Möglichkeiten die Datenschutzrisiken unbeteiligter Verwandter zu reduzieren, die durch die Suchstrategie sog. forensischer oder investigativer genetischer Genealogie entstehen.

**Lösungsansätze für den Gendatenschutz**

Ein angemessenes Schutzniveau für DNA-Daten benötigt eine Kombination aus technischen und gesellschaftlichen Lösungen, die den Kontext berücksichtigen, in dem die Daten angewendet werden (Anmerkung der Redaktion: mehr zu Lösungsansätzen im Originalartikel). Dieses Ziel ist jedoch nicht ohne Weiteres zu erreichen. Aus technischer Sicht ist es herausfordernd, Datenschutz-Technologien, die in wissenschaftlichen Artikeln veröffentlicht oder in einer kleinen Pilotstudie getestet werden, zu einer vollwertigen Lösung im Unternehmensmaßstab zu entwickeln. Darüber hinaus besteht eines der Kernprobleme in der Schwierigkeit, Datenschutz im Nachhinein in eine Infrastruktur zu integrieren. Es sind vielmehr sog. Privacy-by-Design-Ansätze von Nöten, bei dem Datenschutz-Prinzipien zu Beginn eines Projekts oder spätestens bei der Datengenerierung mitgedacht werden. Doch selbst wenn diese Prinzipien klar formuliert sind, gibt es keine Garantie dafür, dass die Technologie den Datenschutz langfristig unterstützt. Z.B. entwickelt sich die sog. homomorphe Verschlüsselung, eine aufkommende Technologie für eine sichere Verarbeitung genetischer Daten, ständig weiter. Dies erschwert es Gendaten, die zu einem bestimmten Zeitpunkt verschlüsselt wurden, mit Daten neuerer Technologie-Versionen zu vergleichen. Außerdem sind Verschlüsselungstechnologien nicht unbedingt ideal für die langfristige Verwaltung von Daten, da z.B. Clouds und Quantencomputer extrem kostengünstig zu knacken sein könnten.

Der Druck wächst, die genetische Privatsphäre mit geeigneten Technologien und rechtlichen Regelungen für die Verwendung von Gendaten zu schützen. Was es braucht ist eine Kombination aus Hinweisen und Wahlmöglichkeiten, rechenschaftspflichtige Kontrollorgane über die Datenverwendung und echte – ökonomische wie imagewirksame – Strafen für die Schädigung von Individuen oder Gruppen. Zusätzlich könnte es erforderlich sein, sichere Datenbanken für spezifische Zwecke zu schaffen (z.B. Forschung oder Abstammung oder Strafjustiz), mit den jeweils geeigneten Mitteln zum Schutz der Privatsphäre und Wahlfreiheit der Betroffenen bei der Aufnahme in diese Datenbanken. Die Entwicklung eines so komplexen Systems wird nicht gänzlich glatt laufen und muss stets darauf reagieren, wie sich neue Datenschutzgesetze und -technologien auf Einzelpersonen und Gruppen auswirken. Einfache Lösungen werden weder ausreichen um einzelne Menschen und Bevölkerungsgruppen vor Schaden zu schützen, noch werden sie der Vergrößerung unseres Wissens über bessere Gesundheitsversorgung dienen können.

*Dieser Text ist ein übersetzter und stark gekürzter Wiederabdruck mit freundlicher Genehmigung von Springer Nature und den Autor*innen: Wan et al. (2022): Sociotechnical safeguards for genomic data privacy, In: Nature Reviews Genetics 23, S.429-445, https://doi.org/10.1038/s41576-022-00455-y. Übersetzung und Redaktion: Isabelle Bartram.*

- 1Sweeney, L./Abu, A./Winn, J. (2013): Identifying participants in the personal genome project by name (a re-identification experiment). Online: www.arxiv.org/abs/1304.7605 [letzter Zugriff: 03.08.2022].
- 2Gymrek, M./McGuire, A. L./Golan, D./Halperin, E./ Erlich, Y. (2013): Identifying personal genomes by surname inference. In: Science 339, S.321-324, www.doi.org/10.1126/science.1229566 [letzter Zugriff: 03.08.2022].
- 3Homer, N. et al. (2008): Resolving individuals contributing trace amounts of DNA to highly complex mixtures using high-density SNP genotyping microarrays. In: PLoS Genet. 4, 8, www.doi.org/10.1371/journal.pgen.1000167 [letzter Zugriff: 03.08.2022].

## Informationen zur Veröffentlichung

Erschienen in:
GID Ausgabe 262 vom August 2022
Seite 7 - 9