No. 262 – August 2022
8,50 Euro

information and criticism
on reproductive
and genetic engineering

GID

Gene Ethics Information Service

GENETIC PRIVACY

Risks and side effects of growing DNA data collections

A rye that belongs to all of us

Together the free ecological strengthen breeding

Small but mighty!

New insights into mutations

Blood test for trisomies is covered by health insurance

Surprisingly critical
Media response to the start of the practice

CONTENTS

Threats to genetic privacy

Current developments in health care, science and the Internet market

Genetic testing has led to a dramatic increase in the amount of genomic data being collected, used, and shared.

From **Dr. Zhiyu Wan, Dr. Dr. James W. Hazel ; Dr. Ellen Wright Clayton . Yevgeny Vorobeychik, Dr. Murat Kantarcioglu** and **Dr. Bradley A. Malin.**

The expansion of genetic data collections raises new and challenging concerns in Be privacy of all people - both legally and technically. In this article, we present existing and emerging threats to genetic privacy.

attacks on privacy

Much research involving genetic data is done in conjunction with datasets - demographic data, social and behavioral health determinants and measurements at the molecular or clinical level (e.g. from electronic health records) – performed with directly identifying information removed. However, there is a lively scientific debate about whether genetic data can be de-identified or de-anonymized alone or in combination with other forms of data. Over the years, a number of researchers have demonstrated their ability to identify individuals whose data has been used for genomic research without personal identifiers.

The use of genetic data at the individual level, even without personal data, includes the possibility of re-identification. For example, data recipients could derive phenotypic information from genetic markers and use this for re-identification. In one study, the researchers identified people by predicting external features such as eye and skin color from genome sequences. Similarly, conversely, genetic traits could potentially be derived from phenotypic traits (e.g. from certain diseases, external features or 3D facial structures) and used for identification purposes - although the actual validity of these analyzes is disputed.

In addition, it is possible to use demographic data, which is often shared with genetic data, to re-identify genetic data when combined with other easily accessible data sources. In 2013, participants in the Personal Genome Project were identified by Sweeney et al. who linked datasets to publicly available voter rolls based on demographic characteristics.(1) In the same year, Gymrek et al. certain participants in the 1000 Genomes Project by deriving their surnames from short tandem repeats (STRs) on the Y chromosome that they share with other demographics Combined data from public sources.(2)

In genotype-phenotype analyses, such as genome-wide association studies (GWAS), researchers usually only publish summary statistics. In 2008, Homer et al. however, that GWAS statistics are also prone to inferring the group membership of individual subjects.

(3) This means that it can be proven that the data of a known target person has been included in such a data set and it can also be deduced that this person belongs to a potentially sensitive group. The efficiency of this strategy was questioned by other scientists, but could be improved in subsequent studies by using additional statistical variables.

In addition, machine learning models trained with individual genetic data sets have the potential to reveal the genotypes and affiliations of the participants.

The similarity in genetic data between biological relatives allows their genotypes and predispositions to certain diseases to be inferred to some degree, even if their own genetic data was never collected. More recently, even better reconstruction strategies have been developed to derive the genotypes and phenotypes of individuals from data on their relatives. In April 2018, the US FBI used genetic data in an unsolved case to arrest the serial killer known as the Golden State Killer. The investigators entered the genetic profile of the suspect, who was still unknown at the time, in the public database GEDmatch.

Using what is known as a ***long-range familiar search,*** in which family members are identified by matching DNA sequences, they found a third cousin of the suspect. Based on this, a family tree with other family members could be reconstructed and the suspect himself then found.

Privacy in context

The sole focus on genomic research does not take into account the potential impact of the increasing availability of genetic data in other areas. A multitude of individuals and organizations are now collecting, using and disseminating genetic data on an unprecedented scale. As a result, this data is becoming an increasingly useful resource for various stakeholders such as employers, insurance companies, law enforcement agencies, etc. Numerous studies suggest that at least some people are concerned about where genetic data about them is going and how it is used become – with possible undesired and unexpected consequences. In addition to the frequently studied fears of discrimination, this information can also influence family relationships, e.g. by confirming or disproving biological paternity, searching for previously unknown relatives or identifying previously anonymous gamete donors. Concerns about the possible use of genetic data and the resulting consequences are usually formulated as a desire for genetic privacy. This can affect an individual's willingness to undergo clinical trials or participate in research projects. Such reluctance, due to privacy concerns, can in turn exacerbate existing health inequalities and hamper scientific knowledge.

So, when scientists plan, conduct, and discuss their studies, they need to consider how genetic data will be used and how how it is used affects whether the data can be controlled outside of the research setting.

Outside the research area

Millions of US citizens have purchased Direct-to-Consumer Genetic Testing (DTC-GT) from companies that provide personal information on a variety of topics, such as health, ancestry, familial relationships (eg, paternity), and lifestyle and well-being , promise. There are now numerous media reports on how consumers use this data to track down biological relatives – with complex consequences, both positive and negative. Some people rejoice at gaining new family members or learning their biological origins, while others worry about the results or unwanted contacts through unknown biological relatives. However, there are virtually no legal restrictions on how customers can use this data, although the resulting legal ramifications can be significant — including divorce and efforts to stop existing child support payments.

Also relevant for the protection of genetic data is that millions of people have downloaded their results from DTC-GT and published them in third-party databases to facilitate the search for relatives or to obtain health-related interpretations. These sites are rarely subject to any type of regulation beyond what they state in their terms of service. In addition, such websites reserve the right to change their practices, which may occur in response to public pressure, but also due to changes in business model. The affected datasets facilitate forensic use and probably pose the greatest risk for re-identification of genetic data.

forensic context

Law enforcement and their potential access to genetic information plays a major role in shaping public opinion about genetic data. High profile unsolved cases that were eventually solved using genetic data have sparked strong interest in this issue. In the United States, there have been efforts over the years to expand federal, state, and local government forensic databases. Law enforcement agencies may also seek to coerce disclosure of genetic information held by an individual or a company such as a healthcare provider, DTC-GT company, or researchers. In addition, investigators may also attempt to use public databases or use the services of a DTC-GT company for forensic genealogy purposes. To date, law enforcement in the US has largely focused on publicly available databases (e.g. GEDmatch) and private databases owned by companies that work voluntarily with law enforcement (e.g

FamilyTreeDNA). At the same time, there has been limited research on ways to reduce the privacy risks of uninvolved relatives posed by the search strategy of so-called forensic or investigative genetic genealogy.

Approaches to solutions for genetic data protection

An adequate level of protection for DNA data requires a combination of technical and societal solutions that take into account the context in which the data is used (Editor's note: more on possible solutions in the original article).
However, this goal is not easy to achieve.
From a technical point of view, it is challenging to transform data protection technologies that are published in scientific articles or tested in a small pilot study into a fully-fledged enterprise-scale solution to develop. In addition, one of the core problems is the difficulty of subsequently integrating data protection into an infrastructure. Rather, there are so-called Privacy-by-design approaches are necessary, in which data protection principles are taken into account at the beginning of a project or at the latest when data is generated. However, even when these principles are clearly stated, there is no guarantee that the technology will support data protection in the long term. For example, the so-called homomorphic encryption, an emerging technology for secure processing of genetic da constantly further. This makes it difficult to compare genetic data encrypted at a specific point in time with data from newer versions of technology. In addition, encryption technologies are not necessarily ideal for long-term management of data, as e.g. clouds and quantum computers could be extremely inexpensive to crack.

There is increasing pressure to protect genetic privacy with suitable technologies and legal regulations for the use of genetic data. What is needed is a combination of notices and choices, accountable controls over data use, and real – economic and image-effective – penalties for harming individuals or groups. In addition, there may be a need to create secure databases for specific purposes (eg, research *or* ancestry *or* criminal justice), with appropriate means of protecting the privacy and freedom of choice of individuals in inclusion in those databases. The evolution of such a complex system will not be entirely smooth and must always respond to how new privacy laws and technologies affect individuals and groups.

Simple solutions will not suffice to protect individuals and population groups from harm, nor will they increase our knowledge about better health care.

This text is a translated and heavily abridged reprint with the kind permission of Springer Nature and the authors: Wan et al. (2022): Sociotechnical safeguards for genomic data privacy, In: Nature Reviews Genetics 23, pp.429-445, https:// doi.org/ 10.1038/ s41576-022-00455-y. Translation and editing: Isabelle Bartram.

Notes and references: (1) Sweeney,
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 [accessed: 2022-08-03].

(2) Gymrek, M./McGuire, AL/Golan, D./Halperin, E./ Erlich, Y. (2013): Identifying personal genomes by surname inference. In: Science 339, pp. 321-324, www.doi.org/10.1126/science.1229566 [accessed: 08/03/2022].

(3) Homer, 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 [last access: 2022-08-03].