MULTI-REGIONAL CLINICAL TRIALS
BIOETHICS
COLLABORATIVE

The Multi-Regional Clinical Trials Center of Brigham and Women's Hospital and Harvard Bioethics Collaborative

Is It Time to Retire the Concept of Deidentification?

Thursday, December 12, 2024

Introduction:

In the U.S., clinical and research data containing the personal information of patients and/or participants are often protected through a process of deidentification. While definitions of "deidentification" vary, it is generally understood as the process of removing information that can be used to identify an individual, most commonly, so-called 'direct identifiers' such as name, address, date of birth, and the like.[1]

There are several regulations in the U.S. that govern the usage and protection of health information. Much data collected during health and research encounters is subject to the Health Insurance Portability and Accountability Act (HIPAA) Privacy Rule of 1996, which is a set of standards that aims to protect individually identifiable health information. There is a widespread belief that the HIPAA privacy rule permits data to be shared without consent under what is known as the 'safe-harbor' provision, which requires data sets to be stripped of 18 identifiers (see Appendix), and that patient authorization is no longer required following removal of the 18 identifiers. This view, however, is not actually correct because in order to be de-identified under HIPAA, the data must not identify a person, and there must be no reasonable basis to believe the data could be used to identify a person. Alternatively, deidentification can be accomplished through a formal determination by a qualified expert who determines that the health information is not individually identifiable.

In addition, data use and sharing for research may also be subject to the Common Rule, a federal policy in place to protect the rights, well-being, and privacy of human research participants. Importantly, the Common Rule only considers research with identifiable data to be human subjects research; research with deidentified datasets does not require ethical review or IRB approval, nor does it require informed consent for data use or risk-mitigation plans. The Common Rule defines identifiable information as "*identifiable private information is private information for which the subject's identity is or may readily be*

ascertained by the investigator or associated with the information." Thus, deidentified data can be understood as information for which the identity of the subject is *not* readily ascertainable, but the definition permits interpretation, particularly as the means to re-identify individuals becomes easier with evolving technologies, data availability, and data access.

A risk of research with deidentified data is the possibility of reidentification of individuals, compromise of privacy and confidentiality, and potential downstream harms. The likelihood of reidentification has changed over time, with the ubiquity of data collection, the availability of datasets from various sources that can be combined and interrogated, and the evolution of technologies such as artificial intelligence (AI). The HIPAA "safe harbor" does not appear to be so "safe," raising the question of whether and how investigators, IRBs, regulatory agencies, and others should consider identifiability, the risks to privacy and confidentiality, and the protections for human research participants that would be appropriate today. Notably, any change to the definition of "identifiable" information would likely have a profound impact on data-driven research, including the ability to reuse previously collected data and biospecimens to advance science and public health.

Presentations and Discussion:

After the introductory presentation, the group considered the concept of deidentification and the risks of reidentification. A member noted that while removing the 18 HIPAA safe harbor identifiers is often considered sufficient, the safe harbor provision includes a secondary clause, which states that the removal of the specified identifiers *"is adequate only if the covered entity has no actual knowledge that the remaining information could be used to identify the individual*." Additionally, for this statement to be true, there must be no reasonable basis to believe that the data could be re-identified. However, it is not clear how often or to what extent the latter clauses are, in fact, considered, but it would appear that the intent of the safe harbor, written almost 30 years ago, was to protect the identity of patients and their health information.

The first speaker acknowledged that just about any dataset can be susceptible to reidentification with sufficient effort but questioned whether this fact suffices to retire the concept of deidentification, given the practical challenges this would involve. The speaker noted the varying definitions of deidentification across multiple jurisdictions,[3] which lead to uncertainty about which standards to apply in what is often multi-national research. Further, IRBs and ethics committees have historically been concerned about the rights and interests of individuals but not the potential for community, group, or societal harms that might come

from reidentification.

The possible benefits and drawbacks of applying more stringent privacy protections than deidentification to data research were then discussed. A more stringent (protective) approach would enable the public to have not only a greater understanding of but also greater control over how their data is used, potentially increasing public trust. The possible drawbacks center around the practicality and feasibility of changing standards and, importantly, potentially chilling secondary research, given the greater burden of ascertaining consent. It would increase the work and burden of ethics committees, investigators, and study teams and increase the regulatory burden on investigators, but these latter considerations pale in comparison to concerns about tempering scientific progress, medical advancements, and health.

The discussion shifted to the assessment of the risks of reidentification and whether those risks are real or overstated. There is the risk *that* data will permit reidentification, but that risk varies based on the nature of the data, what other data are publicly available, and what harms could befall the individual should reidentification take place. The consequences of reidentification of someone who has eczema or high blood pressure differ from those of someone with psychiatric illness, rare disease, or substance use disorder. The likelihood of reidentification varies by the amount and types of data available on individuals, including the data that individuals themselves make available through social media and other means.

Another contributor called for collective public action to put an end to abusive data practices. As a precedent, the speaker pointed to events in the UK in 2012 in response to the UK Parliament enacting the Health and Social Care Act, which authorized the Health and Social Care Information Center to obtain National Health Service patient-identifiable data from general medical practices. Despite ethical protections--such as an option to opt out of being included, to restrict the use of one's data to certain legislatively defined aims, and a data access committee to vet applications for access--the public strenuously objected, leading to the end of the data repository in 2016. Critics charged the initiative with exploitation of patients and the public without adequate reciprocal benefits for them.

In general, it is reasonable to assume that the public expects and deserves a voice in how their data are used and in shaping the privacy policies that govern the use of their data. These expectations are unmet in the current system, where privacy and data-sharing policies are dictated by top-down, expert-led decisions and processes. Empirical evidence indicates that the public cares about how their data are used and want a voice in deciding which uses

to permit, even if data are deidentified.[3]

The HIPAA statute and its deidentification standard were revisited, with emphasis on the two discrete conditions that must be met for data to be considered deidentified: (1) the information cannot readily identify an individual, *and* (2) there should be no reasonable basis to believe it can be used to identify an individual. It was noted that much of what currently passes for deidentification fails to meet this standard, relying entirely instead on whether there is *actual* knowledge that data can be used to reidentify people rather than on whether there is a reasonable basis for believing reidentification is possible or likely. Once the data are deidentified, HIPAA's privacy protections no longer apply. If data are erroneously deemed to be deidentified (i.e., if data are deemed to be deidentified when, in fact, HIPAA's de-identification standard has not, in fact, been met), participants are wrongly denied protections to which they were legally entitled. Privacy rights – such as the right to be asked for authorization or protections against having one's data sold – are lost.

The group discussed how to ensure that the public has a meaningful voice and democratic control over policy regarding the sharing and research use of data. Achieving this in the near term would require pausing current policies and regulations—which, from a legal standpoint, would demand a court injunction—until changes can be implemented. Some states, such as California, are already adopting policy reforms through privacy rights acts, allowing consumers to take legal action against data privacy violations.

One member noted that this is complex in that the public must fully understand the issue in order to have informed views and preferences and play a role in these decisions. Others agreed, noting that fostering public trust requires a respectful approach to educating the public on the risks and benefits while also being willing to have faith in their collective judgment. An analogy to labor unions was used to highlight this point. Labor unions focus on collective issues and inequities rather than individual concerns, advocating for systemic solutions.

The last discussant addressed multiple ethical tensions between the goods of research and deidentification. First, data paired with contextual information is more valuable for research; therefore, the more information we remove from data sets, the less valuable they are. Stripping research data of phenotypic variants, for example, diminishes its utility and plausibly involves the loss of downstream public goods. Second, the speaker noted the difficulty of attempting to impose objective policy standards and protections on privacy problems, which have subjective import and vary in significance between individuals. The

privacy risks involved in sharing information are always contextual and vary based on the extent to which a condition or disease (for example) is stigmatized or accepted in the wider culture. The third tension highlighted how the risks of reidentification are inequitable, in the sense that they often disproportionately impact minoritized populations within datasets. This is due not only to the fact that minoritized populations are typically easier to reidentify within datasets, but also to the fact that they are already disadvantaged, such that harms are likely to compound and leave them worse off than non-minoritized groups who enjoy higher baseline standards of well-being. In addition, even apart from the risks of individual reidentification, the risks of community harm are more salient for minoritized groups, as the Havasupai Tribe research scandal illustrates.[4]

The concept of reciprocity was also discussed in relation to data research and deidentification. Whereas some might believe that patients often benefit from research and the contributions of others before them and, therefore, have a moral obligation to contribute their data to research, others, and minoritized groups in particular, can rightly point out that medical research has not benefited them to the same extent. When this may be taken as grounds not to participate in the present research, a downward ethical spiral can ensue, with non-participation limiting the generalizability of research and the lack of generalizability and lack of benefits for certain groups, in turn contributing to health disparities.

The session ended with a reflection on some possible paths forward. Participants reflected on institutions' role, such as seeking public and patient feedback on policies regarding data sharing and use. The speaker shared an example from their institution, which has established a data and specimen committee to review and assess whether patient consent has been obtained for data sharing. The institution has also launched a visible campaign to inform patients that their data and specimens may be used for research. The group agreed on the importance of making efforts to understand community preferences about data sharing and views about institutional data policies. Further, the positive role of data repositories and data enclaves was highlighted. These can limit bad actors from accessing and misusing data by setting restrictions on access via authorization requirements and encryption, among other measures.

The meeting ended with participants summarizing their thoughts and concerns in light of the far-ranging discussion. One member mentioned that there is a need for interim steps and a transitional framework before completely retiring the concept of deidentification or starting a privacy riot, given that regulatory changes can take years to effect. Another member emphasized the fear that too many privacy restrictions and protections might hinder

important research or, in some cases, stymie it altogether. Finally, the discussion touched on the importance of public transparency, education, and engagement, all of which will be needed to accomplish meaningful reform of data-sharing practices and balance meaningful privacy protections and safeguards against the promise and benefits of data research.

Appendix:

HIPPA Privacy Rule safe harbor method – 18 identifiers must be removed, and, in addition, there must be no actual knowledge that the remaining information could be used to identify an individual:[2]

- Names
- All geographic subdivisions smaller than a State, including street address, city, county, precinct, zip code, and their equivalent geocodes, except for the initial three digits of a zip code if, according to the current publicly available data from the Bureau of Census (1) the geographic units formed by combining all zip codes with the same three initial digits contains more than 20,000 people; and (2) the initial three digits of a zip code for all such geographic units containing 20,000 or fewer people is changed to 000.
- All elements of dates (except year) for dates directly related to the individual, including birth date, admission date, discharge date, date of death; and all ages over 89 and all elements of dates (including year) indicative of such age, except that such ages and elements may be aggregated into a single category of age 90 or older.
- Telephone numbers
- Fax numbers
- Electronic mail addresses
- Social security numbers
- Medical record numbers
- Health plan beneficiary numbers
- Account numbers
- Certificate/license numbers

- Vehicle identifiers and serial numbers, including license plate numbers
- Device identifiers and serial numbers
- Web Universal Resource Locators (URLs)
- Internet Protocol (IP) address numbers
- Biometric identifiers, including finger and voice prints
- Full face photographic images and any comparable images; and ® any other unique identifying number, characteristic, or code, except as permitted for re-identification purposes provided certain conditions are met.
- Any other unique identifying number, characteristic, or code

References:

1. *De-identified patient data*. (n.d.). Toolkit. https://toolkit.ncats.nih.gov/glossary/de-identified-patient-data

2. US Department of Health and Human Services. *Summary of the HIPAA Privacy Rule*. (2008, May 7). https://www.hhs.gov/hipaa/for-professionals/privacy/laws-regulations/index.html

3. Mello, M. M., Lieou, V., & Goodman, S. N. (2018). Clinical Trial Participants' Views of the Risks and Benefits of Data Sharing. *The New England journal of medicine*, *378*(23), 2202–2211. https://doi.org/10.1056/NEJMsa1713258

4. Rothstein, M. A. (2010). Is Deidentification Sufficient to Protect Health Privacy in Research? *The American Journal of Bioethics*, *10*(9), 3–11. https://doi.org/10.1080/15265161.2010.494215