Understanding the full continuum of identifiable to de-identifiable data in the world of big data to minimize privacy risk in database studies

Hye-Chung Kum[1,2,3], Cason Schmit[1], Theodoros Giannouchos[1]

[1] Texas A&M University Health Science Center, Department of Health Policy & Management, [2] Texas A&M University, Department of Computer Science & Engineering
[3] Population Informatics Research Group (https://research.tamhsc.edu/pinformatics/)

Background

- This is a descriptive study on the concept of identifiable data as it relates to human subject research (HSR) in light of the newly issued revisions to the Common Rule in January 2017.
- Federal laws restricting data from disclosure and use (e.g., for research purposes) typically only restrict identifiable data. Consequently, the legal definition of identifiable data is central to questions about whether data can be disclosed or used for research under different federal laws.
- Federal laws differ in how identifiability is defined and determined. Consequently, data that is legally identifiable under one federal law might not be legally identifiable under a different federal law, such as the Common Rule.
- The Common Rule regulates human subjects research, including research involving identifiable private information. The Common Rule was revised in January 2017 to include substantial changes to oversight in secondary research uses of identifiable private information.
- The Health Insurance Portability and Accountability Act of 1996 (HIPAA) and related regulations protect health information held by health care providers, health care clearinghouses and health plans. The regulations contain two different methods to de-identify information: expert determination and a safe harbor method involving the removal of 18 specified identifiers.
- Family Educational Rights and Privacy Act (FERPA) protects personally identifiable information in education records.
- The regulations in 42 CFR Part 2 protect identifiable information held by covered substance abuse disorder treatment programs

Methods

- We conduct a conceptual analysis on the different definitions of identifiable data in various regulations. We limited our analysis to the following laws: the existing Common Rule, revised Common Rule (effective 2018), HIPAA, FERPA, and 42 CFR Part 2) as they relate to human subject research to understand how these regulations are appropriately interpreted during the IRB review process to minimize risk.
- Each legal framework was evaluated on whether the regulatory definition was broad v. narrow and specific v. vague.
- The broadness of a legal definition turns on a conceptualization of how much detail in a record must be present for the record to be legally identifiable. Narrow definitions allow more detail to be considered non-identifiable than broad definitions.
- The specificity of a legal definition signifies the extent that a law contains specified types of data that is deemed identifiable (i.e., lists of identifiers, and specified excluded or included data)

Conclusion

- The most narrow definition of identifiable data is one used in the Common Rule which states, "the identity of the subject is or may be readily ascertained by the investigator or associated with the information". The revised common rule does not change this core definition, but it adds more specificity, including categorical exemptions for certain data uses (e.g., public health surveillance).
- On the other extreme HIPAA defines "individually identifiable health information," in part, as "...information that identifies the individual, or with respect to which there is a reasonable basis to believe the information can be used to identify the individual."
- In practice, HIPAA covered entities consider data identifiable under HIPAA if it contains any of 18 Safe Harbor identifiers, including indirect identifiers (e.g., birthdate, zipcode). The HIPAA expert determination de-identification method lacks the specificity of the HIPAA Safe Harbor method, but this vagueness can lead to confusion in implementation.
- Identifiable data under the 42 CFR Part 2 regulations is defined using broad language (i.e., "information by which the identity of a patient … can be determined with reasonable accuracy either directly or by reference to other information)", but only provides a few examples of identifiers in a non-exhaustive list (i.e., name, address, SSN, fingerprints, photographs).
- FERPA provides specific, but non-exhaustive, lists of direct and indirect identifiers. The regulations also contain two broader and less-specific categories of identifiable information: 1) "information that… would allow a reasonable person … to identify the student with reasonable certainty," and 2) where there is a reasonable belief a person requesting information "knows the identity of the student"

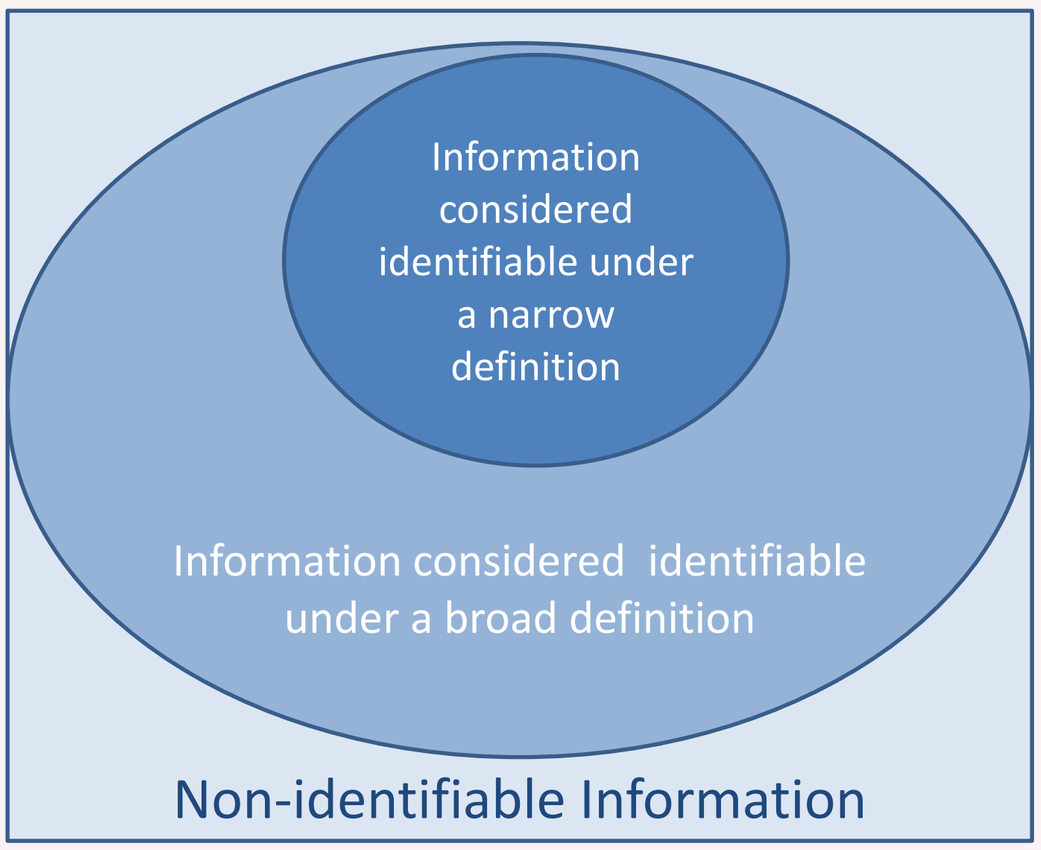

Figure 1 – Effect of broad versus narrow legal definitions on the quantity of data deemed non-identifiable

Understanding the full continuum of identifiable to de-identifiable data in the world of big data to minimize privacy risk in database studies

Hye-Chung Kum[1,2,3], Cason Schmit[1], Alva O. Ferdinand[1]

[1] Texas A&M University Health Science Center, Department of Health Policy & Management, [2] Texas A&M University, Department of Computer Science & Engineering
[3] Population Informatics Research Group (https://research.tamhsc.edu/pinformatics/)

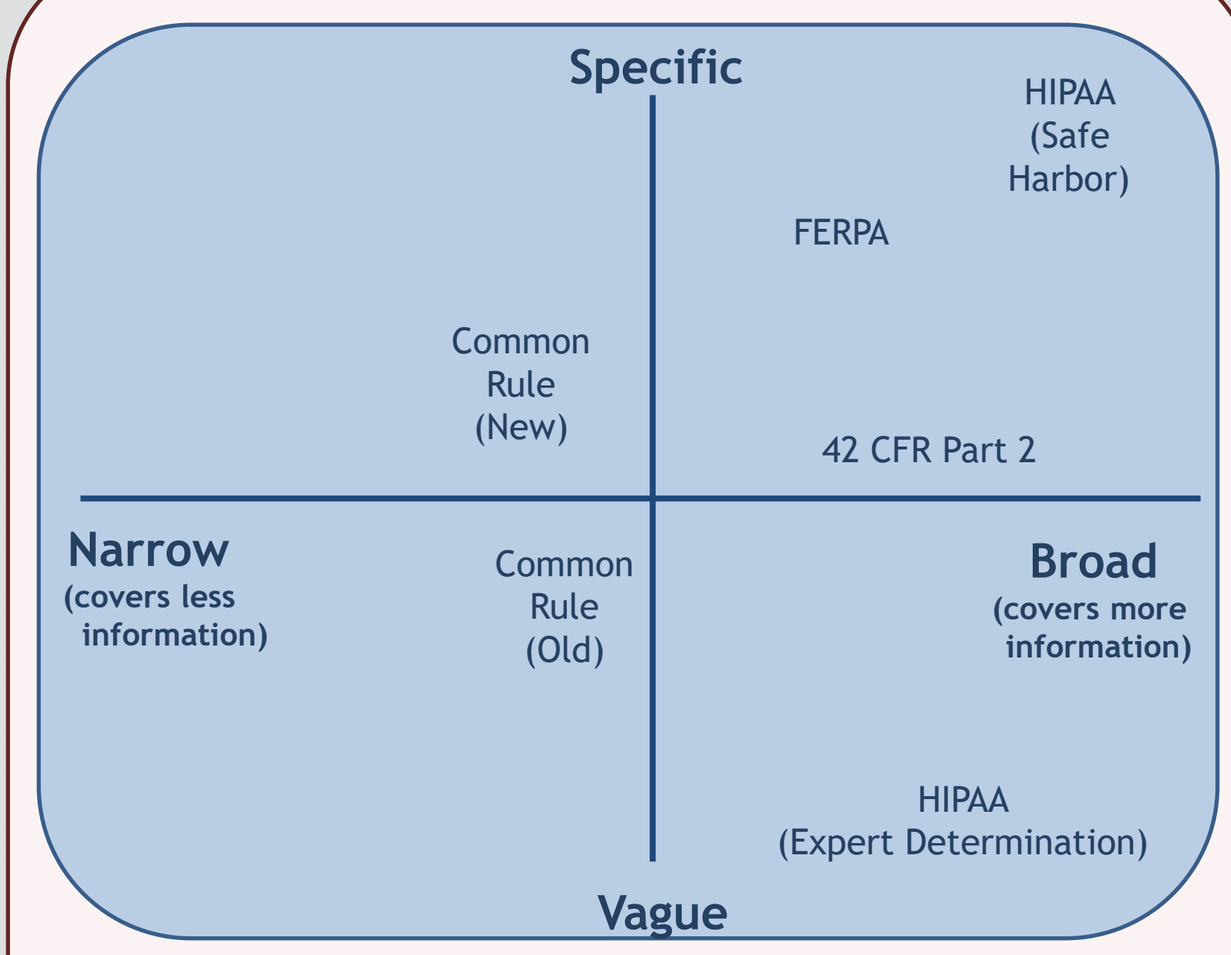

Figure 2 – Depiction of two dimensions of legal definitions of identifiable information (broadness and specificity)

Pitfalls of Definitional Inconsistencies

- Many data used in database studies fall in between the HIPAA and Common Rule definitions of identifiable data (e.g., limited datasets), often confusing the determination of HSR or exempt status of database studies where HIPAA regulations are also applicable.
- The SEER-Medicare data available from CMS is a good example. It contains dates of services, an indirect identifier listed in HIPAA that classifies the data as identifiable PHI.
- However, with many direct identifiers removed from limited datasets, investigators cannot readily ascertain the identities of subjects in limited datasets.
- Thus, many IRBs, including NIH's Office of HSR, have classified some research with limited datasets, such as the SEER-Medicare data, as exempt research under the common rule (45 CFR 46.101(b)(4)).
- IRBs that incorrectly apply the HIPAA identifiability standard in place of the Common Rule standard severely and unnecessarily inhibit database research by placing increased scrutiny on critical data.

Limitations

- This project only investigates identifiability of real datasets. [These two limitations still confuse me. We only discuss one real dataset (SEER) as an example. I think we can generalize this. See below.
- Generating realistic synthetic data is an active area of research in information privacy. Sometimes called derived data, the question of how different synthetic data must be from real data to be classified as de-identifiable data is an open research question. [These two limitations still confuse me. I am not clear where synthetic data is incorporated in the earlier sections. This seems like it will hit the reader out of the blue]
- This is a conceptual analysis of statutes and regulations. Judicial opinions and federal agency guidance can provide additional guidance for similar factual situations or for specific datasets (e.g., SEER data).
- Reasonable minds will sometimes differ in legal analysis. The graphical representations of the dimensions of legal identifiability are intended to illustrate the differences between laws. However, the continuum of identifiablity is highly dependent on the specific factual context, relevant judicial opinions, and regulatory guidance.
- This material is intended to be educational and is not a substitute for legal advice.

Discussion

- Identifiability is central to IRB determinations of whether research is exempt or even if it involves “human subjects.”
- For objectivity, IRBs, and not PIs, must make this determination. Yet, identifiability standards are often misunderstood by IRBs that can delay the IRB review process.
- Moreover, confusion between identifiable data under different legal standards (e.g., HIPAA and Common Rule) can result in difficult, inaccurate, or inappropriate assessments of privacy risk.
- Given the recent increase in research using big data about people, it is critical for IRB staff to appreciate the full continuum between identifiable and de-identifiable data to make the correct determinations and minimize risk.

References

- Federal Policy for Protection of Human Research Subjects (Common Rule). See 45 C.F.R. § 46.102.
- The Health Insurance Portability and Accountability Act of 1996 (HIPAA). See 45 C.F.R. § 160.103
- Family Educational Rights and Privacy Act (FERPA). See 34 C.F.R. § 99.3
- Confidentiality of Substance Use Disorder Patient Records, 42 C.F.R. Part 2. See 42 C.F.R. § 2.11