Session 3: Privacy and Confidentiality Challenges in Health Survey Research

Chair: Stephen Blumberg (CDC/NCHS) / Brad Edwards (Westat)
Discussant: Doug Williams (Westat)
Rapporteur: Lisa Mirel (CDC/NCHS); Jeanette Ziegenfuss (HealthPartners)

After four papers were reported, the discussant highlighted three challenges related to participant privacy assurances including respondent beliefs, contrasting protections (between consent designed to protect the respondent and terms of services designed to protect the company or organization) and respondent's general lack of knowledge about when, where and who is collecting data as well as how it is shared. It was acknowledged that myriad experiences outside of a given researcher's or researchers' control in general influence a potential participants privacy concerns. The primary strategy comprised of mathematical mechanisms for reducing privacy concerns is Differential Privacy (DP). DP comes from the computer science framework and involves the calculation of epsilon to estimate the "acceptable privacy loss" or the balance between privacy and the accuracy of the data. Group discussion followed and included citing new/reinforcing strengths and limitation of DP, public sentiment on privacy, confidentiality and further considerations. The text below highlights each one of these areas.

Strengths of DP:

- Permits more public use of data by protecting confidentiality thereby potentially allowing more data to be available to researchers.
- Allow for data to be released that may be critical in answering health related questions.
- Best for Census data and large datasets. Not necessarily needed when sampling (and subsampling) is used as that is a form of privacy protection in itself.
- Other large data sources include Twitter. The use of Twitter data has greatly increased for health-related research in recent years.

Limitations of DP:

- Use of DP could engender overconfidence of data privacy by researchers and potentially result in unnecessary/over-collection of data. However, it is difficult to communicate what DP is to potential respondents. Respondents seem to be more concerned with hacking and data breach threats rather than re-identification. The approach to communicate that an individual's data will not be intact in the database may have the unintended consequence of leading people to believe that they need not participate in the research at all.
- The storing of data and a data breach is not addressed by DP.
- There is not a one size fits all approach for DP and subject matter experts should be involved in the process.
- It is difficult to obfuscate certain types of data with DP such as medical claims data or some types of economic data. DP does not address issues with consent to link data, hierarchical and/or nested data. It is not clear how it would be used when studying rare diseases but may not be as much of an issue since there are fewer sources with this information out there.
- Not clear how the need for transparency is met when using DP. If one needs to release data and methods for transparency, doesn't the objective of DP fail?
- If DP is implemented, and a study participant has self-identified, then people may infer wrong opinions, behaviors, etc. The area of self-identification may need to be addressed within the consent process.

- Use of sources like Twitter may be limited in their representativeness (note: 80% of tweets come from 10% of users)

Public sentiment on privacy and security:

- Up to 31% of people report not being willing to put their data at privacy risk. But, asking about willingness is difficult particularly when the context while important, is hard to explain. It is not good practice to ask people opinions about something where they likely do not have an opinion already formed. If necessary, to do so, vignettes can help. Typically, respondents are most worried about a neighbor or spouse finding out information about them. A good reference for this is Contextual Integrity by Nissenbaum (1998).
- Generally, people seem to act in conflicting ways.
    - Individuals want their data to be private and protected, but think that it is a lost cause and they have no control.
    - Individuals do not want people they know to see their private information but have less concern over people they do not know or corporations gaining access to this information.
- People want their data used for the right purpose in the case of medical records and Fitbit/other tracking devices. They just do not want it used for other purposes or in ways they do not know.
- Most people are willing to give data to trusted entities, so we need to increase level of trust if we want to increase access to data.
- We are in a state of competing demands.

Further considerations:

- DP is not going away. DP is from the computer science field. Terms such as histograms imply data tables and cross tabs. One question to consider is if DP provides over confidence in privacy?
- Some of the techniques being implemented are similar to a randomized response approach.
- When data are released there may be unforeseen usages. Should we be considering data enclaves and validation servers? Or do we just ask people to agree to make all their data public?
- Individuals self-disclose research participation on social media – or otherwise, not realizing that in doing so they risk reidentification. It is important that this risk is communicated in consent or otherwise.
- Generally, we make too broad of assurances as there is much about assuring privacy that is outside of our control and is truly an issue of data security. This was noted in an example where an NHANES participant self-identified by posting on social media after completion of the NHANES medical exam.
- We need to determine just how much we need to know for decision making.
- Generally, we make too broad of assurances as there is much about assuring privacy that is outside of our control and is truly an issue of data security.