June 27, 2022

Authors, Chris Barker, Ph.D, Monica Johnson, M.S.

Dear Statistical Consulting Section and all ASA colleagues,

As chair-elect of the Statistical Consulting Section, I present my motivation for one of my several forthcoming Section initiatives. I will outline other initiatives when I am section chair. The short working title for my note below: "Finding the de-anonymization Needle in the SEER Haystack." SEER is the Surveillance and Epidemiology End Results database administered by the Division of Cancer Control and Population Sciences (DCCPS) at the National Cancer Institute (NCI). My initiative arises because I needed a crash course in concepts entirely new to me about data privacy, anonymization/de-anonymization and identification/de-identification and re-identification) and statistical disclosure. Based on what I learned in my crash course, I am inviting an interested group of statisticians to help develop a "data privacy toolbox" that members of the consulting section and certainly any/all statisticians at ASA can use in day-to-day work. The toolbox may be used at a leisurely pace rather than a crash course. Volunteers need not be Section members, though I encourage joining the Section. Defining and measuring success of the toolbox is an additional objective for the group.

Privacy in the 21st century may no longer exist. Bill Gates stated, "That historically, privacy was almost implicit, because it was hard to find and gather information." (Wired, 2013). Today, de-anonymization (i.e., re-identification), the practice and relative ease of "adversaries" of matching anonymized data (i.e., De-identified data) with publicly available, or auxiliary data, may lead to identifying that "anonymous" person in terms of actual name address, employer, etc.. The ethics of de-anonymization and implications to those de-anonymized has been studied by independent ethicists and other experts including the US Census Bureau. One 'data use" ethics issue is the unambiguous violation of explicitly-stated Terms of Use (TOU) for Surveillance and Epidemiology End Results (SEER) and clinicaltrials.gov (CTG) database. The specific paper in a prominent economics journal, clearly states the researchers and organizations linked SEER and CTG with no reference to the TOU (SEER 2022). The authors themselves, and the current Nobel prize winning Journal Editor (Dr. Duflo) and Nobel prize winning president of the America Economic Association (AEA) Dr. Card, when asked, did not provide proof that the authors had permission of any kind from the Federal agencies overseeing SEER and CTG to violate TOU. This creates a risk "adversaries" re-identify oncology patients by linking to auxiliary data. Statisticians working with any data from humans may need to update their understanding of anonymization, de-anonymization and statistical disclosure.

The Needle in the SEER Haystack

Paraphrasing an article (Wired, 2007) about anonymization, much we may have learned about anonymization of patient data and "statistical disclosure" may be completely outdated or possibly wrong. SEER data are "anonymized," and information permitting identification (name, address, credit cards, salary, etc.) of individual patients has been removed. However, in 2008 Netflix created the NETFLIX prize and provided anonymized customer data to the public. A team of computer scientists were able to link NETFLIX database with Internet movie database (IMDb), identify the Netflix client's actual name and address, and received confirmation of correctly identifying the individuals from Netflix management (IEEE, 2008). A critical caveat emptor to my work here is that there is no direct proof of a de-anonymization, since that can only be achieved by directly contacting the patients involved. Briefly, I inspected (using SAS and R) datasets prepared by the authors, available for public download by anyone with internet access and not password protected or any method to track the download. I found forty (40) unique clinical trials with exactly one patient (sample size n=1) linked to SEER patient level data with a large number of covariates that can be used by an adversary to link to auxiliary data for de-anonymization. I have no way to contact the individual patients and I turned over my discovery entirely to the experts at SEER and CTG. The patient data I found is at potentially very high risk of de-anonymization. Given the detailed data in both CTG, investigator, institution, and many variables in SEER, in principle it may be possible to de-

anonymize the forty unique cases of patients in clinical trials n=1. As a courtesy, I specifically informed the Federal DCCPS and NLM director and their privacy experts at SEER and CTG that I did not expect or need to know how the matter was handled. And I specifically asked the authors and journal editor and AEA president to carry out the turnover – in order to be in compliance with the SEER TOU to notify SEER of de-anonymizations. In the absence of their replies, I intervened, guided by the ethical principles of ASA and reported the matter to SEER and CTG.

Proof of Concept of De-anonymization of SEER using certain CTG trials

My background is in pharmaceutical clinical trials where we routinely blind patients, investigator and sponsor. Anonymization and de-anonymization differ from blinding and unblinding. The two share a common characteristic that individual patient identifiers are removed by an anonymization algorithm, sometimes referred to in the privacy literature as "catch and release." The two differ in that de-anonymization may occur only for a single patient, several patients, or possibly all patients. Clinical trial Unblinding is applied to all patient data, at one time "database unlock." The patient identifiers are not included in publications in journals publication of pharmaceutical clinical trials data by the EMA (2019), European transparency laws (GPRDS).

I believe I have discovered the first-ever publication of the "proof of concept" (POC) for de-anonymization algorithm in a prominent peer reviewed journal of economics policy, the "American Economic Review." (AER). The POC of "de-anonymization" of SEER anonymized data is caused by combining two crown Jewel datasets of the Federal statistical system, SEER and CTG in violation of the TOU. In fact, as many as four Federal level databases may be involved. Clinicaltrials.gov in certain small number of situations has a type of "patient level" data, specifically a small number of clinical trials where the final sample size (number of patients) is one (n=1). Based on my experiences, I pose a broader question; I do not attempt to answer. How important is the discovery of a "proof of concept" to say, the on-going initiatives, by pharmaceutical companies to provide datasets of pristine anonymized clinical trial data to external experts. And does the POC increase the risk that some of those experts may attempt to "de-anonymize" the data. Lastly, I do not attempt to answer the question whether the POC "scales up" to larger clinical trials. And I completely turned over the matter to DCCPS/SEER and NLM/CTG for the privacy experts to address the matter.

Synergism with other ASA committees

At the outset I recognized that the concept of a data privacy toolbox may overlap with the work of other ASA committees. To avoid duplication of effort, I invite Consulting Section members and other ASA committees, such as Data Privacy, Record Linkage, Epidemiology, and Ethics to collaborate on this initiative.

My remaining concern is with discovery of a "proof of concept" paper in the peer reviewed literature, and the possible encouragement that the possibility of the first published algorithm, and violation of TOU, some very specific circumstances, uses two "Big Data" - clinicaltrials.gov and SEER to de-anonymize or place at high risk of de-anonymization exactly forty patients that are present in both these big data. Lastly by "big data" I refer to the fact that as of about June 14 2022, clinicaltrials.gov has 418,148 clinical trials, and SEER captures approximately 400,000 cases of cancer annually and stores data on about 30% (about 99 million) of the US population (332,403,650 as of 2022)

SEER Research Data use agreement, accessed July 8, 2022

https://seer.cancer.gov/data-software/documentation/seerstat/nov2021/seer-dua-nov2021.html

Wired, 2013. https://www.bartleby.com/essay/Privacy-Of-The-21st-Century-No-Longer-PKKHYPQ3PVDX

Narayanan, Arvind, and Vitaly Shmatikov. "Robust de-anonymization of large sparse datasets." *2008 IEEE Symposium on Security and Privacy (sp 2008)*. IEEE, 2008.

Bruce Schneier, Why 'Anonymous' Data Sometimes Isn't, Wired DEC 12, 2007,

EMA https://www.ema.europa.eu/en/human-regulatory/marketing-authorisation/clinical-data-publication