BDK20 Exercise

Suppose you have two datasets – a voter registration list and records of patients with diabetes. Using these two sources, how easy is it to re-identify patients?

Download the data file BDK20\_Dataset.xlsx which contains two sheets: “Diabetes Database” and “Voter Registration” (note: these are synthetic datasets and are not meant to reflect any actual information; names were created from a random generator, as were the zip codes and dates). Use the data to answer the following questions.

1. Can you uniquely re-identify any of the patients in the diabetes database from the voter registration list? Are there any similar records? (That is, multiple patient records that map to one registered voter, or multiple voters that map to the same patient record.)

Noah Barrett: Male, zip = 97630, DOB = August 8, 1937.

1. If you were able to identify anybody, how sure are you that this is the same person? What additional information do you think you would need to be sure?

You would need to know if there are any other males born on August 8, 1937 who live in the 97630 zip code, other than those in the data sets.

1. Ideally, patient information wouldn’t be re-identifiable. What information do you think should be removed from the diabetes database to protect the patients’ health information?

Less specificity in birthday or location would be an improvement.

In 1997, Governor William Weld of Massachusetts was identified from supposedly de-identified data by an MIT graduate who cross-referenced a publicly available state insurance database with a Massachusetts voter registration list that she purchased for $20. She did this by linking zip code, gender, and date of birth (Sweeney, 1997). Current HIPAA legislation suggest that zip codes should be 3 digits instead of 5, and that only birth year should be given rather than full date of birth (http://www.hhs.gov/ocr/privacy/hipaa/understanding/coveredentities/De-identification/guidance.html).

Now download the dataset BDK20\_Dataset\_AfterHIPAA.xlsx. This is the same data as before, but zip codes and dates now follow current HIPAA practices.

1. Using the new datasets, can anyone be identified? Are there any similar records?

Many one-to-one records, but it is much less likely that they are actually correct matches. For example, the 3 digit zip code 976 is a frequent one-to-one match, but there are over 20 zip codes that start with 976.

(Many similar records, too many to list.)

1. Do you think that the use of 3 digit zip codes and birth years is enough to sufficiently de-identify patient data? In your opinion, should more information be hidden and if so, what?