

Ellen Wright Clayton, MD, JD  
Vanderbilt University

# **Data Access v. Confidentiality: Balancing Risks and Benefits**

# What are the risks?

- ⓘ Re-identification

- § Why would someone want to do this?

- § What is the likelihood that someone would seek to re-identify?

- § What would be the impact of re-identification?

- ⓘ The case of hacking

- ⓘ The case of returning individual results

# What are the risks?

## Social science data

- ⓘ Social science data can be stigmatizing
- ⓘ Why is there a risk of re-identification?
  - § Existence of other publicly available databases
  - § Risk of re-identification depends on
    - Extent to which the research data set is redacted
    - Number of people in the research dataset
    - Completeness of the comparison dataset(s)

# What are the risks?

## Genetic data

- ⓘ Existence of known collections of identified DNA
  - § Identified research datasets
  - § Pathology laboratories
  - § Forensic
  - § Military
  - § Paternity determination – private and public
- ⓘ Vary in terms of existing analysis

# What are the risks?

## Genetic data

- ! Identification that someone is in a study might lead one to infer phenotype
- ! The ability to determine whether a particular individual is in a dataset depends on:
  - § The number of people in the study ( $\uparrow$  number  $\rightarrow$   $\downarrow$  power)
  - § The number of SNPs reported ( $\uparrow$  number  $\rightarrow$   $\uparrow$  power)
  - § Minor allele frequency ( $\uparrow$  number  $\rightarrow$   $\uparrow$  power)
  - § Willingness to accept misidentification ( $\uparrow$  tolerance for error  $\rightarrow$   $\uparrow$  power)
- ! Homer N, Szelling S, Redman M, Duggan D, Tembe W, et al. (2008) Resolving Individuals Contributing Trace Amounts of DNA to Highly Complex Mixtures Using High-Density SNP Genotyping Microarrays. PLoS Genet 4(8): e1000167. doi:10.1371/journal.pgen.1000167

## Homer, et al.

“it is now clear that further research is needed to determine how to best share data while fully masking identity of individual participants. However, since sharing only summary data does not completely mask identity, greater emphasis is needed for providing mechanisms to confidentially share and combine individual genotype data across studies, allowing for more robust meta-analysis such as for gene-environment and gene-gene interactions.”

# Balancing risks and benefits – two approaches

- ⌚ Limiting access to data
  - § Redacting data
  - § Firewalls with oversight
- ⌚ Controlling misuse
  - § Role of NIH
  - § Role of investigator's institution