The Community as a Learning System for Health:
Using Local Data to Improve Community Health

National Center for Health Statistics, CDC
May 12, 2011

Are there limits to privacy preserving sharing of data?

Staal A. Vinterbo

Associate Professor, Division of Biomedical Informatics

UCSD

Take Home Point

Context: Sharing of patient information for research.

A general and purely technological solution to privacy preserving sharing of patient data might not be possible.

Current State

- Sharing of data
  - Complete (needs oversight)
  - Limited data set ("almost" de-identified, needs oversight)
  - De-identified data
- De-identification by HIPAA standard
  - Safe Harbor (removal of 18 predefined information items)
  - Statistical Standard (expert declares data re-identification risk as "very small")

Problems

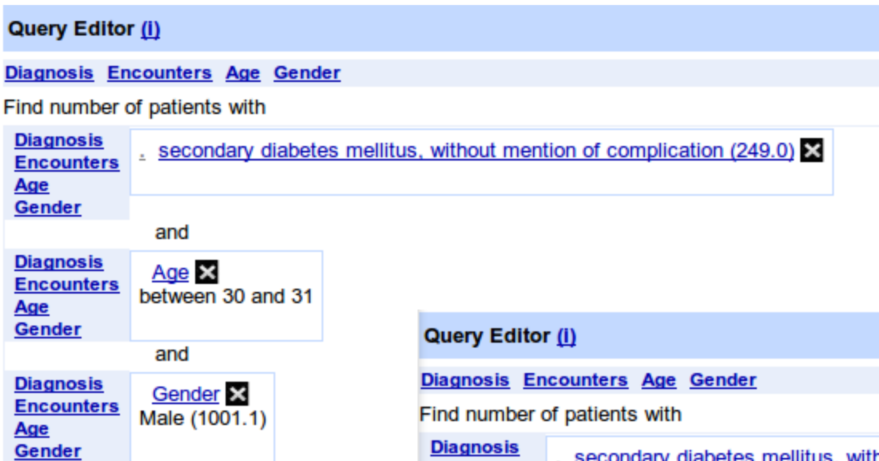
- Oversight (IRB)
  - Costly (administration, time)
    - Researcher: write IRB protocol and wait for approval
    - Institution: process protocol and administrate it
  - Difficult across institutions

Problems

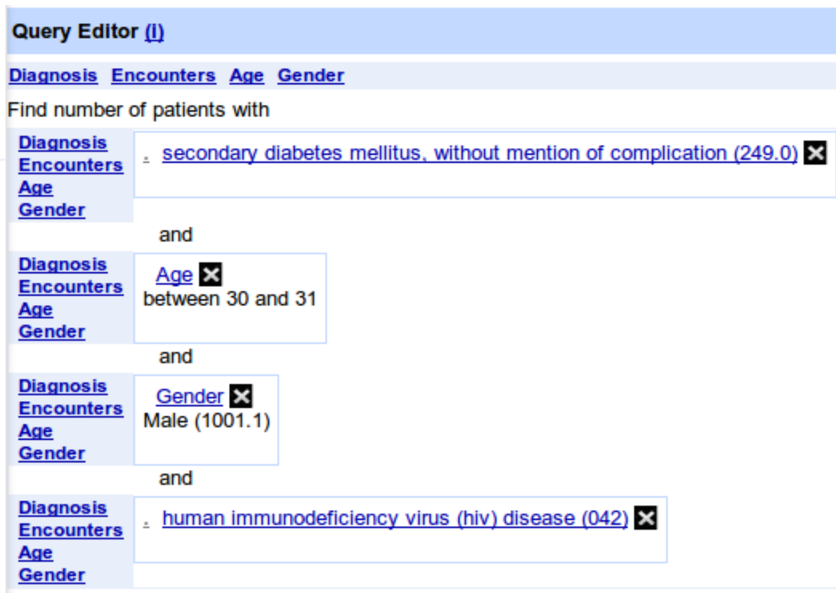
- De-identification
  - by Safe Harbor yields data with limited utility[1]
  - by Safe Harbor does not prevent re-identification[2]
  - by Statistical Standard is vaguely defined:
    - "A person with appropriate knowledge" [...] "determines that the risk [of re-identification] is very small"
- Inferences about sensitive information can be made without re-identification

[1]Beyond the HIPAA Privacy Rule: Enhancing Privacy, Improving Health Through Research. IOM 2009
[2]The disclosure of diagnosis codes can breach research participants' privacy. Loukides G, Denny JC, Malin B. J Am Med Inform Assoc. 2010 May 1;17(3):322-7

Insufficiency of de-identification:
inferences about known individuals



= 3



= 3

(We know that neighbor Bob is 30 and has secondary diabetes)

Points of Discussion

- Are insufficiencies of de-identification too esoteric to be of practical concern?
  - Is heuristic and empirical risk assessment[1,2] convincing?
    - "we were able to re-identify x %": not a valid upper bound!
  - Can we use media attention as a guide?
    - Note:
      - HITECH breach reporting does not apply to de-identified data
      - There are no tracking requirements for de-identified data

[1]Evaluating re-identification risks with respect to the HIPAA privacy rule. K Benitez, B Malin. *JAMIA 2010;17:169-177*
[2]A method for managing re-identification risk from small geographic areas in Canada. El Emam et al. *BMC Medical Informatics and Decision Making* 2010, **10:**18

A possible alternative view?

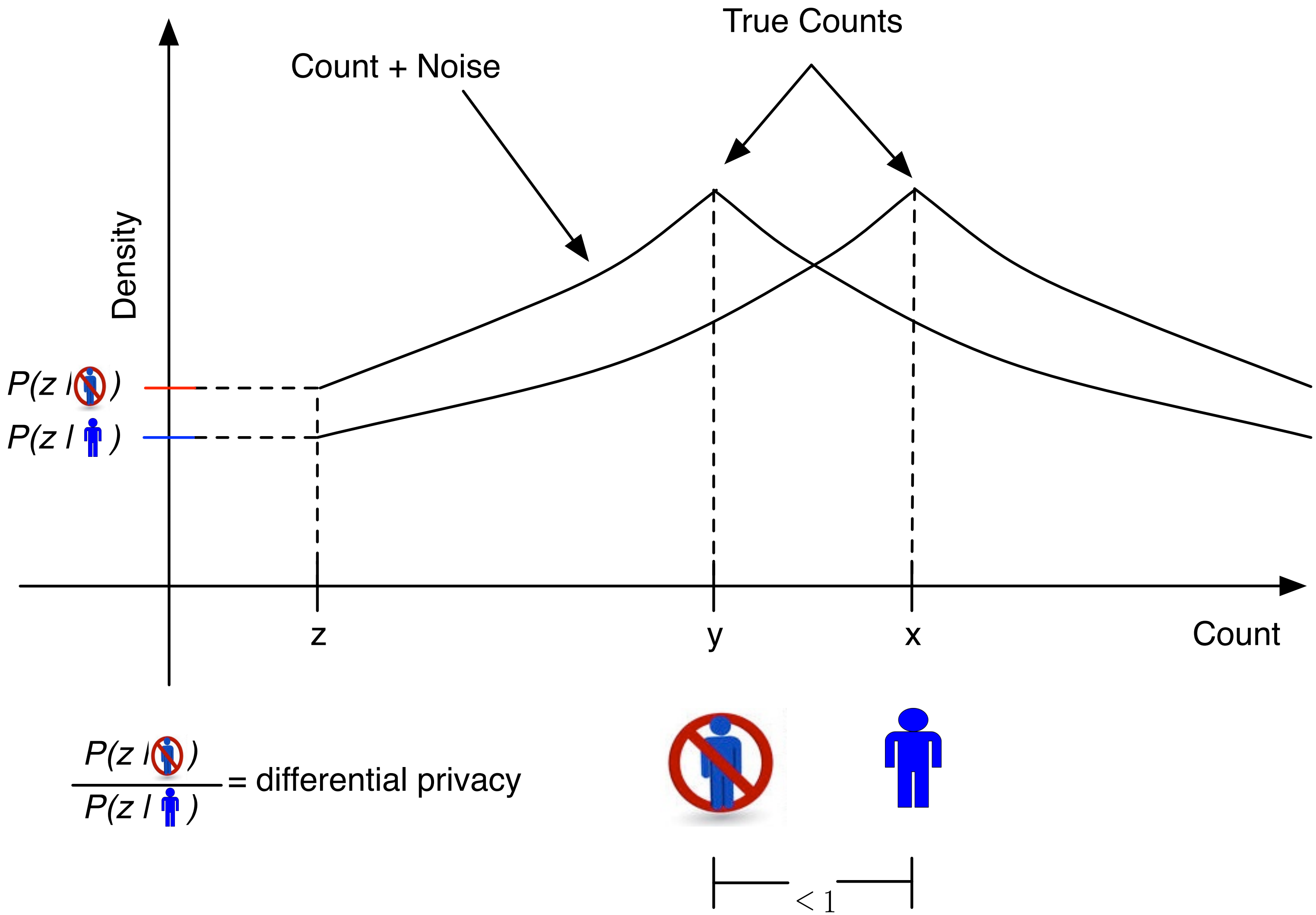
- Ideal for individual privacy: "*information is privacy preserving if what can be learned about any **individual** is independent of this information*"
- Consequence: we are allowed to share information about *populations.*
- Implies de-identification
- Complete independence not feasible: requires infinite populations

Towards the ideal: Differential Privacy

- Differential Privacy* bounds the change in likelihood of learning anything about an individual by his inclusion in the data
- Is a property of an access method (as opposed to a property of data)
- Access methods to data that provably guarantee differential privacy exist

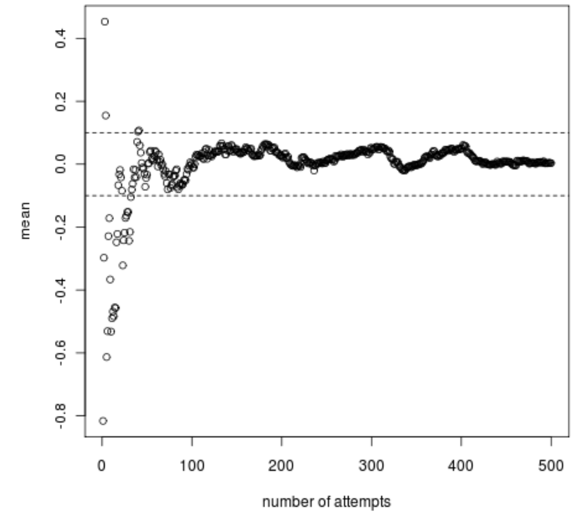
*Dwork, C.; McSherry, F.; Nissim, K. & Smith, A. Calibrating Noise to Sensitivity in Private Data Analysis Proceedings of the Conference on Theory of Cryptography, 2006

Differential Privacy and Noisy Counts



The Finite Privacy Budget

Example: Adding noise to counts does not Protect against averaging multiple trials



Suggests* a general property of a finite "privacy budget": only small # of privacy preserving accesses can be allowed, beyond which privacy can no longer be guaranteed

Increase information about each patient: decrease in budget!

*New Efficient Attacks on Statistical Disclosure Control Mechanisms. Cynthia Dwork and Sergey Yekhanin CRYPTO 2008

Dealing with a finite budget

- Use all allowed information accesses up front to extract all privacy preserving information
  - Never allow privacy preserving access again
  - Different uses might need different information
  - Very high-dimensional data: budget very small
- Leverage environment to “extend budget”
  - Principle: substitute some “treatment” (punishment) for some “prevention”
  - Requires:
    - Detection of misuse and perpetrator
    - Effective sanctioning of perpetrator (aka. “teeth”)

Conclusion

- De-identification as a definition of privacy seems insufficient for believable privacy
- Current theoretical research suggests that there are limits to truly privacy preserving sharing of data using technological means alone

Acknowledgements

- Collaborators:
  - Kamalika Chaudhuri
  - Anand Sarwate
  - DBMI/iDASH members
- Support
  - NIH R01 LM07273
  - NIH U54 HL108460