An Overview
Heritage Health Prize

Jonathan Gluck

General Counsel, Heritage Provide Network

The De-identification of the Heritage Health Prize Data Set

Khaled El Emam

CEO, Privacy Analytics

Senior Scientist, CHEO Research Institute

Canada Research Chair, University of Ottawa

Outline

- General observations
- About the data
- Technical issues that we faced
- If we were to do it again

Reasonableness Criterion

- "Health information that does not identify an individual and with respect to which there is **no reasonable basis** to believe that the information can be used to identify an individual is not individually identifiable health information."
- "... generally accepted statistical and scientific principles ..."
- " .. the risk is **very small** that the information could be used, alone or in combination with **other reasonably available inform**

Data Set

Age (members)	Date of claim (claim)
Sex (members)	Diagnosis (claim)
Days in Hospital (Outcome)	Length of stay (claim)
Specialty of provider (claim)	Provider ID (claim)
Place of service (claim)	Vendor ID (claim)
CPT Code (claim)	Pay delay (claim)

Technical Issues

- "very small" was defined as a maximum probability of a single record being re-identified of 0.05
- At the outset removed patients with highly sensitive values
- Evaluated matches with California voter registration list and SID
- The problem of correlated domains
- Truncation of outliers with a large number of claims
- The concept of adversary power for longitudinal data
- The concept of patient diversity
- We used the OLA algorithm to optimally generalize and suppress
- Sub-sampling was used to provide some contingency
- Additional perturbation to protect provide confidentiality (not really a privacy issue)

Simulated Attacks

Power	5	10	15
Original	0.84%	0.94%	1.17%
Multiple	3.67%	3.72%	3.87%
Ordered	0.96%	1%	1.2%

An adversary with a power of 15 will know more than 100 pieces of information about an individual accurately

Matching with SID (%)

Age	LOS	Sex	# of Visits	PCG	CPT	Year 1	Year 2	Year 3	All Years
X	X	X	X			0.161	0.147	0.151	0.514
X	X	X			X	0.71	0.568	0.596	0.973
X	X	X		X		1.333	1.015	1.092	1.357
X	X	X		X	X	1.727	1.270	1.379	1.599

kelemam@uottawa.ca

www.privacyanalytics.ca
www.ehealthinformation.ca