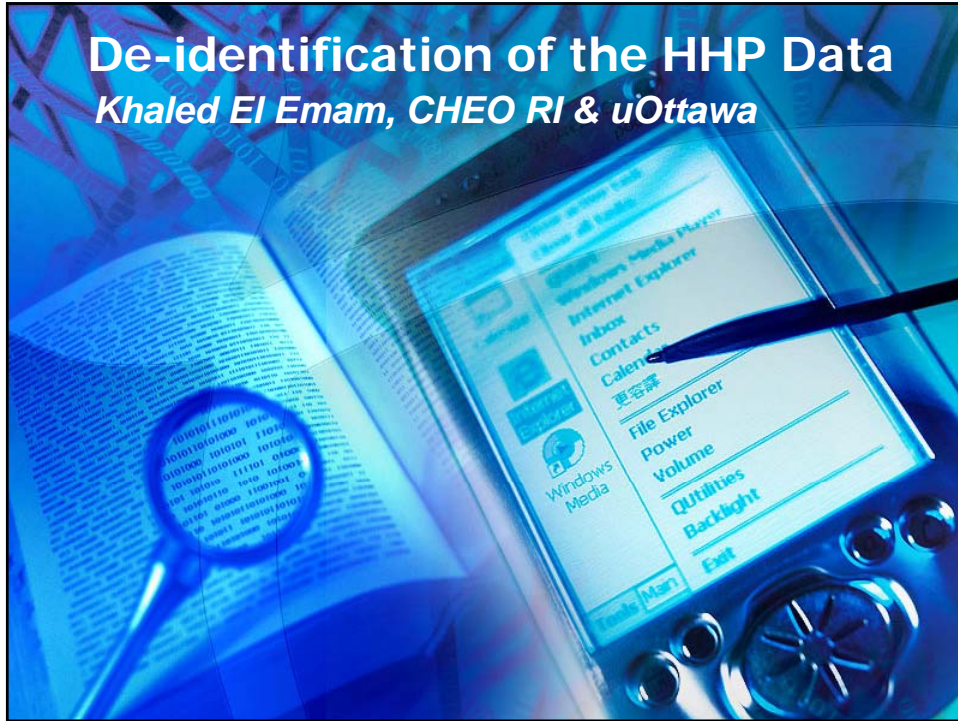


# De-identification of the HHP Data

*Khaled El Emam, CHEO RI & uOttawa*



## Today's Presentation

- Provide overview of rationale and methods used to de-identify the HHP data set, as well as lessons learnt
- The complete details have been published in a recent article in JMIR:  
<http://www.jmir.org/2012/1/e33/>
- Address questions from different communities:
  - entrants in the competition
  - disclosure control community
  - other competition organizers

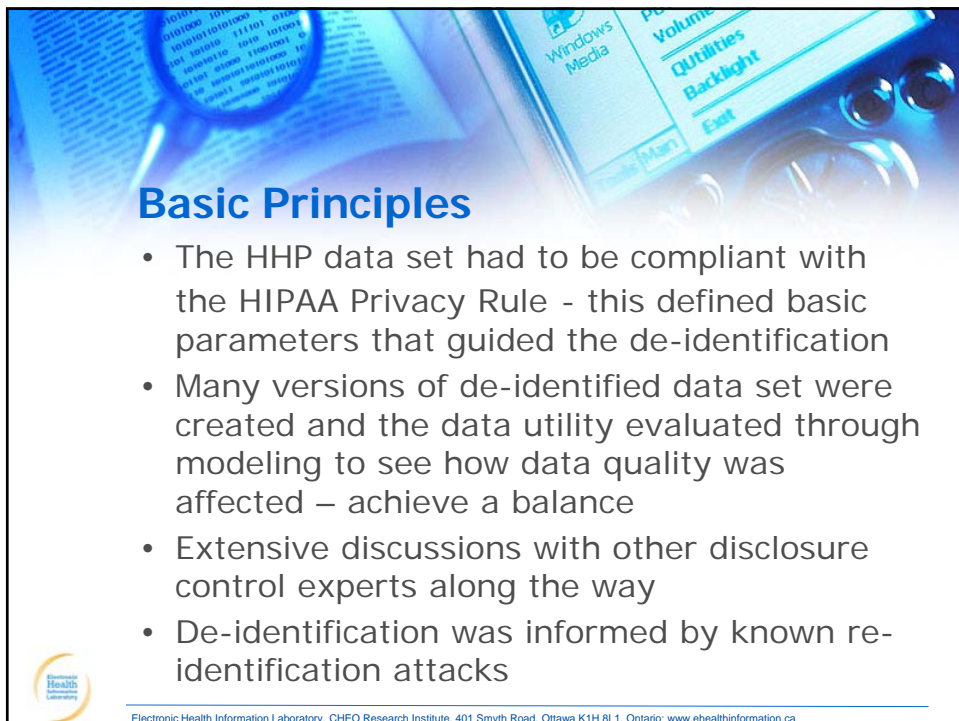




## Caveats


- Certain manipulations are not revealed
- We do not represent HPN or Kaggle – questions about the competition rules should be posted on the HHP forum for the Kaggle team to respond to

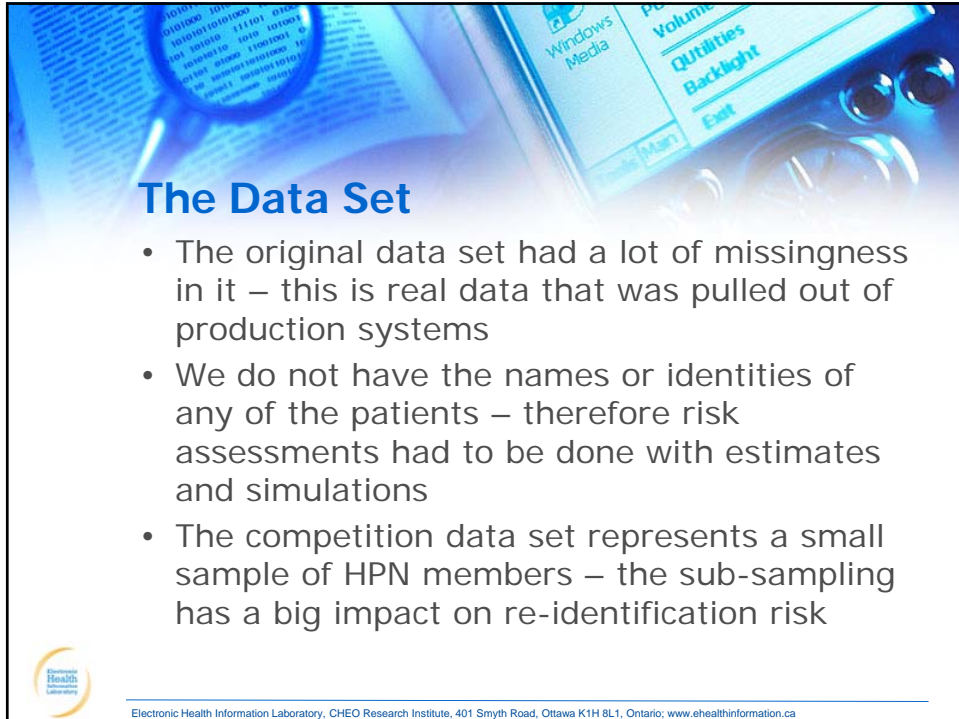
 Electronic Health Information Laboratory, CHEO Research Institute, 401 Smyth Road, Ottawa K1H 8L1, Ontario: [www.ehealthinformation.ca](http://www.ehealthinformation.ca)



## Basic Principles


- The HHP data set had to be compliant with the HIPAA Privacy Rule - this defined basic parameters that guided the de-identification
- Many versions of de-identified data set were created and the data utility evaluated through modeling to see how data quality was affected – achieve a balance
- Extensive discussions with other disclosure control experts along the way
- De-identification was informed by known re-identification attacks

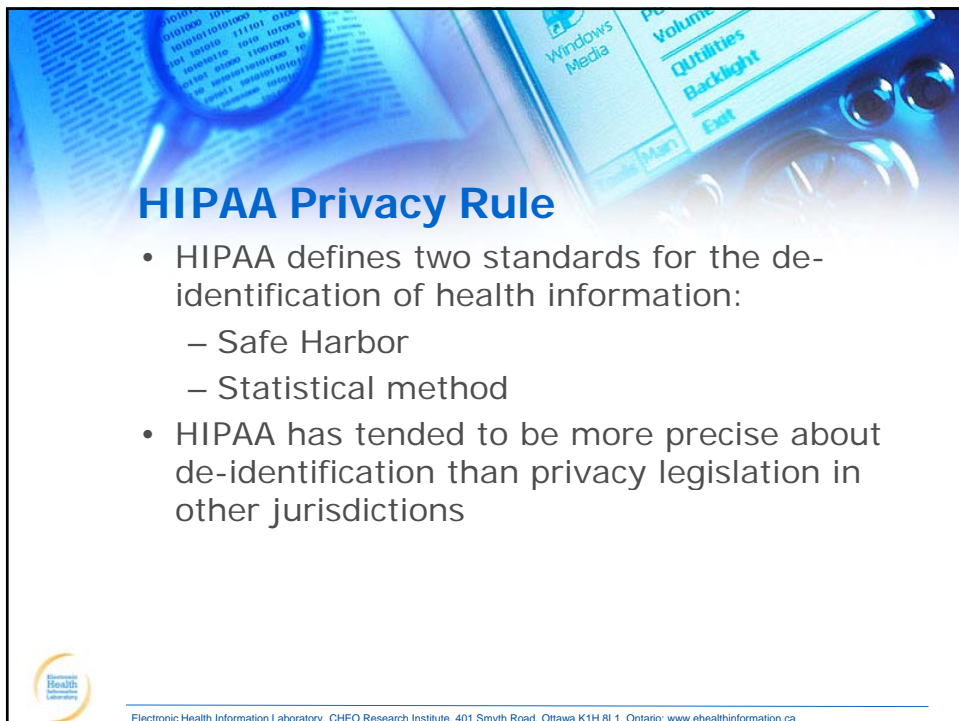
 Electronic Health Information Laboratory, CHEO Research Institute, 401 Smyth Road, Ottawa K1H 8L1, Ontario: [www.ehealthinformation.ca](http://www.ehealthinformation.ca)



## The Data Set


- The original data set had a lot of missingness in it – this is real data that was pulled out of production systems
- We do not have the names or identities of any of the patients – therefore risk assessments had to be done with estimates and simulations
- The competition data set represents a small sample of HPN members – the sub-sampling has a big impact on re-identification risk

 Electronic Health Information Laboratory, CHEO Research Institute, 401 Smyth Road, Ottawa K1H 8L1, Ontario, [www.ehealthinformation.ca](http://www.ehealthinformation.ca)



## HIPAA Privacy Rule

- HIPAA defines two standards for the de-identification of health information:
  - Safe Harbor
  - Statistical method
- HIPAA has tended to be more precise about de-identification than privacy legislation in other jurisdictions

 Electronic Health Information Laboratory, CHEO Research Institute, 401 Smyth Road, Ottawa K1H 8L1, Ontario, [www.ehealthinformation.ca](http://www.ehealthinformation.ca)

# HIPAA Safe Harbor

## Safe Harbor Direct Identifiers and Quasi-identifiers

<ol style="list-style-type: none"> <li>1. Names</li> <li>2. ZIP Codes (except first three)</li> <li>3. All elements of dates (except year)</li> <li>4. Telephone numbers</li> <li>5. Fax numbers</li> <li>6. Electronic mail addresses</li> <li>7. Social security numbers</li> <li>8. Medical record numbers</li> <li>9. Health plan beneficiary numbers</li> <li>10. Account numbers</li> <li>11. Certificate/license numbers</li> </ol>	<ol style="list-style-type: none"> <li>12. Vehicle identifiers and serial numbers, including license plate numbers</li> <li>13. Device identifiers and serial numbers</li> <li>14. Web Universal Resource Locators (URLs)</li> <li>15. Internet Protocol (IP) address numbers</li> <li>16. Biometric identifiers, including finger and voice prints</li> <li>17. Full face photographic images and any comparable images;</li> </ol>	<ol style="list-style-type: none"> <li>18. Any other unique identifying number, characteristic, or code</li> </ol>
--	--	--



Electronic Health Information Laboratory, CHEO Research Institute, 401 Smyth Road, Ottawa K1H 8L1, Ontario: [www.ehealthinformation.ca](http://www.ehealthinformation.ca)

# HIPAA Safe Harbor

## Safe Harbor Direct Identifiers and Quasi-identifiers

<ol style="list-style-type: none"> <li>1. Names</li> <li>2. ZIP Codes (except first three)</li> <li>3. All elements of dates (except year)</li> <li>4. Telephone numbers</li> <li>5. Fax numbers</li> <li>6. Electronic mail addresses</li> <li>7. Social security numbers</li> <li>8. Medical record numbers</li> <li>9. Health plan beneficiary numbers</li> <li>10. Account numbers</li> <li>11. Certificate/license numbers</li> </ol>	 <ol style="list-style-type: none"> <li>12. Vehicle identifiers and serial numbers, including license plate numbers</li> </ol>	<ol style="list-style-type: none"> <li>13. Device identifiers and serial numbers</li> <li>14. Web Universal Resource Locators (URLs)</li> <li>15. Internet Protocol (IP) address numbers</li> <li>16. Biometric identifiers, including finger and voice prints</li> <li>17. Full face photographic images and any comparable images;</li> <li>18. Any other unique identifying number, characteristic, or code</li> </ol>
--	---	---



Electronic Health Information Laboratory, CHEO Research Institute, 401 Smyth Road, Ottawa K1H 8L1, Ontario: [www.ehealthinformation.ca](http://www.ehealthinformation.ca)



## 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 information**, by an anticipated recipient to identify an individual who is a subject of the information .. ”



Electronic Health Information Laboratory, CHEO Research Institute, 401 Smyth Road, Ottawa K1H 8L1, Ontario: [www.ehealthinformation.ca](http://www.ehealthinformation.ca)



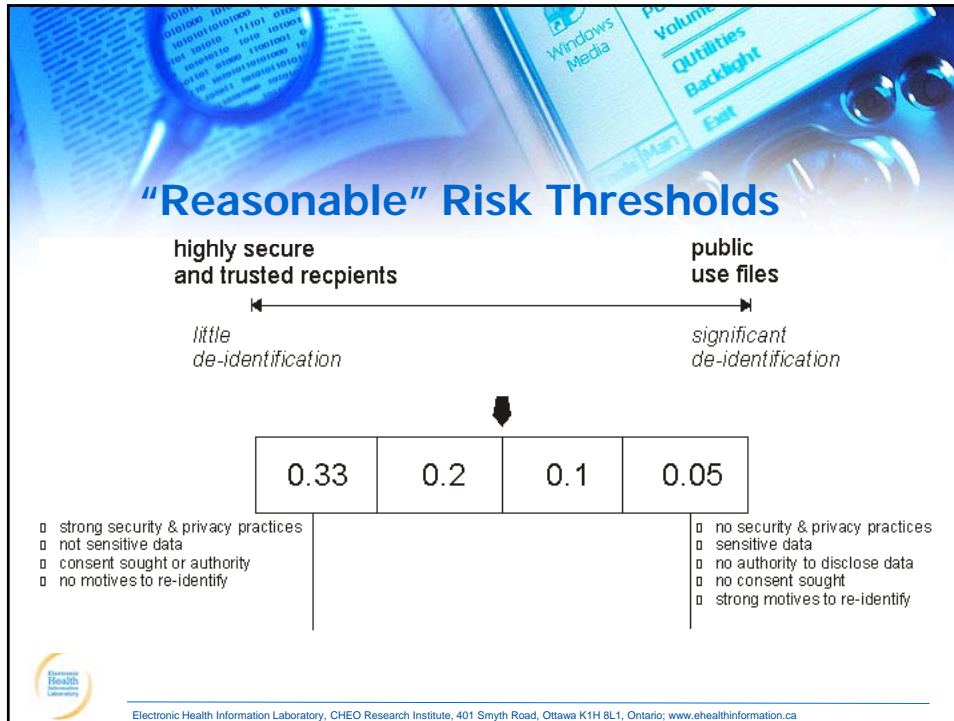
## Statistical Method

- Need to ensure that the risk of re-identification is very small

$$\left( \frac{1}{N} \sum_i I(\theta_i \leq \tau) \right) \geq \alpha$$



Electronic Health Information Laboratory, CHEO Research Institute, 401 Smyth Road, Ottawa K1H 8L1, Ontario: [www.ehealthinformation.ca](http://www.ehealthinformation.ca)



**Precedents - I**

- The value of  $\tau$  represents the probability that a record can be correctly re-identified
- There are many precedents for setting this value to 0.2, 0.1, and 0.05 for the public release of health data (as well as other types of data)
- For the HHP data it was decided to err on the conservative side and use a threshold value of 0.05
- This is under ideal conditions – real value likely lower

Electronic Health Information Laboratory, CHEO Research Institute, 401 Smyth Road, Ottawa K1H 8L1, Ontario, www.ehealthinformation.ca



## Precedents - II

- HIPAA Safe Harbor estimated risk is that 0.04% of the population is unique:

$$\left( \frac{1}{N} \sum_i I(\theta_i \leq 1) \right) \geq 0.9996$$



Electronic Health Information Laboratory, CHEO Research Institute, 401 Smyth Road, Ottawa K1H 8L1, Ontario: [www.ehealthinformation.ca](http://www.ehealthinformation.ca)



## Risk Exposure

$$\text{Risk Exposure} = \text{Loss} \times \text{Probability}$$

- In the case of Safe Harbor:

$$\text{Risk Exposure} \leq N \times 0.0004 \times 1$$

- Equivalent HHP risk exposure:

$$\text{Risk Exposure} \leq N \times 0.008 \times 0.05$$



Electronic Health Information Laboratory, CHEO Research Institute, 401 Smyth Road, Ottawa K1H 8L1, Ontario: [www.ehealthinformation.ca](http://www.ehealthinformation.ca)



## Risk Management

- Ensure that no more than 0.8% of members have a probability of re-identification greater than 0.05
- A combination of technical and legal approaches used to manage the overall risk
- Legal limits:
  - Prohibition on re-identification
  - Agreements with HPN service providers (e.g., labs and insurers)



Electronic Health Information Laboratory, CHEO Research Institute, 401 Smyth Road, Ottawa K1H 8L1, Ontario: [www.ehealthinformation.ca](http://www.ehealthinformation.ca)



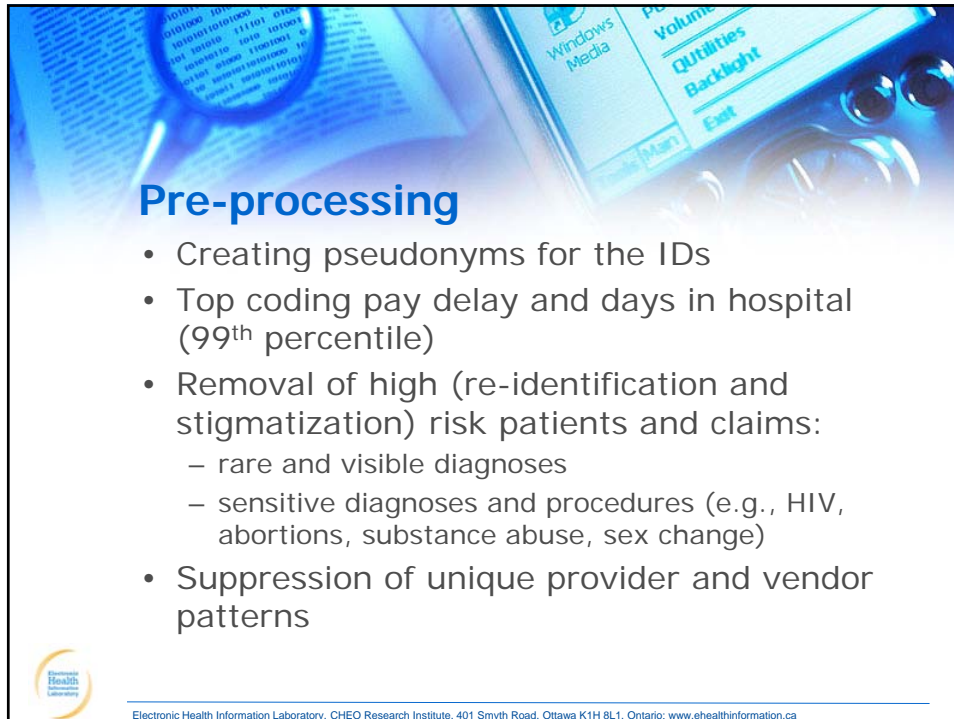
## Data Set

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)




Electronic Health Information Laboratory, CHEO Research Institute, 401 Smyth Road, Ottawa K1H 8L1, Ontario: [www.ehealthinformation.ca](http://www.ehealthinformation.ca)

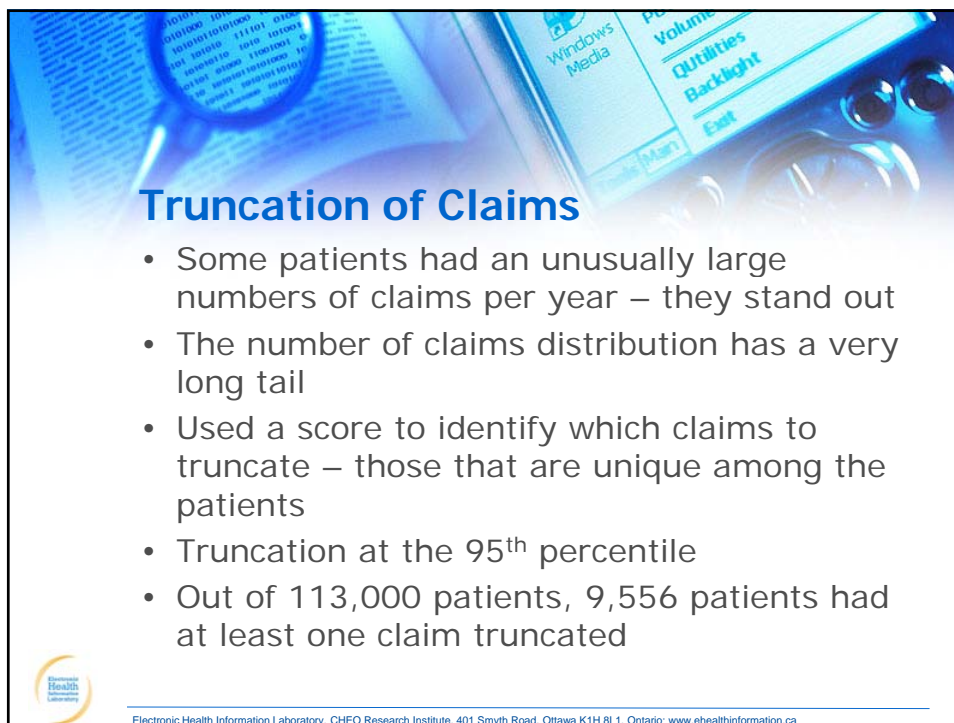




## Pre-processing


- Creating pseudonyms for the IDs
- Top coding pay delay and days in hospital (99<sup>th</sup> percentile)
- Removal of high (re-identification and stigmatization) risk patients and claims:
  - rare and visible diagnoses
  - sensitive diagnoses and procedures (e.g., HIV, abortions, substance abuse, sex change)
- Suppression of unique provider and vendor patterns

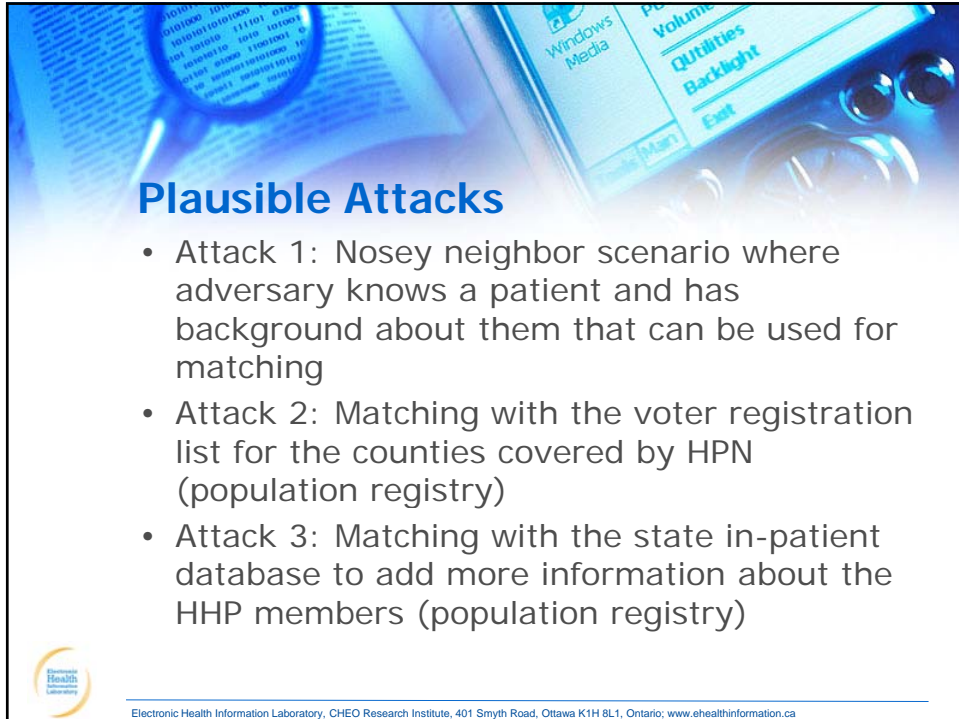
 Electronic Health Information Laboratory, CHEO Research Institute, 401 Smyth Road, Ottawa K1H 8L1, Ontario: [www.ehealthinformation.ca](http://www.ehealthinformation.ca)



## Truncation of Claims


- Some patients had an unusually large numbers of claims per year – they stand out
- The number of claims distribution has a very long tail
- Used a score to identify which claims to truncate – those that are unique among the patients
- Truncation at the 95<sup>th</sup> percentile
- Out of 113,000 patients, 9,556 patients had at least one claim truncated

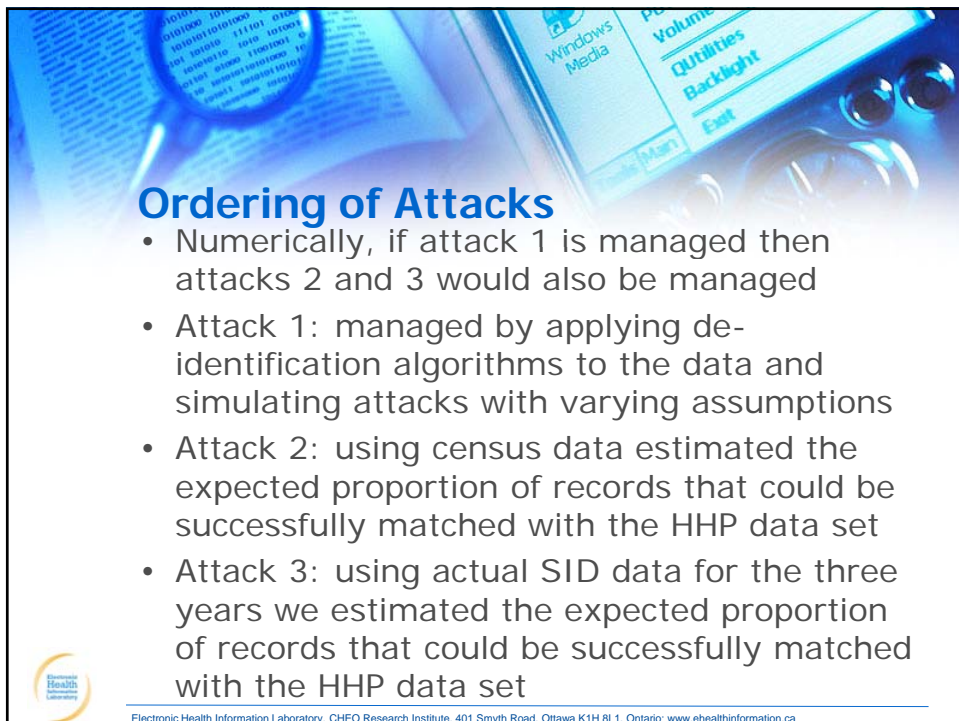
 Electronic Health Information Laboratory, CHEO Research Institute, 401 Smyth Road, Ottawa K1H 8L1, Ontario: [www.ehealthinformation.ca](http://www.ehealthinformation.ca)



## Plausible Attacks


- Attack 1: Nosey neighbor scenario where adversary knows a patient and has background about them that can be used for matching
- Attack 2: Matching with the voter registration list for the counties covered by HPN (population registry)
- Attack 3: Matching with the state in-patient database to add more information about the HHP members (population registry)

 Electronic Health Information Laboratory, CHEO Research Institute, 401 Smyth Road, Ottawa K1H 8L1, Ontario, [www.ehealthinformation.ca](http://www.ehealthinformation.ca)



## Ordering of Attacks

- Numerically, if attack 1 is managed then attacks 2 and 3 would also be managed
- Attack 1: managed by applying de-identification algorithms to the data and simulating attacks with varying assumptions
- Attack 2: using census data estimated the expected proportion of records that could be successfully matched with the HHP data set
- Attack 3: using actual SID data for the three years we estimated the expected proportion of records that could be successfully matched with the HHP data set

 Electronic Health Information Laboratory, CHEO Research Institute, 401 Smyth Road, Ottawa K1H 8L1, Ontario, [www.ehealthinformation.ca](http://www.ehealthinformation.ca)

## Generalizations

- Practical approach to reduce the probability of re-identification that has advantages over other common approaches
- Examples: diagnosis codes to primary condition groups & Charlson index and procedure codes to higher level codes

d, ○ {year}

d, ○ {month/year}

d, ○ {day/month/year}

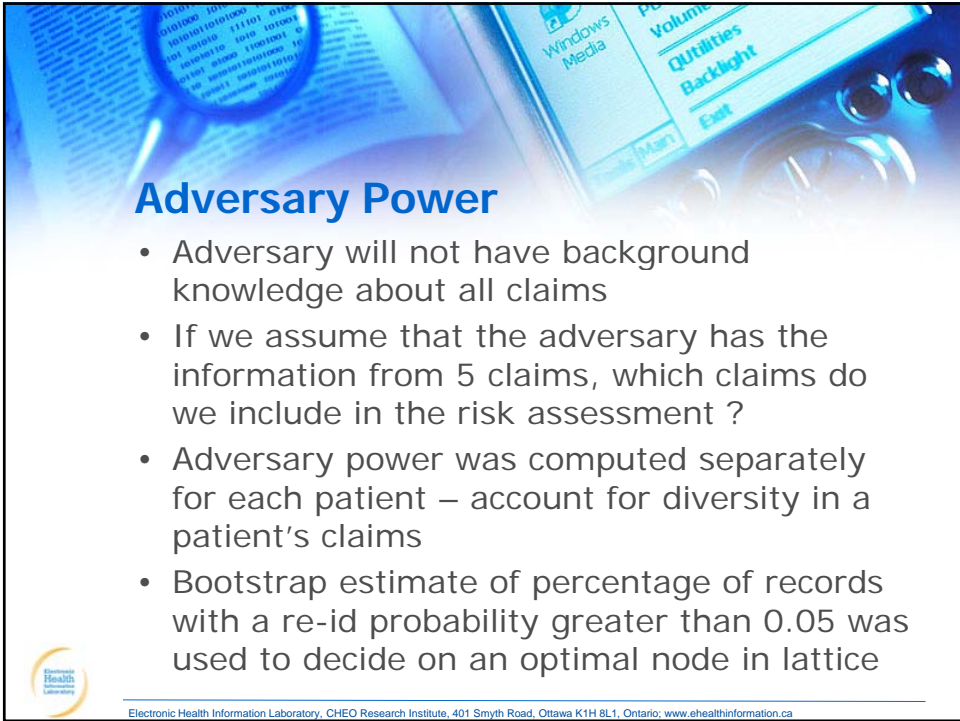
g<sub>2</sub> ○ {person}

g<sub>1</sub> ○ {M/F}

Electronic Health Information Laboratory, CHEO Research Institute, 401 Smyth Road, Ottawa K1H 8L1, Ontario, www.ehealthinformation.ca


## Optimal Generalizations

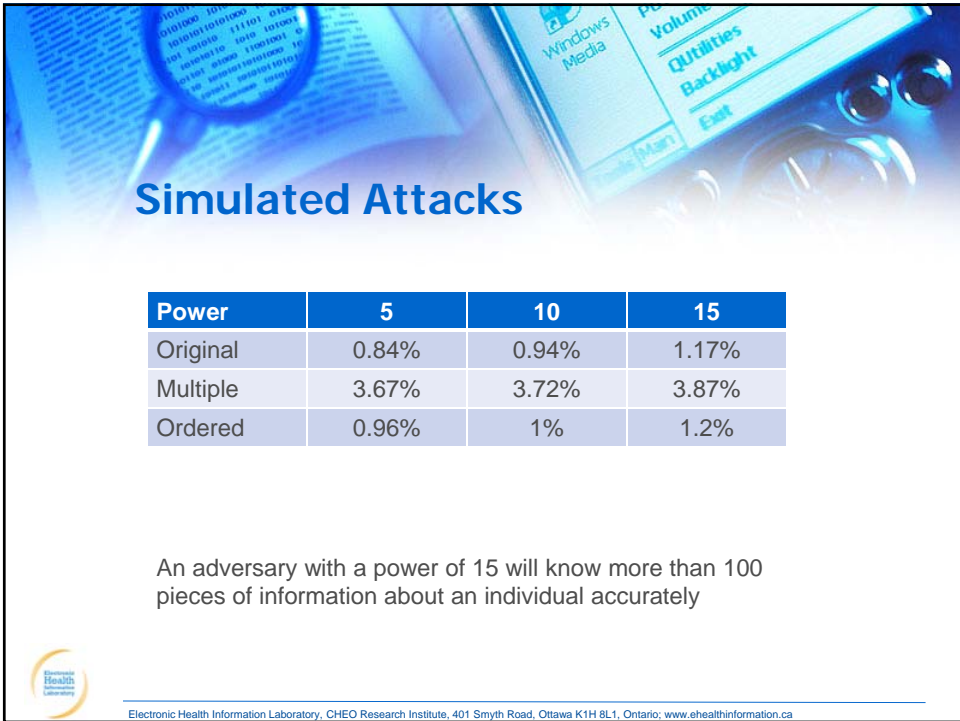
Electronic Health Information Laboratory, CHEO Research Institute, 401 Smyth Road, Ottawa K1H 8L1, Ontario, www.ehealthinformation.ca



## Adversary Power

- Adversary will not have background knowledge about all claims
- If we assume that the adversary has the information from 5 claims, which claims do we include in the risk assessment ?
- Adversary power was computed separately for each patient – account for diversity in a patient’s claims
- Bootstrap estimate of percentage of records with a re-id probability greater than 0.05 was used to decide on an optimal node in lattice



 Electronic Health Information Laboratory, CHEO Research Institute, 401 Smyth Road, Ottawa K1H 8L1, Ontario: [www.ehealthinformation.ca](http://www.ehealthinformation.ca)

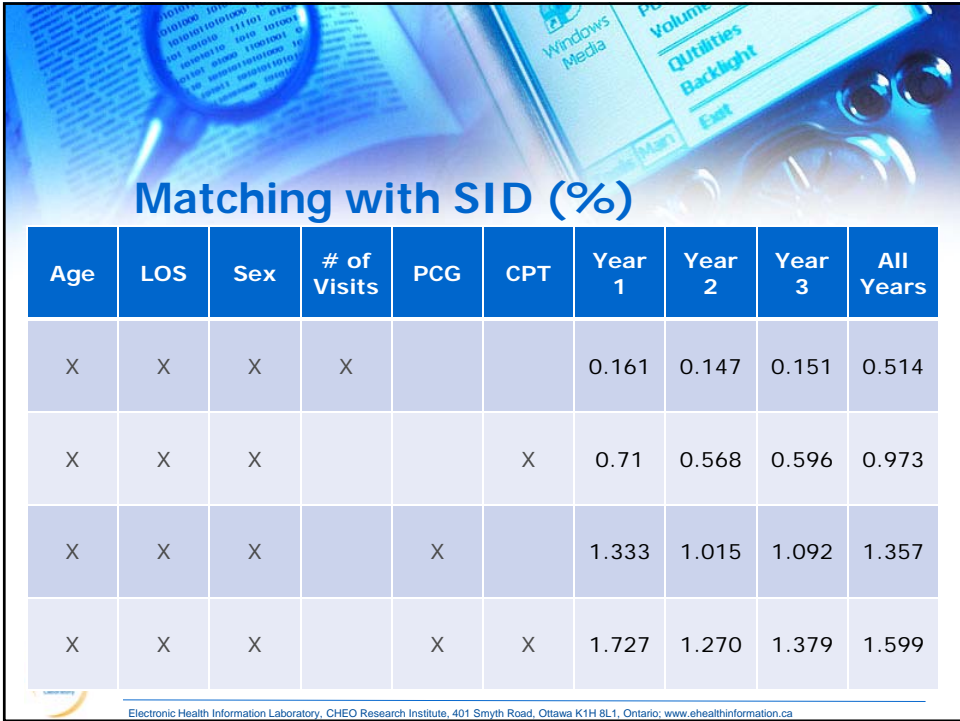


## Simulated Attacks

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


 Electronic Health Information Laboratory, CHEO Research Institute, 401 Smyth Road, Ottawa K1H 8L1, Ontario: [www.ehealthinformation.ca](http://www.ehealthinformation.ca)



## 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

Electronic Health Information Laboratory, CHEO Research Institute, 401 Smyth Road, Ottawa K1H 8L1, Ontario: [www.ehealthinformation.ca](http://www.ehealthinformation.ca)



## Things we would do differently

- We have since developed more sophisticated ways for claim truncation that would result in less information loss
- We need more sophisticated ways that are less computationally intensive for estimating re-identification risk at different adversary power levels

Electronic Health Information Laboratory, CHEO Research Institute, 401 Smyth Road, Ottawa K1H 8L1, Ontario: [www.ehealthinformation.ca](http://www.ehealthinformation.ca)



[kelemam@uottawa.ca](mailto:kelemam@uottawa.ca)

[www.ehealthinformation.ca](http://www.ehealthinformation.ca)

[www.ehealthinformation.ca/knowledgebase](http://www.ehealthinformation.ca/knowledgebase)

