

Sharing Health Data for Research

Khaled El Emam, CHEO RI & uOttawa





Context

- The disclosure of health information for secondary purposes, such as research
- Many practical challenges to obtaining express individual consent from patients for large databases
- Even if express consent can be obtained, there is compelling evidence that consenters differ from non-consenters – introduces bias
 - age, sex, race, marital status, educational level, socioeconomic status, health status, mortality, lifestyle factors, functioning



Variable Distinctions

- Directly identifying
 - Can uniquely identify an individual by itself or in conjunction with other readily available information
- Quasi-identifiers
 - Can identify an individual by itself or in conjunction with other information
- Sensitive variables



Examples of Direct Identifiers

- Name, address, telephone number, fax number, MRN, health card number, health plan beneficiary number, license plate number, email address, photograph, biometrics, SSN, SIN, implanted device number



Examples of Quasi-Identifiers

- sex, date of birth or age, geographic locations (such as postal codes, census geography, information about proximity to known or unique landmarks), language spoken at home, ethnic origin, aboriginal identity, total years of schooling, marital status, criminal history, total income, visible minority status, activity difficulties/reductions, profession, event dates (such as admission, discharge, procedure, death, specimen collection, visit/encounter), codes (such as diagnosis codes, procedure codes, and adverse event codes), country of birth, birth weight, and birth plurality



Methods

- Masking
 - Deals with the directly identifying variables
- De-identification
 - Deals with the quasi-identifiers



Masking - I

- Suppression
 - Removal of directly identifying fields
- Pseudonymization
 - Replace direct identifiers with unique keys that cannot be reversed
- Randomization
 - Replace direct identifiers with random values (eg, random names, MRNs, telephone numbers, postal codes)



Masking - II

- Adding Noise
 - Sometimes people add noise to data
 - This is risky because filters can be applied to the data to remove the noise and recover the original signal



Masking is not enough

- Removing names and addresses from a data set does not de-identify it
- It is possible to re-identify individuals using residual information, such as date of birth and postal code
- Consider uniqueness in the Canadian population

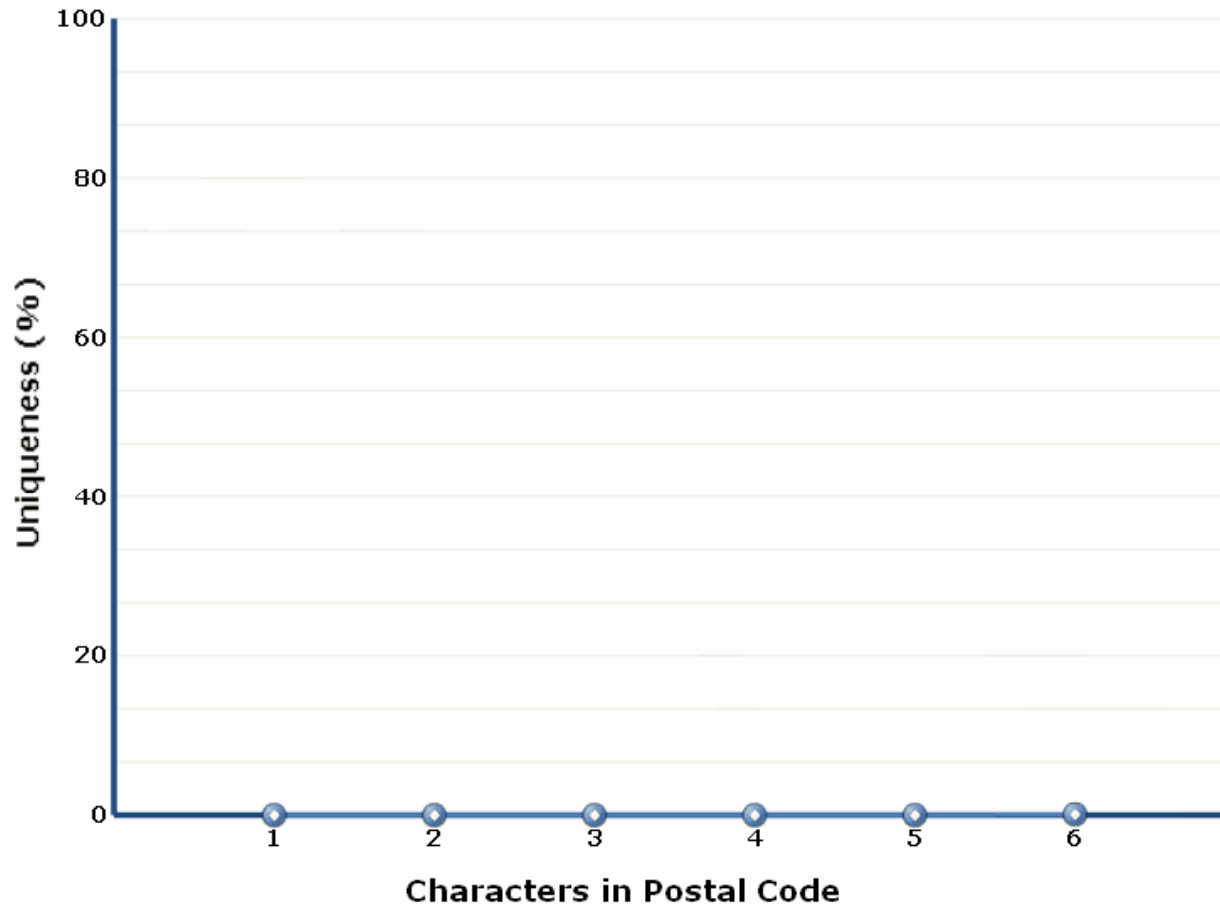
Residence Trails

DoB

None

Gender

None



Trace

- 1 year
- 2 years
- 3 years
- 4 years
- 5 years
- 6 years
- 7 years
- 8 years
- 9 years
- 10 years
- 11 years

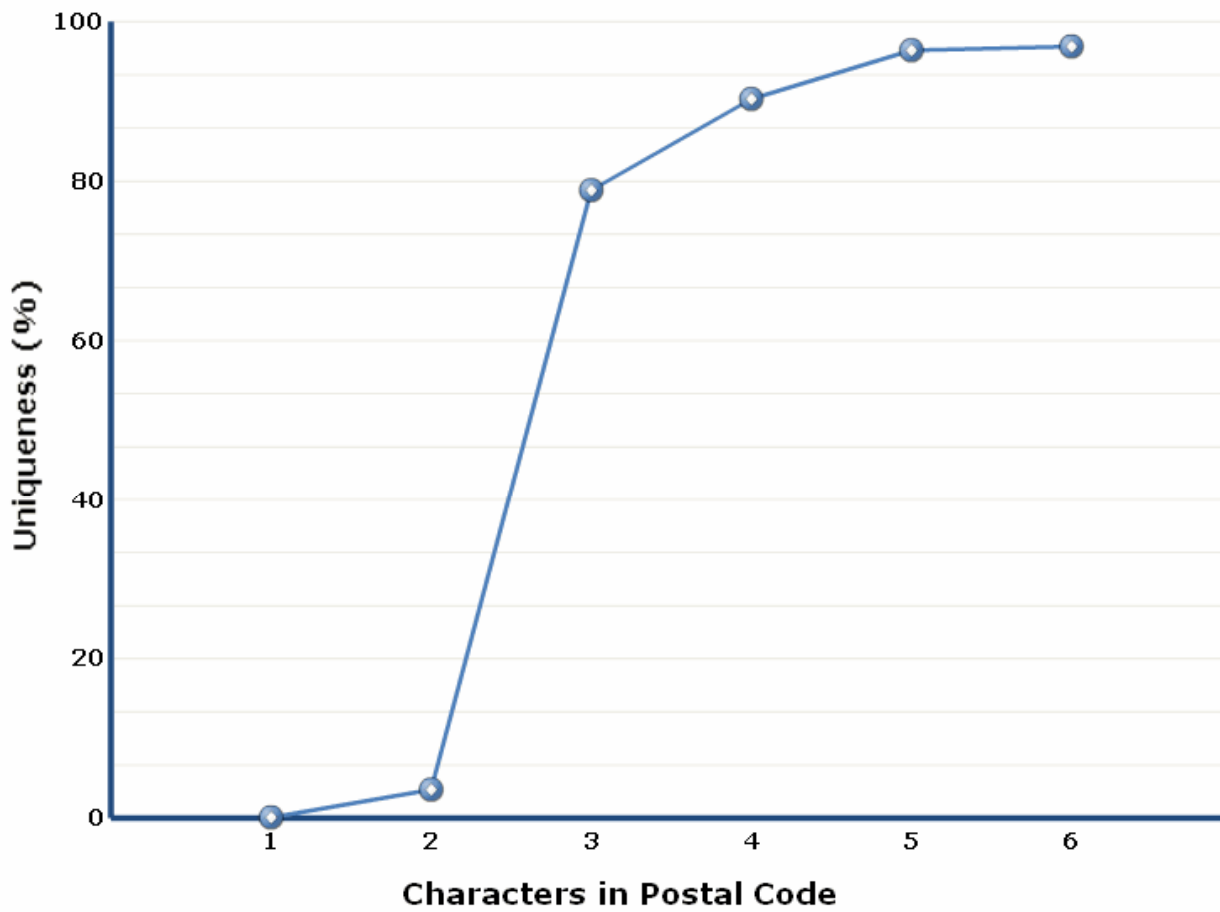
Residence Trails

DoB

Day/Month/...

Gender

None



Trace

- 1 year
- 2 years
- 3 years
- 4 years
- 5 years
- 6 years
- 7 years
- 8 years
- 9 years
- 10 years
- 11 years

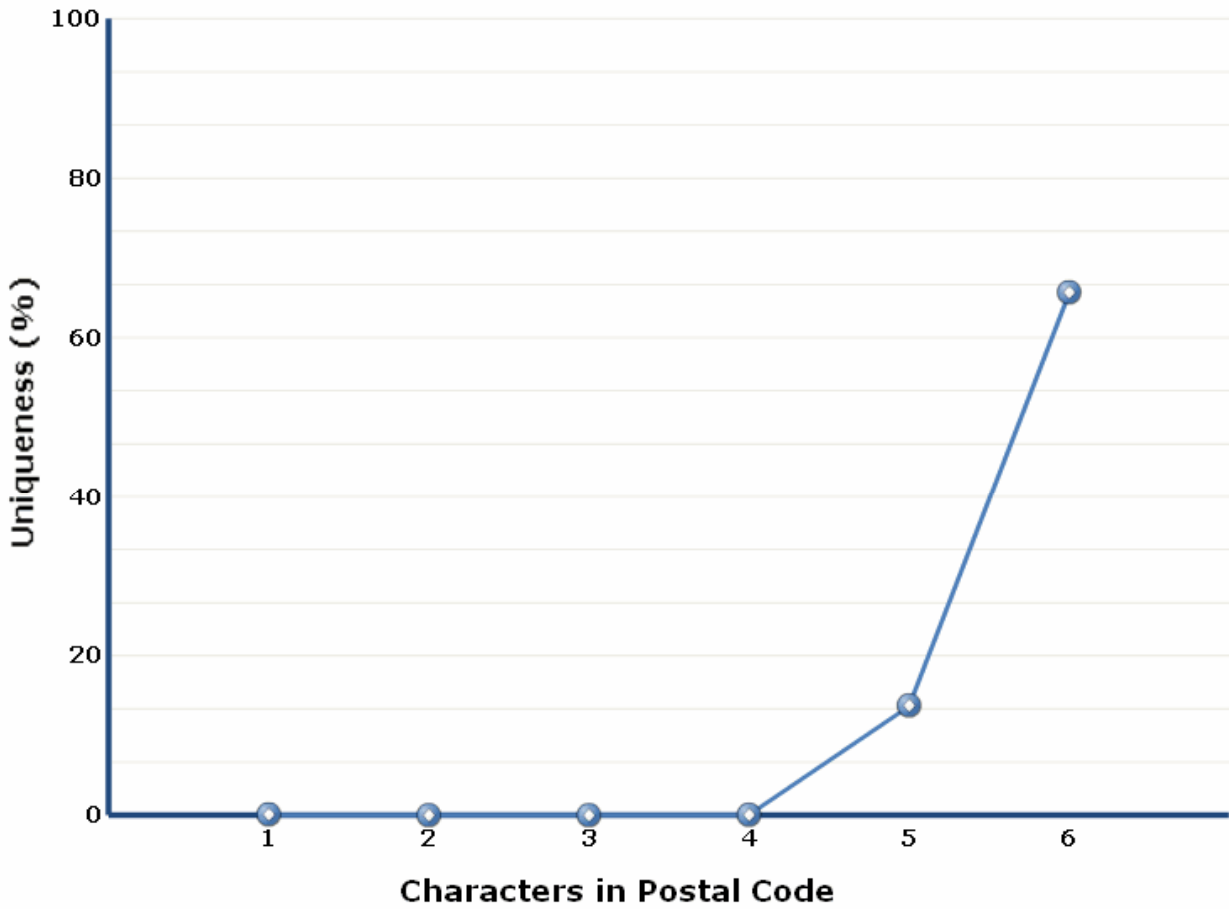
Residence Trails

DoB

Year

Gender

None



Trace

- 1 year
- 2 years
- 3 years
- 4 years
- 5 years
- 6 years
- 7 years
- 8 years
- 9 years
- 10 years
- 11 years

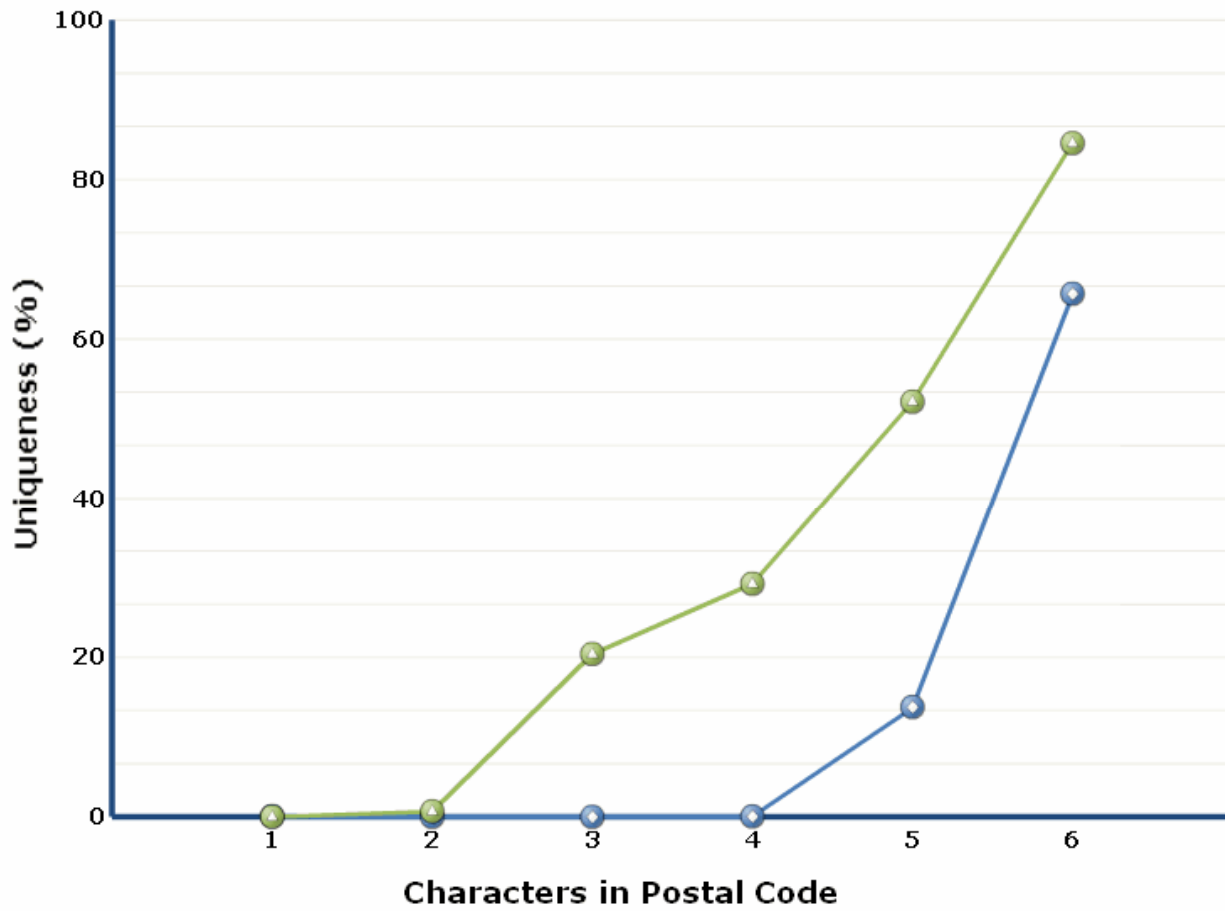
Residence Trails

DoB

Year

Gender

None



Trace

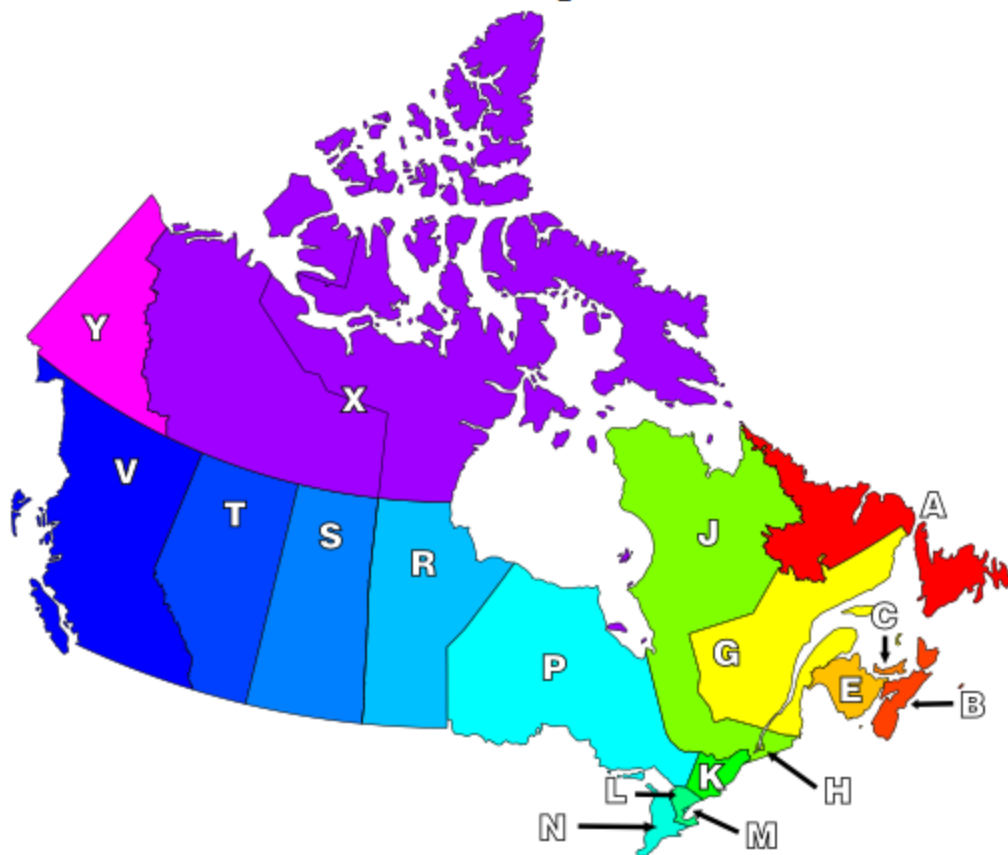
- 1 year
- 2 years
- 3 years
- 4 years
- 5 years
- 6 years
- 7 years
- 8 years
- 9 years
- 10 years
- 11 years



Example Query

- Consider a protocol with basic demographics about the patients:
 - Age
 - Gender
 - Language spoken at home
 - Visible minority status
- The REB Wizard tool is here:
<http://www.ehealthinformation.ca/rebwizard/ca/>

Selected Region: K



A | B | C | E | G | H | J | K | L | M | N | P | R | S | T | V | X | Y

Postal Code Digits

3

Uniqueness Threshold

0%

Variables

Combinations: 1800

Name	Values
Gender	2
Age	90
Language	5
Visible Minority	2

Add Row

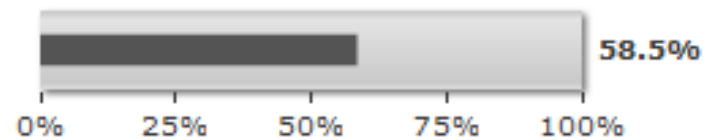
Delete Row

Clear

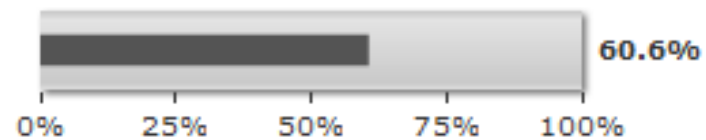
Submit



Population at Risk

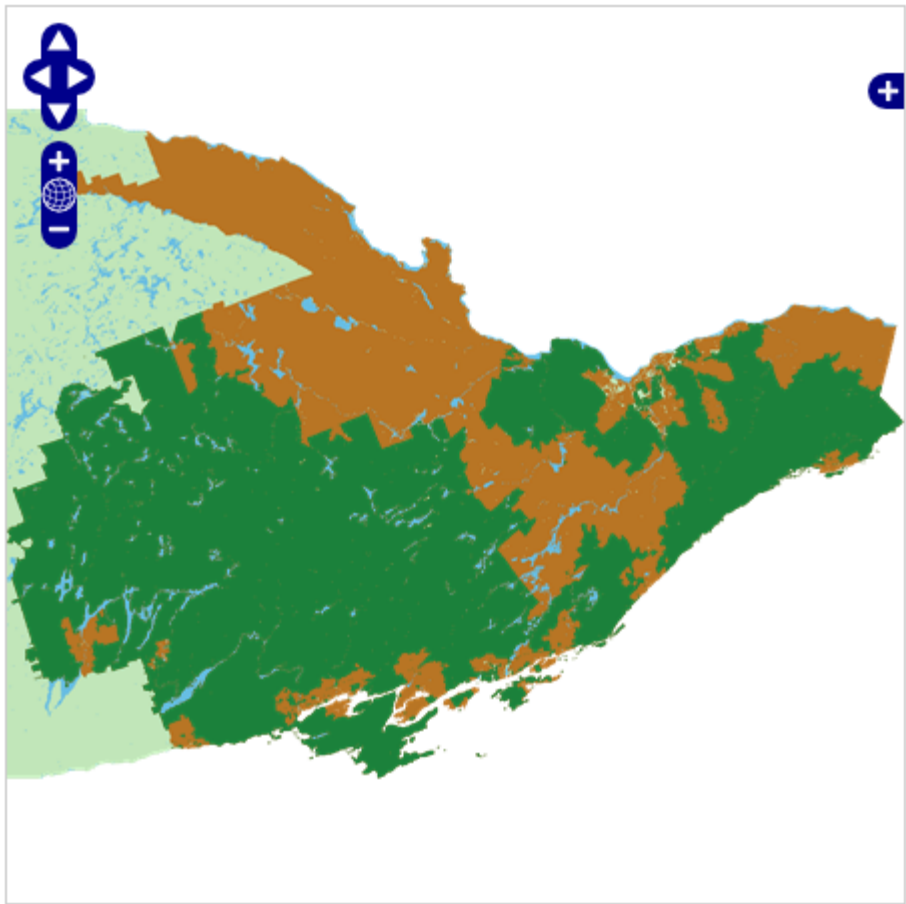


Households at Risk



Variables

Name	Values
Gender	2
Age	90
Language	5
Visible Minority	2



Postal Code Digits

2

Selected Region: K

Variables

Combinations: 1800

Name	Values
Age	90
Gender	2
Language	5
Visible Minority	2

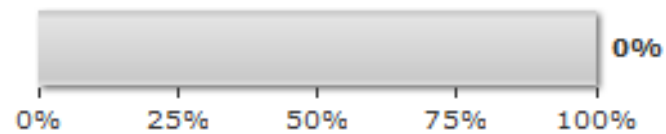
Add Row Delete Row

Clear

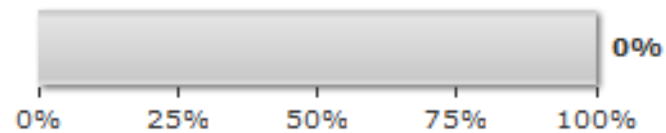
Submit

A | B | C | E | G | H | J | K | L | M | N | P | R | S | T | V | X | Y

Population at Risk



Households at Risk



Variables

Name	Values
Gender	2
Age	90
Language	5
Visible Minority	2

Selected Region: K

Postal Code Digits: 3

Uniqueness Threshold: [Slider]

Variables

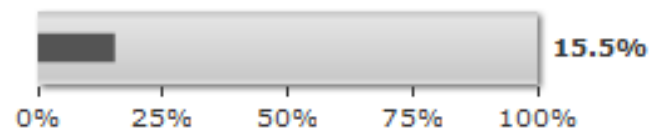
Combinations: 360

Name	Values
Gender	2
Age	18
Language	5
Visible Minority	2

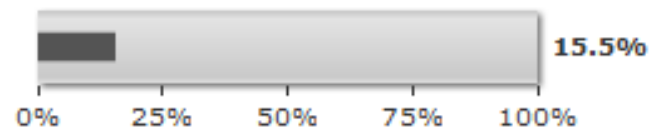
Submit

A|B|C|E|G|H|J|K|L|M|N|P|R|S|T|V|X|Y

Population at Risk



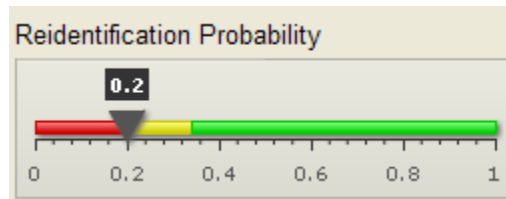
Households at Risk



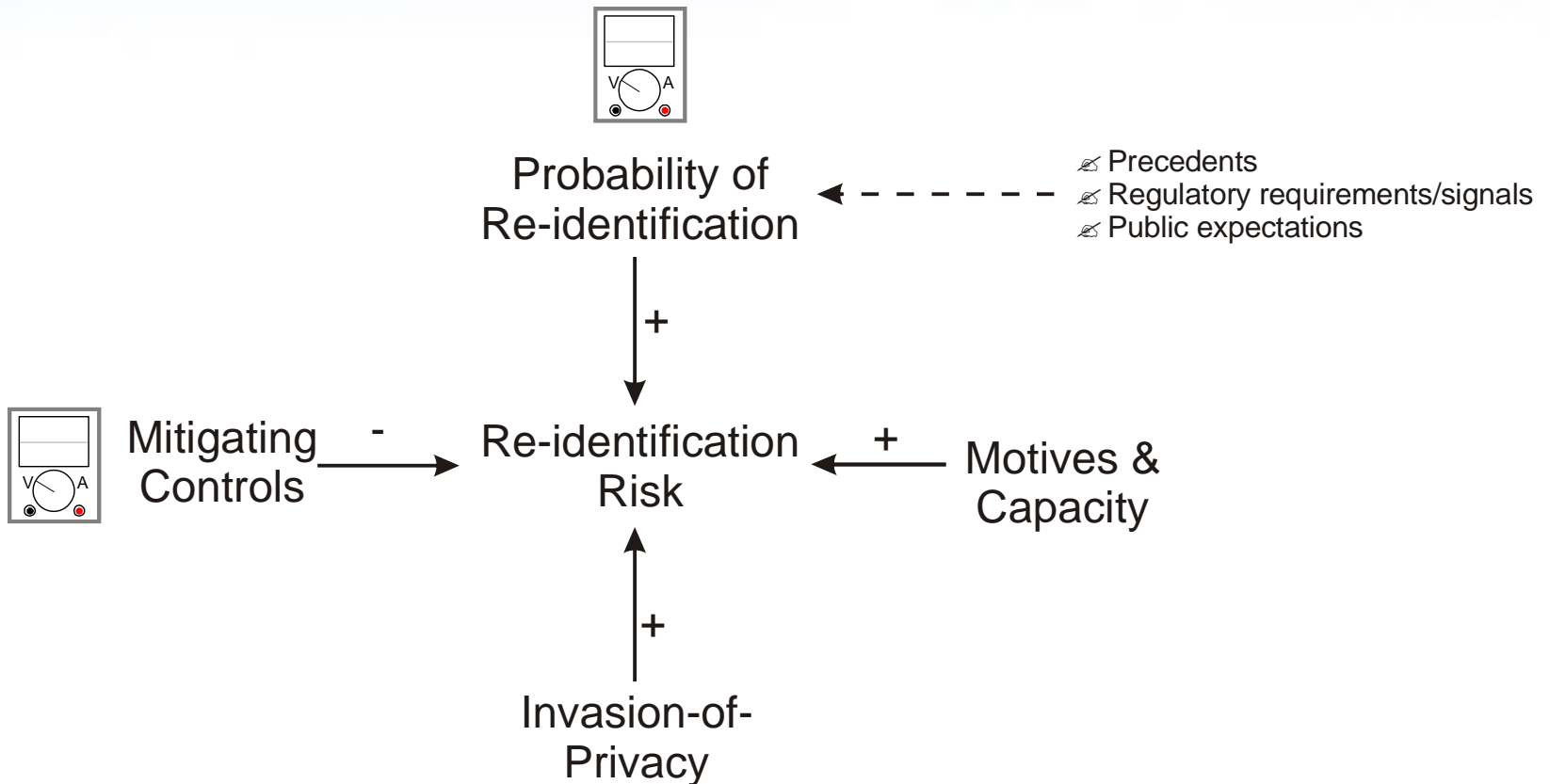
Variables

Name	Values
Gender	2
Age	18
Language	5
Visible Minority	2

Re-identification Risk Spectrum



Managing Re-identification Risk



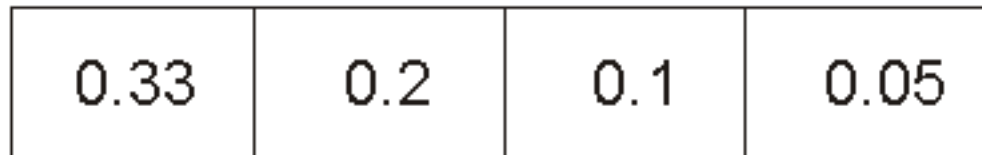
The Tradeoffs

highly secure
and trusted recipients

public
use files

*little
de-identification*

*significant
de-identification*



- strong security & privacy practices
- not sensitive data
- consent sought or authority
- no motives to re-identify

- no security & privacy practices
- sensitive data
- no authority to disclose data
- no consent sought
- strong motives to re-identify



Examples

- BORN Ontario – provincial birth registry
- Cancer Data Link (ICES/OICR)
- Rick Hansen SCI Registry
- Vancouver Coastal Health
- CHEO
- CPCSSN – public health
- CMS – public use files

kelemam@uottawa.ca

www.ehealthinformation.ca

www.ehealthinformation.ca/knowledgebase





References

- El Emam K, Mercer J, Moreau K, Grava-Gubins I, Buckeridge D, Jonker E: **Physician Privacy Concerns when Disclosing Patient Data for Public Health Purposes During a Pandemic Influenza Outbreak.** *BMC Public Health* (to appear).
- El Emam K, Hu J, Mercer J, Peyton L, Kantarcioglu M, Malin B, Buckeridge D, Samet S, Earle C: **A Secure Protocol for Protecting the Identity of Providers When Disclosing Data for Disease Surveillance.** *Journal of the American Medical Informatics Association*, 18:212-217, 2011.
- El Emam K, Brown A, AbdelMalik P, Neisa A, Walker M, Bottomley J, Roffey T: **A method for managing re-identification risk from small geographic areas in Canada.** *BMC Medical Informatics and Decision Making*, 10, 2010.
- El Emam K, Brown A, Abdelmalik P: **Evaluating Predictors of Geographic Area Population Size Cutoffs to Manage Re-identification Risk.** *Journal of the American Medical Informatics Association*, 16:256-266, 2009.
- El Emam K, Dankar F, Issa R, Jonker E, Amyot D, Cogo E, Corriveau J-P, Walker M, Chowdhury S, Vaillancourt R, Roffey T, Bottomley J: **A Globally Optimal k-Anonymity Method for the De-identification of Health Data.** *Journal of the American Medical Informatics Association*, 16(5):670-682, 2009.
- El Emam K, Jabbouri S, Sams S, Drouet Y, Power M. **Evaluating common de-identification heuristics for personal health information.** *Journal of Medical Internet Research*, 2006; 8(4):e28.