

Protecting Patient Privacy

Khaled El Emam, CHEO RI & uOttawa





Context

- In Ontario data custodians are permitted to disclose PHI without consent for public health purposes
- What is the problem then ?
 - This disclosure is discretionary (with exceptions)
 - Data custodians are reluctant to do so without de-identification
 - Data custodians are concerned about patient and *their own* privacy
 - General limiting principles
- How do we de-identify data but still ensure that it is useful ?

Collection/Control/Awareness

- ✍ De-identify data where possible (e.g., when identifiable information not required for contact tracing) [C1]
- ✍ Provide notice to patients [C2]
- ✍ There is a clear link to patient benefit [C3]
- ✍ Provide actionable, regular and clear feedback to physicians after the data is collected [C4]
- ✍ Obtain support from the professional college(s) [C5]
- ✍ Put in place data sharing agreements with the physicians [C6]

Trusting Beliefs

- ✍ Public health has good data handling practices [T1]
- ✍ There are no unanticipated uses of data different from the stated purpose(s) [T2]
- ✍ Data will not be shared with other third parties [T3]

-ve

Risk Beliefs

- ✍ Damage to physician-patient relationship [R1]
- ✍ Patient complaints or legal action [R2]
- ✍ Negative consequences due to exposure of physician data (e.g., compliance audits or performance reviews) [R3]
- ✍ Use up valuable resources with no return/benefit [R4]

+ve

Intention to Disclose Patient Information

-ve





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

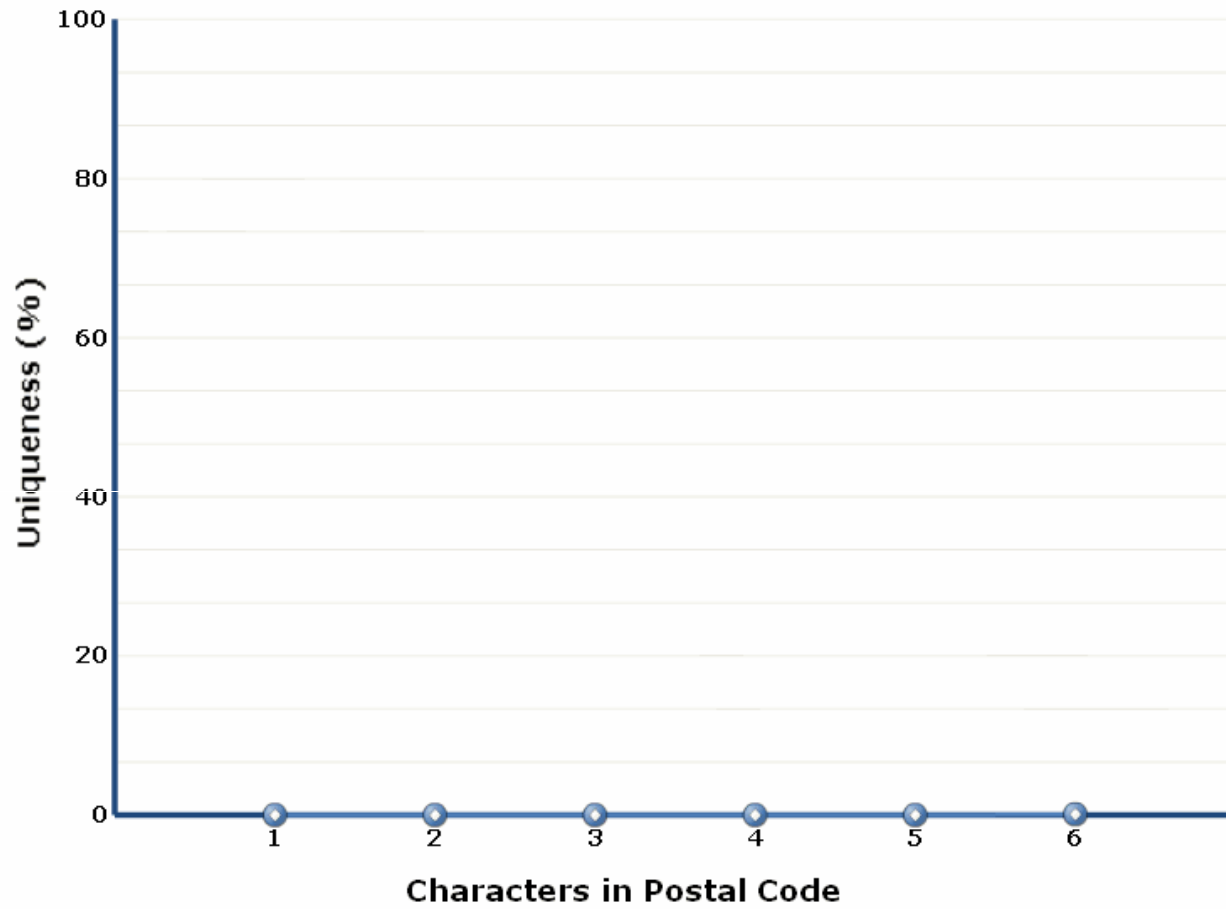
Residence Trails

DoB

None

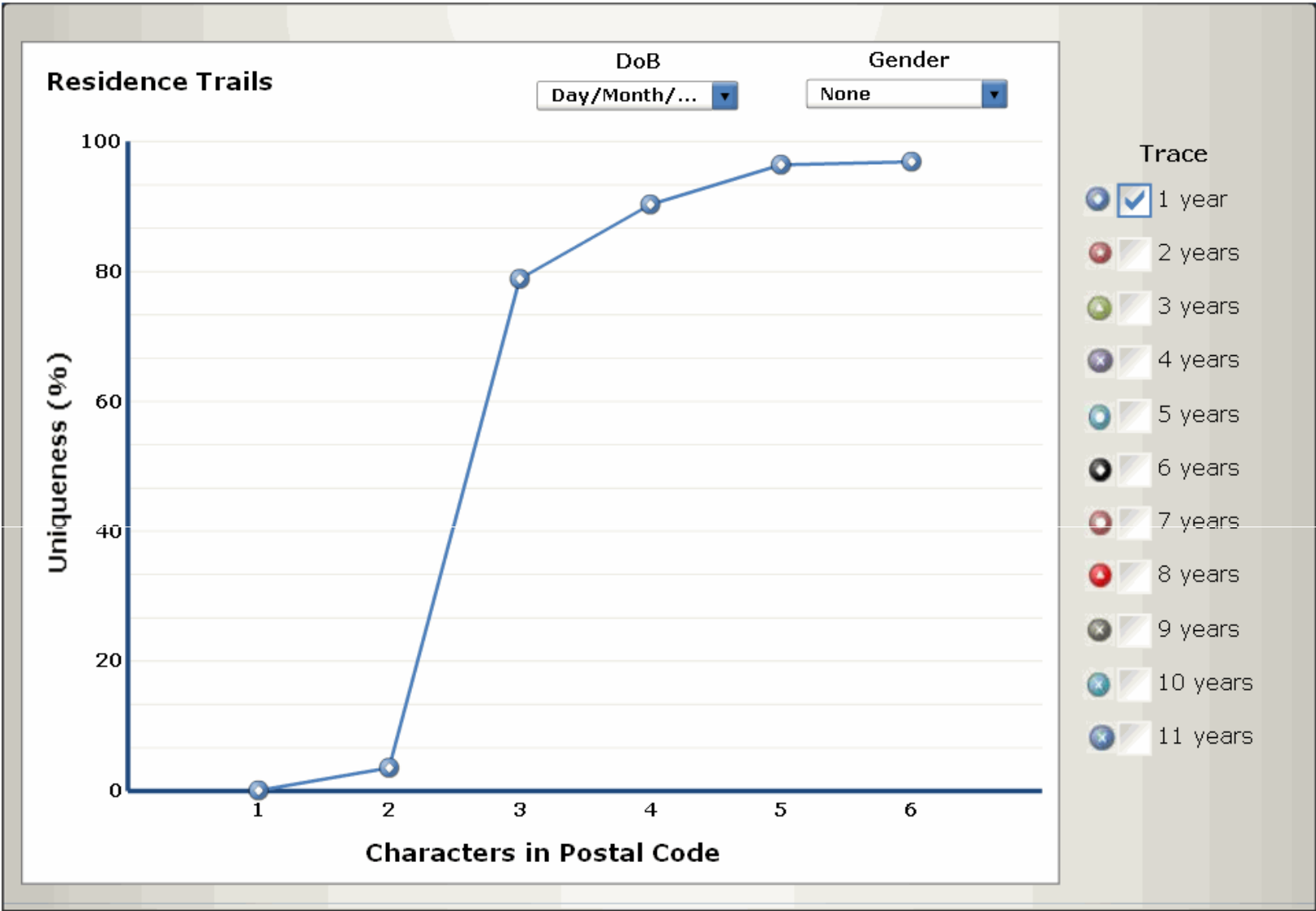
Gender

None



Trace

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



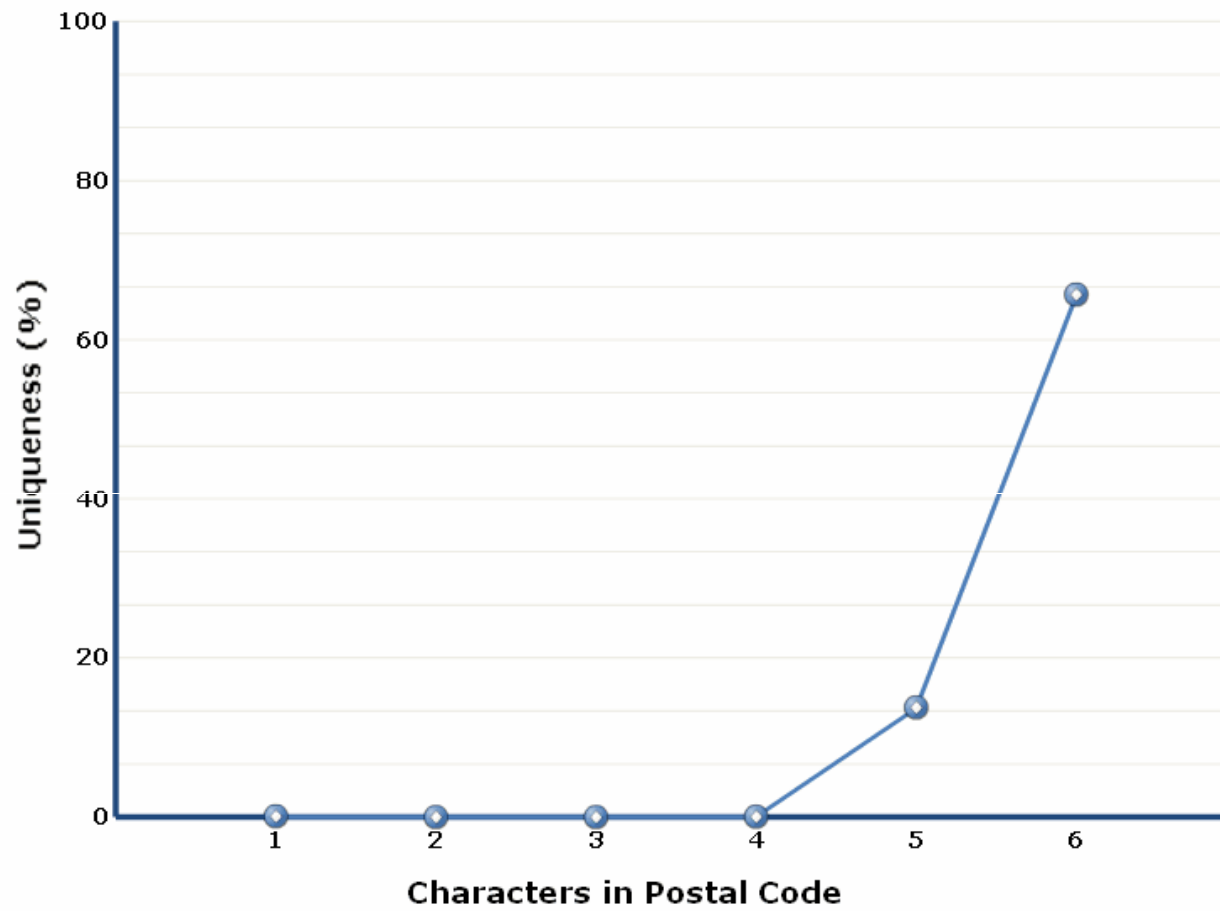
Residence Trails

DoB

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

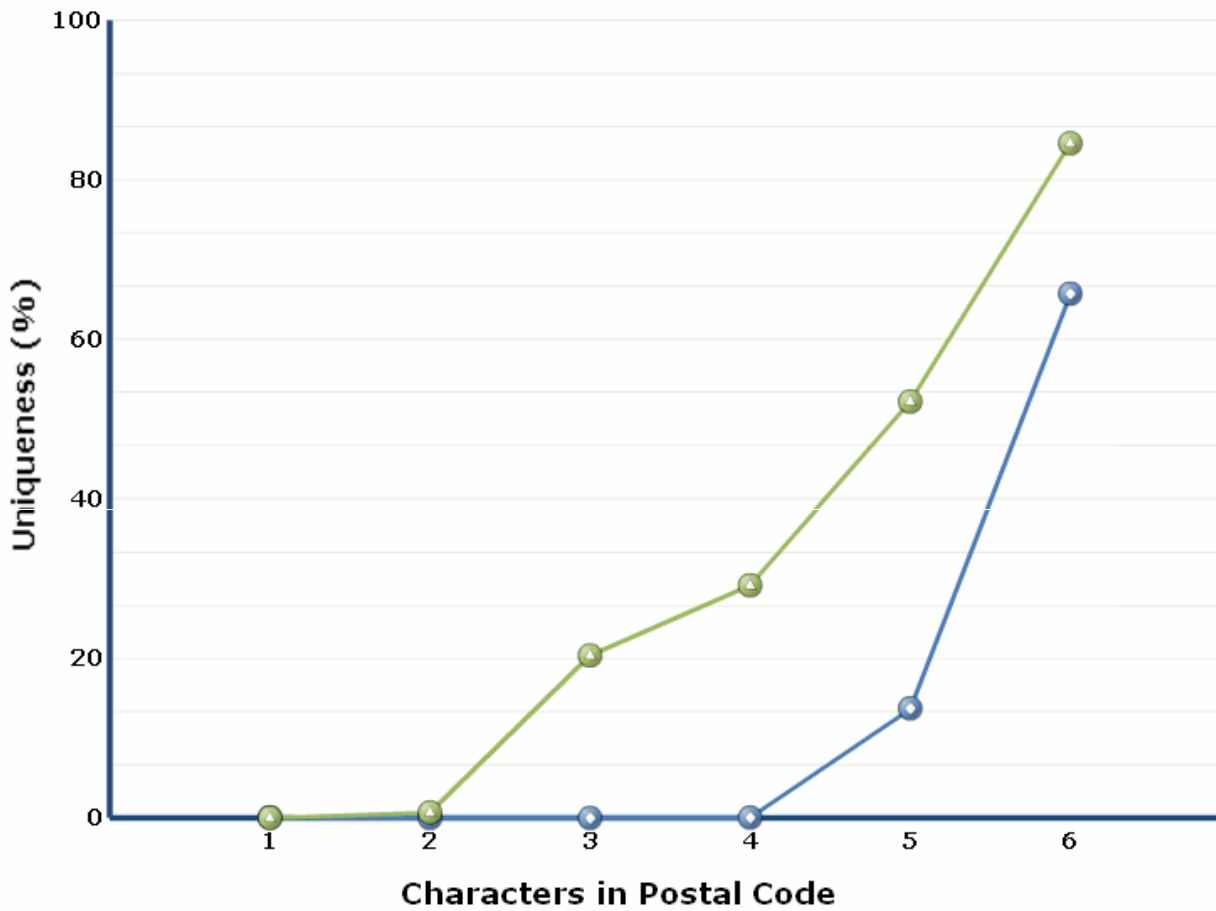
Residence Trails

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



REB Wizard Tool

- Allows risk assessment with more variables (demographic and socio-economic)
- Based on an analysis of census data
- 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% ▾

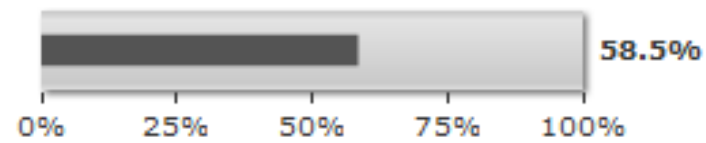
Variables

Combinations: 1800 ⌵

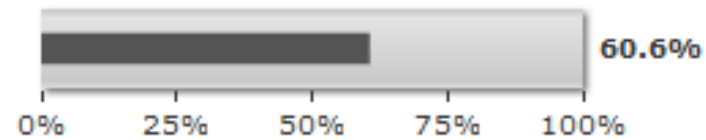
Name	Values
Gender	2
Age	90
Language	5
Visible Minority	2



Population at Risk



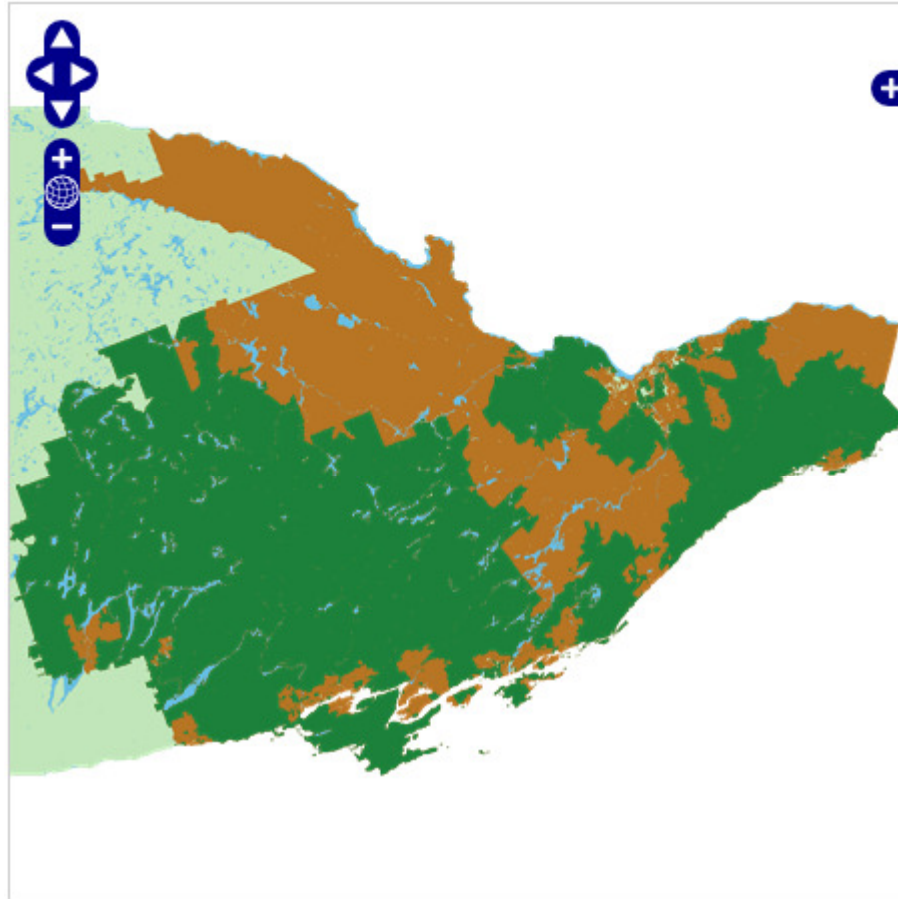
Households at Risk



Variables

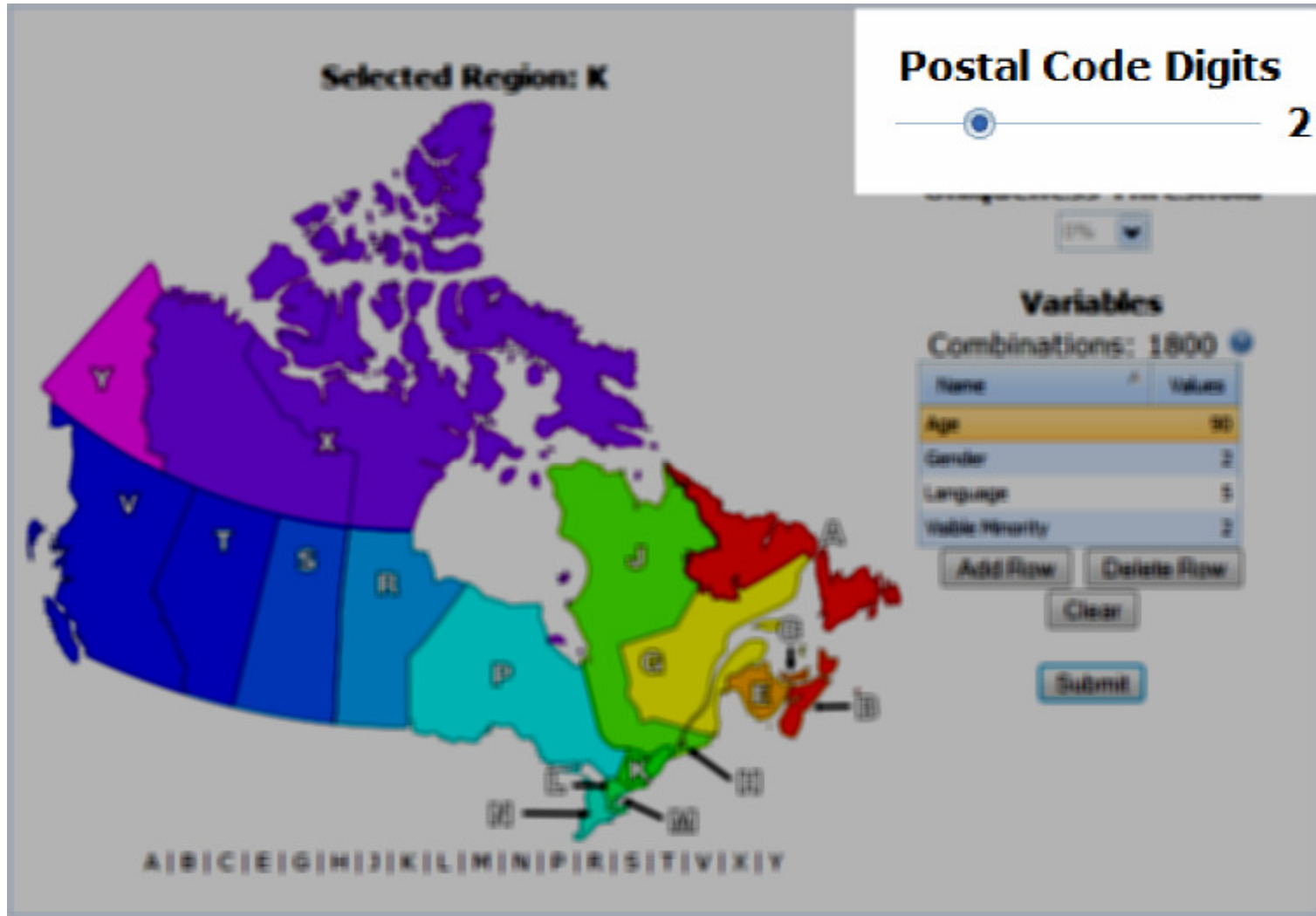
Name	Values
Gender	2
Age	90
Language	5
Visible Minority	2



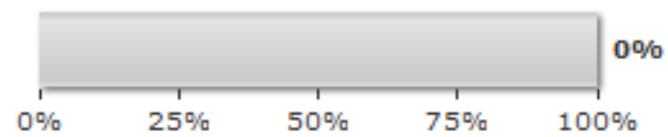


Postal Code Digits

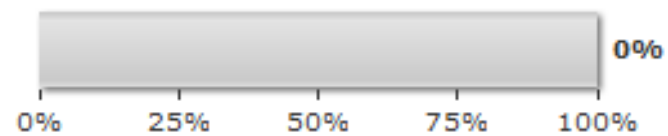
2



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

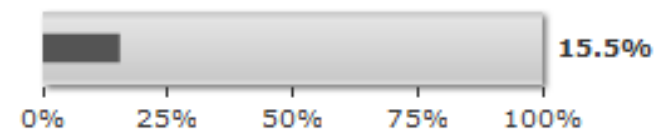
Combinations: 360

Name	Values
Gender	2
Age	18
Language	5
Visible Minority	2

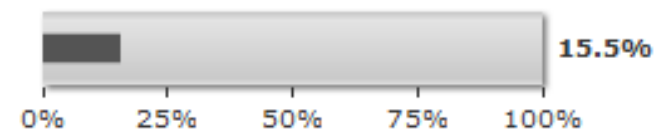
Submit



Population at Risk



Households at Risk



Variables

Name	Values
Gender	2
Age	18
Language	5
Visible Minority	2





Example: Syndromic Surveillance

- Data collected from emergency department and sent to the public health unit
- Data set of all presentations at CHEO emergency from June 2007 to June 2009, 108,344 records
- Basic information: DoB, time and date of presentation, gender, postal code, chief complaint
- What is the re-identification risk ?

Example: Syndromic Surveillance

ID	Quasi-identifier Precision				Percent of Records with Cell Suppression	
	Date of Presentation	Postal Code	Date of Birth	Gender	risk threshold at 0.2	risk threshold at 0.33
1	day/month/year	6 char	day/month/year	M/F	100%	100%
2	quarter/year	1 char	quarter/year	M/F	<5%	-
3	month/year	1 char	quarter/year	M/F	-	<5%
4	day/month/year	?	?	?	no solution	no solution
5	day/month/year	3 char	year	M/F	100%	99.7%
6	day/month/year	1 char	year	M/F	60.6%	46.73%
7	day/month/year	1 char	5 year interval	M/F	11.22%	8.3%
8	day/month/year	1 char	10 year interval	M/F	6%	5.1%
9	month/year	1 char	year	M/F	2.3%	2.15%
10	month/year	3 char	year	M/F	54.6%	43.72%



Options for Syndromic Surveillance

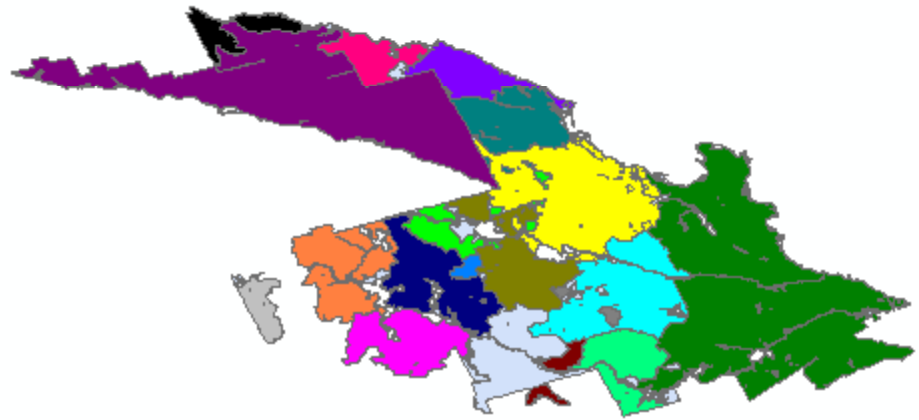
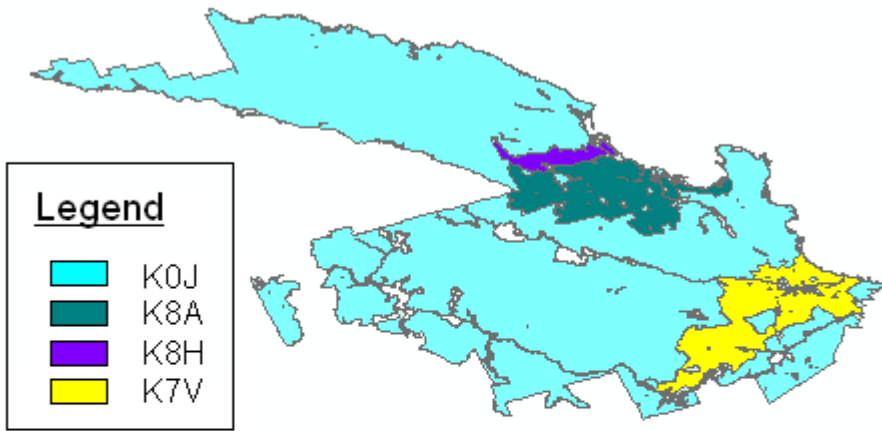
Data Options

		Disclose data on all patients	Disclose data on affected patients
Identifiability Options	Disclose identifiable data	Common Practice	Limited identifiable data is disclosed
	Disclose de-identified data	Standard de-identification approach	Limited de-identified data is disclosed
	Hybrid disclosure	Standard de-identified data with override	Limited de-identified data with override



Aggregating Postal Codes

- Postal code information always increases re-identification risk
- If postal code will be used to link to a SES file, then often the data custodian can link and then disclose the data without postal code
- Otherwise, common aggregation methods are quite crude
- Optimal aggregation can produce small areas that maintain acceptable re-identification risks
- Optimal aggregation can maintain homogeneous SES values

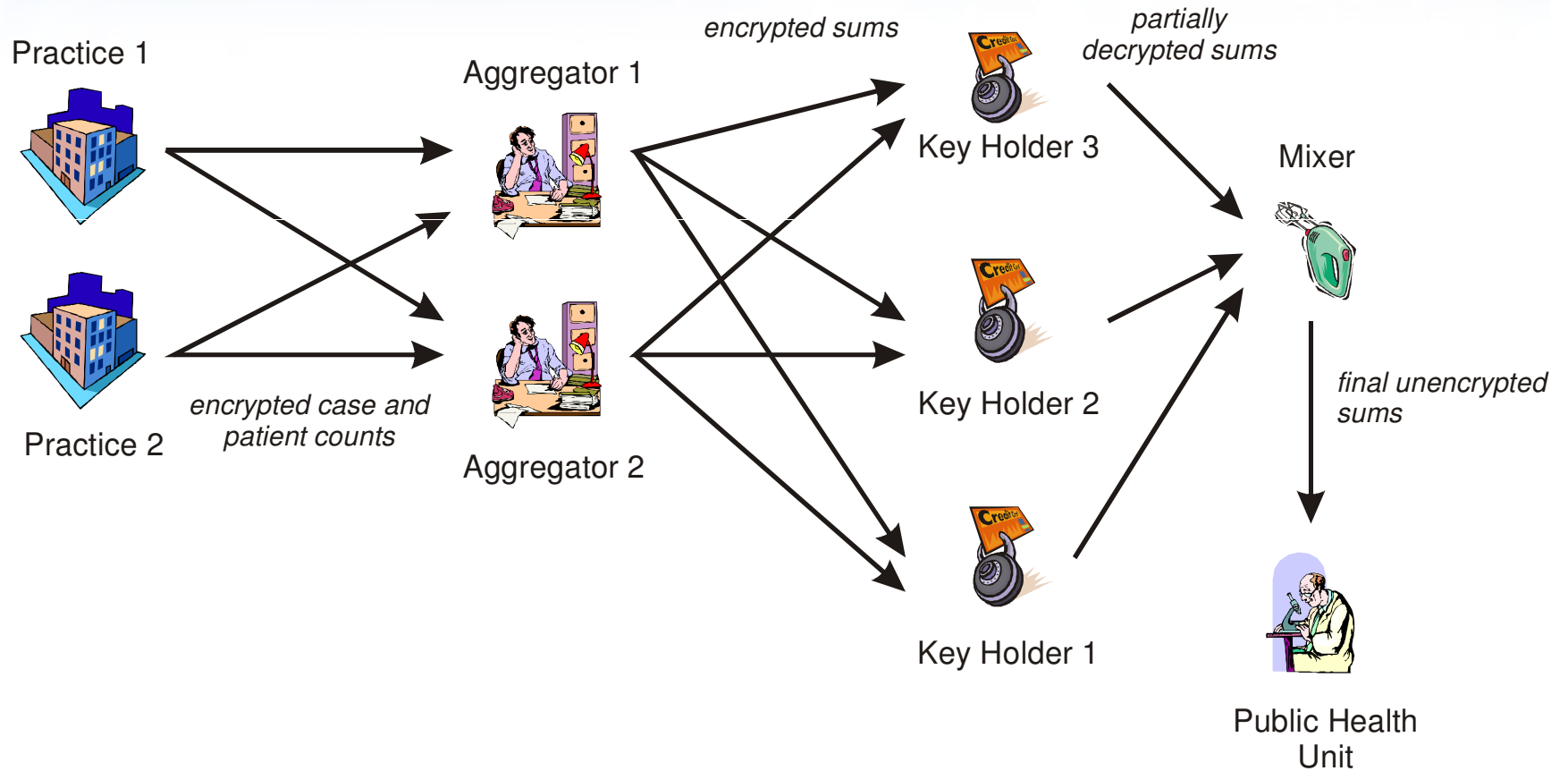




Secure Multi-Party Computation

- Allows the computation of statistics in a secure manner without the sharing of raw data
- Only encrypted data is shared and basic or complex statistical analyses can be performed on these
- The public health unit gets the final encrypted statistical results and decrypts these
- The protocol can:
 - can provide strong privacy guarantees without compromising data granularity
 - protect the identity of providers, with an override to allow investigations

Secure Counts Protocol





kelemam@uottawa.ca

www.ehealthinformation.ca

www.ehealthinformation.ca/knowledgebase





References

- 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.