Method and Experiences of Risk-Based De-identification

Khaled El Emam

www.ehealthinformation.ca

www.ehealthinformation.ca/knowledgebase

Background

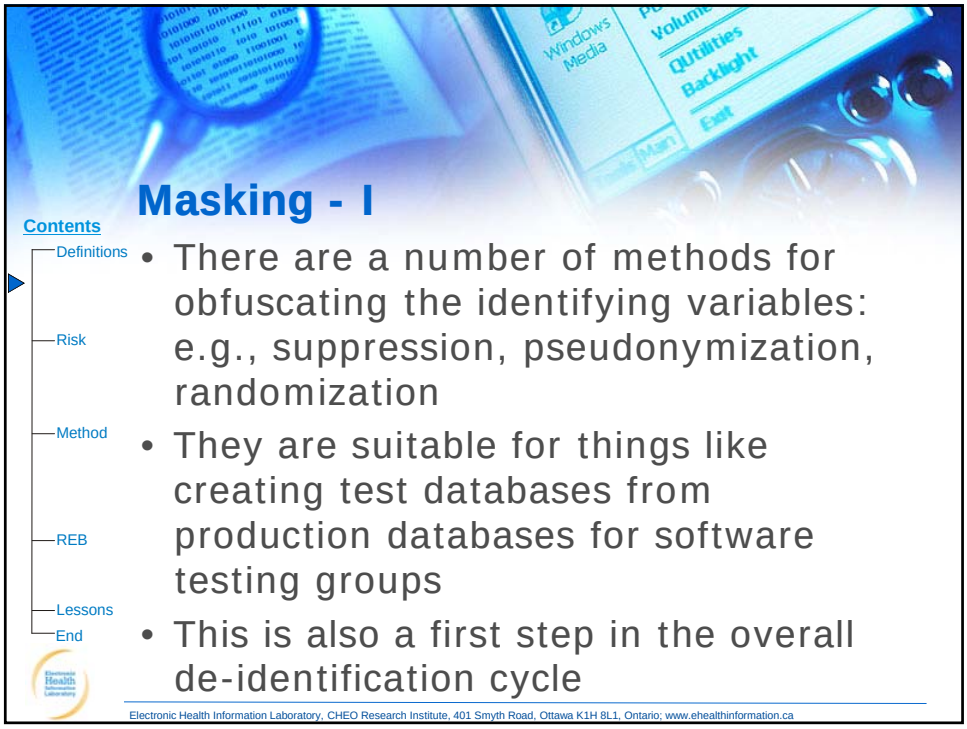
Contents
- Definitions
- Risk
- Method
- REB
- Lessons
- End

- We have been measuring re-identification risk and de-identifying clinical data sets for a few years (national and provincial registries, physician practice data, hospital data, as well as demographics and SES variables in census data)
- Developed and adopted a risk-based approach to de-identification

Variable Distinctions

Contents
- Definitions
- Risk
- Method
- REB
- Lessons
- End

- 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

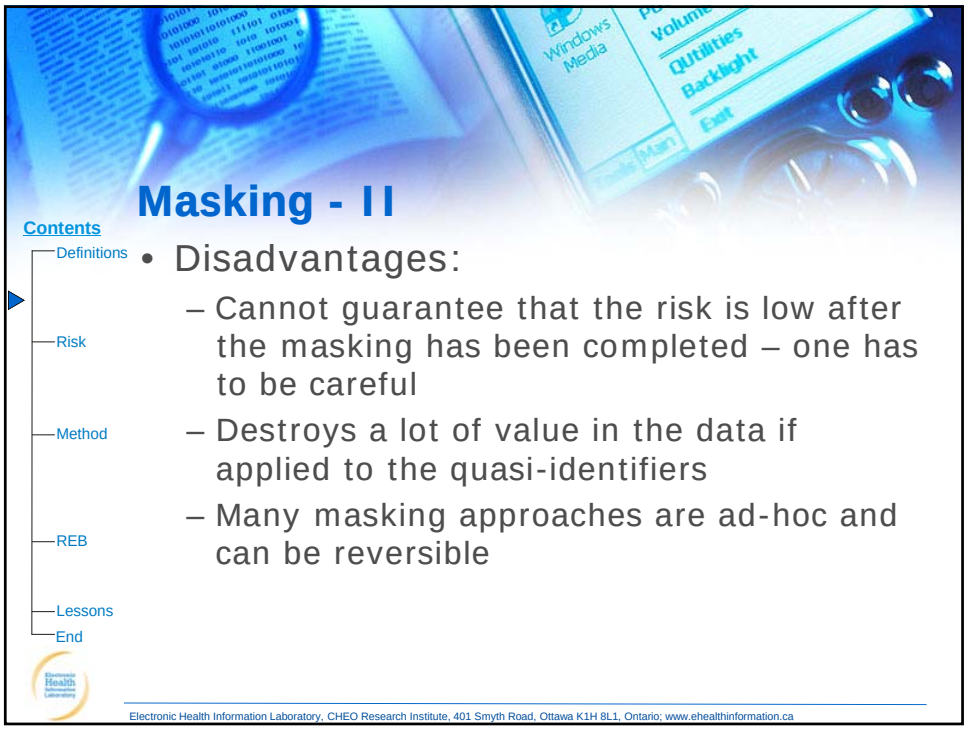


Masking - I

Contents

- Definitions
- Risk
- Method
- REB
- Lessons
- End

- There are a number of methods for obfuscating the identifying variables: e.g., suppression, pseudonymization, randomization
- They are suitable for things like creating test databases from production databases for software testing groups
- This is also a first step in the overall de-identification cycle



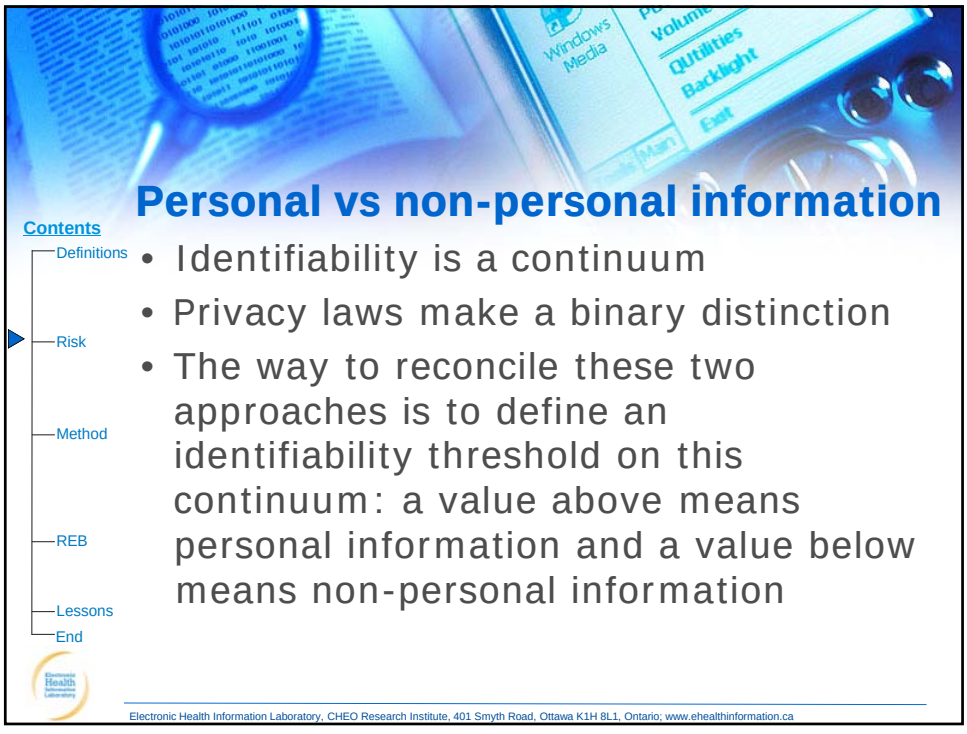
Masking - II

Contents

- Definitions
- Risk
- Method
- REB
- Lessons
- End

- Disadvantages:
  - Cannot guarantee that the risk is low after the masking has been completed – one has to be careful
  - Destroys a lot of value in the data if applied to the quasi-identifiers
  - Many masking approaches are ad-hoc and can be reversible

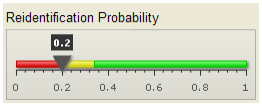
Electronic Health Information Laboratory, CHEO Research Institute, 401 Smyth Road, Ottawa K1H 8L1, Ontario; www.ehealthinformation.ca


Windows Media
Volume
Utilities
Backlight
Exit
Main
Personal vs non-personal information
Contents
Definitions
Risk
Method
REB
Lessons
End
• Identifiability is a continuum
• Privacy laws make a binary distinction
• The way to reconcile these two approaches is to define an identifiability threshold on this continuum: a value above means personal information and a value below means non-personal information
Electronic Health Information Laboratory, CHEO Research Institute, 401 Smyth Road, Ottawa K1H 8L1, Ontario; www.ehealthinformation.ca

Identifiability Spectrum

Contents

- Definitions
- Risk
- Method
- REB
- Lessons
- End



Five Levels of Identifiability

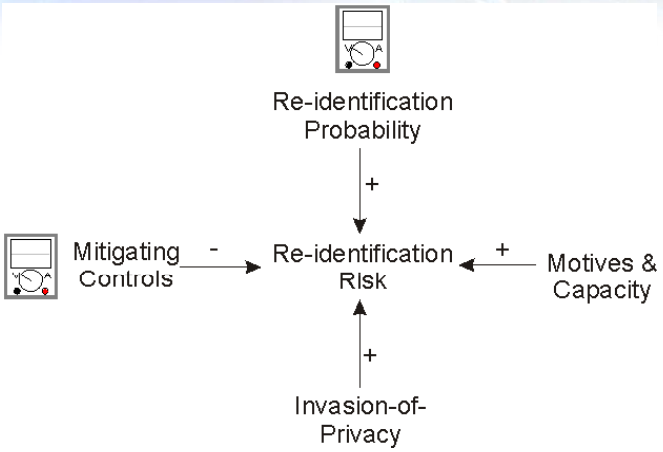
greater risk of re-identification

greater effort, cost, time & skill to re-identify

Level 5	Aggregate Data		not personal information
Level 4	Managed Data	*identifiability below threshold* / *identifiability above threshold*	
Level 3	Exposed Data		personal information
Level 2	Masked Data	*irreversibly masked data* / *reversibly masked data*	
Level 1	Readily Identifiable Data		

Managing Re-identification Risk

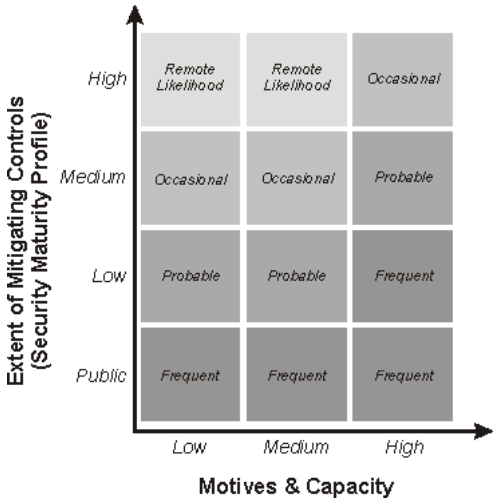
Contents
- Definitions
- Risk
- Method
- REB
- Lessons
- End



K. El Emam, F. Dankar, R. Vaillancourt, T. Roffey, and M. Lysyk: "Evaluating Patient Re-identification Risk from Hospital Prescription Records." In the Canadian Journal of Hospital Pharmacy, 62(4):307-319, 2009.

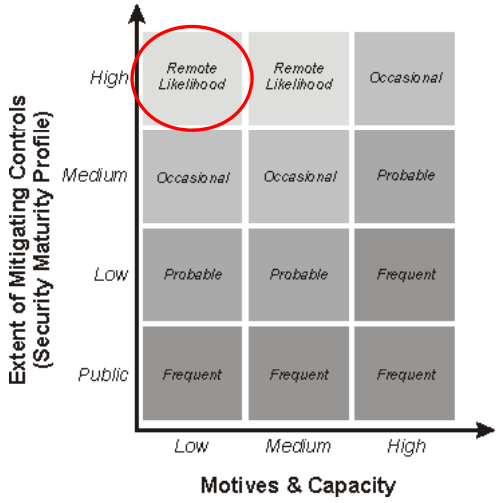
Determining Pr Re-identification Attempts

Contents
- Definitions
- Risk
- Method
- REB
- Lessons
- End



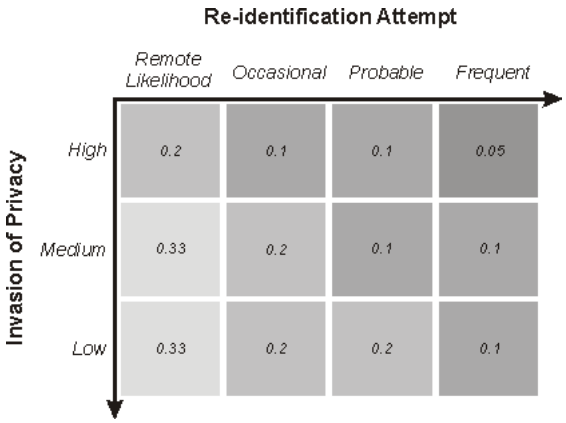
Determining Pr Re-identification Attempts

Contents
- Definitions
- Risk
- Method
- REB
- Lessons
- End

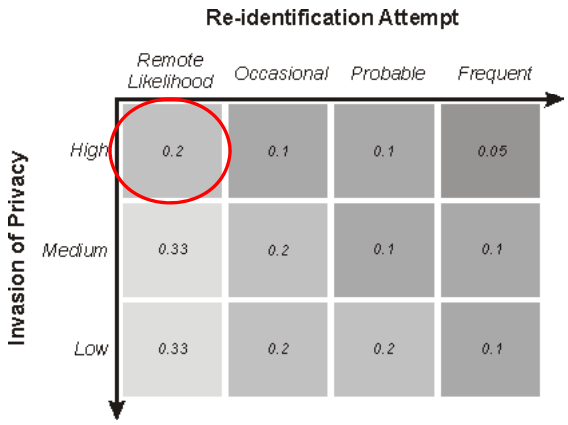


Determining Risk Threshold to Use

Contents
- Definitions
- Risk
- Method
- REB
- Lessons
- End



Determining Risk Threshold to Use



Role of the REB

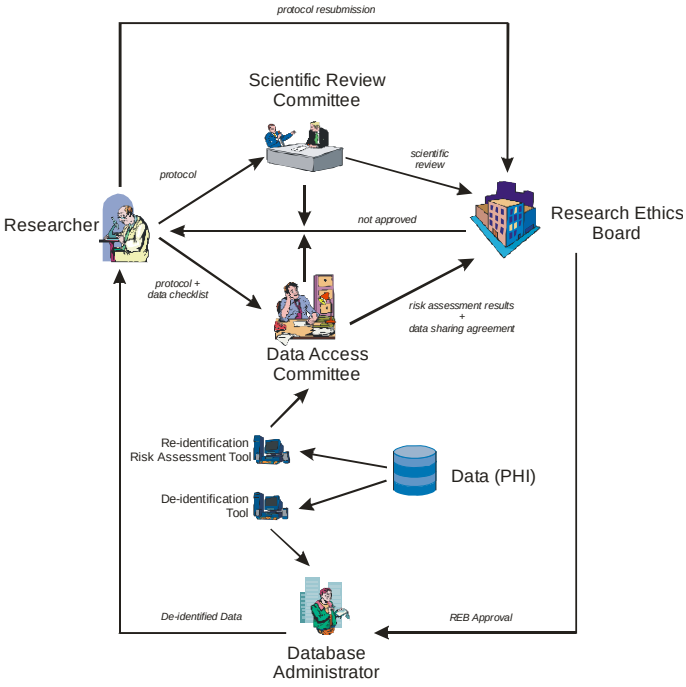
- The REB should still review protocols that claim to use de-identified data:
  - Not possible to determine whether it is de-identified without a proper risk assessment
  - Investigator have an inherent conflict when making self declarations of identifiability
  - There may be other ethical considerations beyond just the privacy question (e.g., group harms)

Incorporating De-identification

Contents

- Definitions
- Risk
- Method
- REB
- Lessons
- End

- Re-identification risk assessment should be done before protocols are submitted to the REB
- A “certificate” with re-identification risk assessment results is produced and submitted with the protocol
- The REB focuses on ethical issues if the “certificate” shows that the data will be de-identified



Lessons Learned

Contents

- Definitions
- Risk
- Method
- REB
- Lessons
- End

- This approach provides an incentive for the data requestor to improve their security and privacy practices
- Ensures that the amount of de-identification is proportionate to the risk
- Improves the relationship between the custodian and the data requestor

www.ehealthinformation.ca

www.ehealthinformation.ca/knowledgebase