

Implementing Information Governance Controls

Nathan Lea

CHIME, UCL
UK CAB Community Training Day
18th June 2010, MRC Clinical Trials Unit

Adapted from Materials by Nathan Lea, Dipak Kalra and Peter Singleton

Introduction

- Sensitive data use in an enterprise is governed by policy based controls
- The controls involve private networks, encryption, firewalls, access controls, authentication, data masking and so forth
- There has been an increasing emphasis on stakeholder engagement over the last decade to paint a more complete picture on how these controls should be enacted
- Stakeholders, of course, include community reps!
- The focus today will be on De-identification and Statistical Disclosure Control

Introduction



- De-Identification - Anonymisation, Pseudonymisation and Masking
- Statistical Disclosure Control
- An artist was commissioned to work with the concept of security
- The result was to produce a pastel piece which emphasises contrast, juxtaposition and balance
- You may find this to be an eloquent expression of the challenge that faces projects that use sensitive data where security is concerned

De-Identification

- De-Identification is the process by which an identifying record is rendered “unidentifying”
- It enacts the methods you may have heard about:
 - ***anonymisation***, where the association between an individual and their data has been severed so that they cannot be identified from it
 - ***pseudonymisation***, where the associations remain, but are represented in an unrecognisable fashion (like an alias)

De-Identification

- There are a number of ways to implement the process
- Available techniques depend on what kind of data needs to be treated in order to remove or mask identifying features
- It may be helpful to think of masking or obfuscation
- It is something of a balancing act between protecting individuals' identities and not compromising the purpose of data use

De-Identification Techniques

- String pattern matching and removal
- Date removal or masking - releasing only the year of birth, or a Julian Date
- Numeric Value ranges (CD4s below 200 as opposed to exact CD4 Count)
- Pseudonymous ID generation
- Aggregate data release for population queries across a cohort

Statistical Disclosure Control

- Methods to calculate the statistical likelihood of individual identification exist depending on:
 - what data is being sought
 - what data has already been released
 - whether there is a querying pattern that is attempting to identify an individual from a pool of records.

Statistical Disclosure Control

- Analysis involves Bayesian modelling of cell counts and graphical analyses
- The results of the analysis can then be used to help decide whether the data should be released at all, or whether stronger de-identification coupled with aggregation of results should be employed.
- Not unlike an email spam filter where a score is applied to certain emails to flag them as being potentially unsolicited

Thank You!

Any Questions?

n.lea@chime.ucl.ac.uk