



PRIVACYANALYTICS
Data Anonymization Solutions



Standards and Guidelines for Risk-Based De-identification of Clinical Trial Data

Khaled El Emam

www.privacy-analytics.com
613.369.4313

251 Laurier Avenue, Suite 200
Ottawa, Ontario, Canada K1P 5J6

Legal Perspective



- The anonymization methods used must be acceptable to regulators around the world, and they must be capable of passing an audit, investigation, or litigation by participants
- Following existing standards becomes important under such circumstances
- Most jurisdictions today follow a “reasonableness” standard

Standards & Guidelines



A number of contemporary standards and guidelines recommend a risk-based approach to de-identification

© 2015 Privacy Analytics, Inc.

3

Risk-based De-identification



- Key characteristics of risk-based de-identification:
 - The standard approach in the disclosure control community
 - Re-identification **risk measurement** is an integral part of the process
 - The **context of the data disclosure** is taken into account when deciding what perturbations or transformations to apply to the data
 - The process is evidence-based and builds on strong precedents
 - Easier to justify and defend in multiple jurisdictions

© 2015 Privacy Analytics, Inc.

4

Dimensions of Data Release



- Three dimensions to consider (representing current practices):
 - Mechanism of data release (microdata vs portal)
 - Approach to data release (public, quasi-public, non-public)
 - Approach to anonymization (proactive vs on-demand)
- Each of these options has advantages and disadvantages
- For example, Project Data Sphere:
 - Microdata
 - Quasi-public
 - Proactive

© 2015 Privacy Analytics, Inc.

5

Data Release Mechanism



- **Microdata release:** individual-level participant data (IPD) is being provided to data recipients as flat files (CSV or SAS) or database files
 - Microdata can be public or available through controlled access
- **Online portal:** data recipients can access IPD through a portal and perform their analysis through the portal only
 - No raw data download allowed (different control mechanisms used)
 - Online portal registration can be public or through a qualification process

© 2015 Privacy Analytics, Inc.

6

Approach to Data Release



	Public Data Release	Quasi-Public Data Release	Non-public Data Release
Anyone can access data	X	X	
Verify identity		X	X
Sign a contract		Limited Terms	Extensive Controls

© 2015 Privacy Analytics, Inc.

7

Approaches to Anonymization



- Proactive Anonymization:
 - Anonymization performed before data requests
 - Can be very efficient
 - Unable take into account the specifics of the data analysis
- On-demand Anonymization:
 - Can calibrate the anonymization to the anticipated analysis – higher quality data
 - Anonymization must be performed for each data release of the same data set (higher cost)

© 2015 Privacy Analytics, Inc.

8

Anonymizing Portal Access



- Is it necessary to anonymize data if it is on a portal ?
 - There are three types of attack:
 - **Deliberate attack by recipient** – manage that risk through contracts and audit trails
 - **Data breach** – managed by manufacturer through portal controls
 - **Inadvertent re-identification** – could happen if data recipient lives in the same geography as some the participants
 - It is mostly inadvertent disclosure risk that needs to be managed in a portal – anonymization is still needed
 - But there may be residual risks from other attacks – this needs to be actually assessed not assumed

© 2015 Privacy Analytics, Inc.

9

Anonymizing Public Data

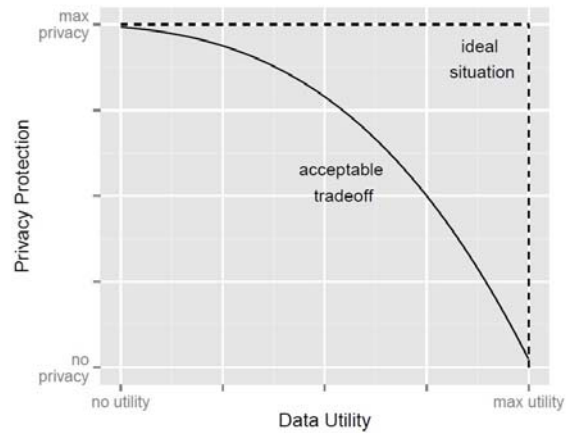


- The key consideration is a **demonstration attack**
- Adversary uses public or private information to re-identify individuals in the data set
- There are multiple examples of demonstration attacks occurring – this is somewhat common
- There are no controls, therefore the anonymization perturbs the data quite a bit

© 2015 Privacy Analytics, Inc.

10

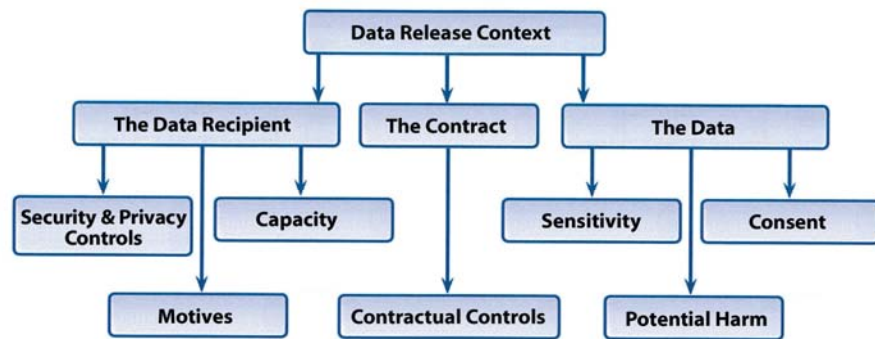
Data Quality Balance



© 2015 Privacy Analytics, Inc.

11

Data Release Context



© 2015 Privacy Analytics, Inc.

12

Direct & Quasi Identifiers



Examples of direct identifiers: MRN, health card number, health plan beneficiary number, VID, photograph, biometrics, SSN, SIN, device number, clinical trial record 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, total years of schooling, marital status, criminal history, total income, visible minority status, profession, event dates, number of children, high level diagnoses and procedures

Direct & Quasi Identifiers



Examples of direct identifiers: MRN, health card number, health plan beneficiary number, VID, photograph, biometrics, SSN, SIN, device number, clinical trial record 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, total years of schooling, marital status, criminal history, total income, visible minority status, profession, event dates, number of children, high level diagnoses and procedures

FOLLOW PHUSE STANDARD

CT Classification of Variables



- **Identifiers:** Subject ID, Device ID, Reference ID, Sponsor ID, Investigator ID, Site ID
- **Demographics:** Country, Sex, Race, Ethnicity, Date of Birth, Date of Death
- **Visits:** Dates
- **Serious Adverse Events:** HLT/HLGT, Date, Outcome, Death, Hospitalization, Life Threatening
- **Medical History:** Condition or event, Date
- **Free Text:** Remove or Anonymize (95% recall)



Equivalence Classes



Gender	Year of Birth (10 years)	NDC
Male	1970-1979	0009-0031
Male	1980-1989	0023-3670
Male	1970-1979	0074-5182
Female	1990-1999	0078-0379
Female	1980-1989	65862-403
Male	1990-1999	55714-4446
Male	1990-1999	55714-4402
Female	1980-1989	55566-2110
Male	1970-1979	55289-324
Female	1990-1999	54868-6348
Male	1980-1989	53808-0540

Equivalence Classes



Gender	Year of Birth (10 years)	NDC
Male	1970-1979	0009-0031
Male	1980-1989	0023-3670
Male	1970-1979	0074-5182
Female	1990-1999	0078-0379
Female	1980-1989	65862-403
Male	1990-1999	55714-4446
Male	1990-1999	55714-4402
Female	1980-1989	55566-2110
Male	1970-1979	55289-324
Female	1990-1999	54868-6348
Male	1980-1989	53808-0540

Equivalence Classes



Gender	Year of Birth (10 years)	NDC
Male	1970-1979	0009-0031
Male	1980-1989	0023-3670
Male	1970-1979	0074-5182
Female	1990-1999	0078-0379
Female	1980-1989	65862-403
Male	1990-1999	55714-4446
Male	1990-1999	55714-4402
Female	1980-1989	55566-2110
Male	1970-1979	55289-324
Female	1990-1999	54868-6348
Male	1980-1989	53808-0540

Risk Measurement Example

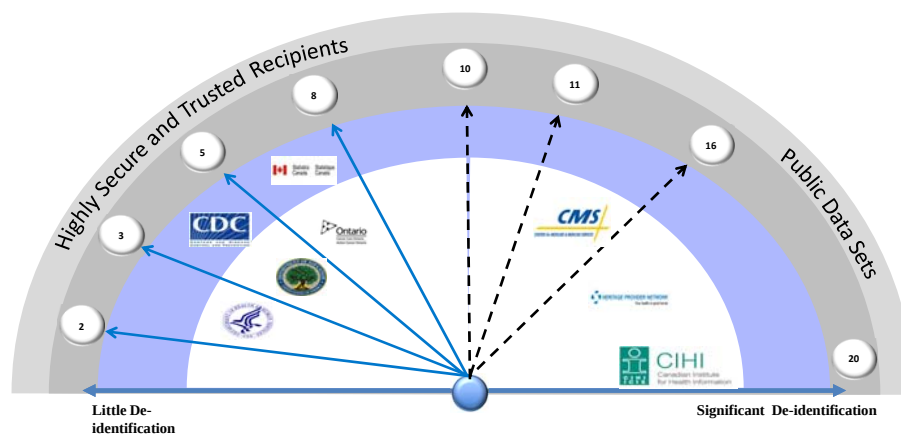


DoB	Date of Recruitment	Location	Sex	Probability of Re-identification
X				0.014
	X			0.005
X	X			0.88
X	X	X		0.98
X	X	X	X	1.0
X	X		X	0.91
	X	X		0.98
X		X		0.85
		X		0.19

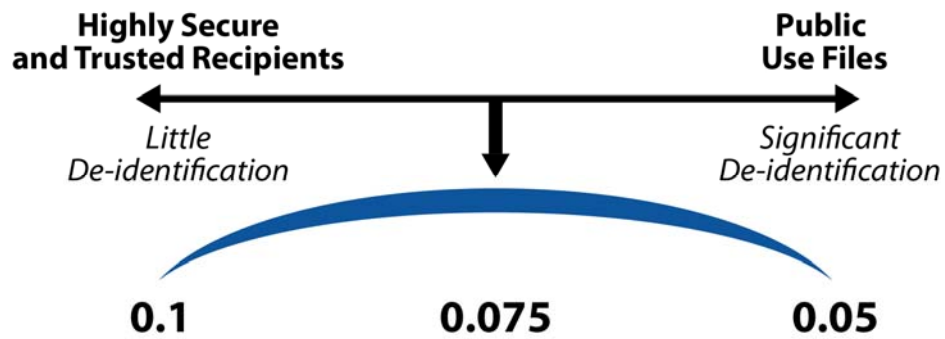
Impact of Generalization

	DoB	Date Rec	Location	Sex	Probability of Re-identification
S1	Year	dd/mm/yy	City	No Change	0.973
S2	Year	mm/yy	City	No Change	0.677
S3	5-year	mm/yy	City	No Change	0.327
S4	19<; 10-year; >40	mm/yy	City	No Change	0.23
S5	19<; 10-year; >40	mm/yy	Country	No Change	0.007
S6	Year	mm/yy	Country	No Change	0.034
S7	5-year	quarter/yy	City	No Change	0.152
S8	19<; 10-year; >40	quarter/yy	City	No Change	0.1

Spectrum of Identifiability

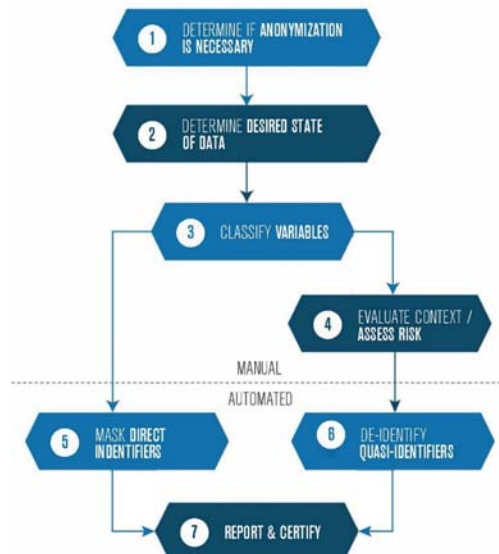


Re-identification Risk Spectrum



© 2015 Privacy Analytics, Inc.

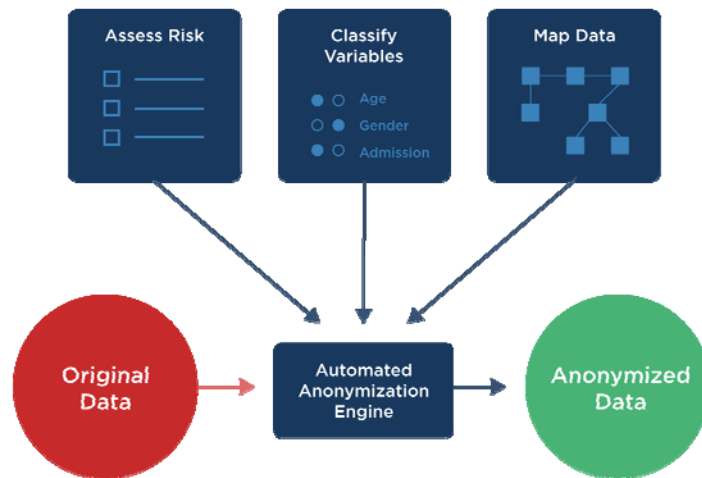
23



© 2015 Privacy Analytics, Inc.

24

Automation



Implementation

- At this point, typically one week turn around time for a trial data set, with a few days of effort, including the production of documentation and certification
- Knowledge of clinical trial data is necessary to perform the anonymization, ie, some technical background needed
- Training and certification programs available
- Some companies outsourcing most of the anonymization work rather than develop in-house expertise

Thank You



A black and white photograph of a hand-drawn "Thank you" in cursive script. A pen is shown in the process of writing the final stroke of the word "you".

kelemam@privacy-analytics.com