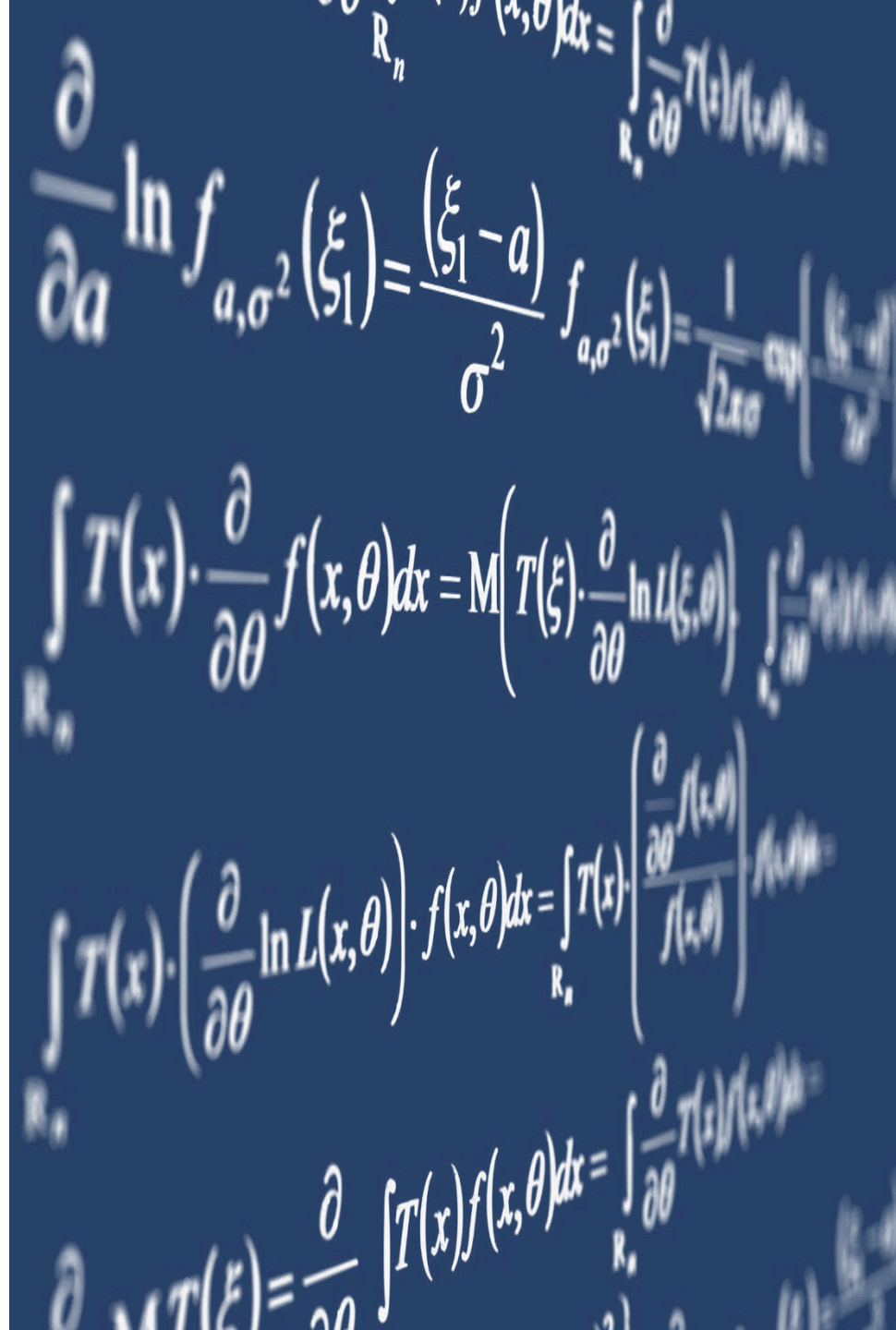


DH09

Calculating
RISK
of
Re-Identification
of
Patient-Level Data
using
Quantitative
Approach

Lukasz Kniola

Biogen
Maidenhead
UK



DISCLAIMER

The scope of this paper is to present the opinions and suggestions of the author. The interpretations of standards and procedures contained in this paper are those of the author and are not necessarily correct. Any views and recommendations stated within this paper are those of the author, and they do not represent the positions of their employer.

Graphics and illustrations:
designed by  **freepik.com**

AGENDA

1. **CONTEXT** of data release

2. **IDENTIFIERS** in the data

3. **PROBABILITY**
of **RE-IDENTIFICATION**

Attempt

Success

Uniqueness

4. **THRESHOLDS & BALANCE**

5. **EMA Policy 0070**



PLAN

Data (MS study):

USUBJID	SEX	AGE	COUNTRY
101-002	M	30	USA
101-006	F	31	USA
101-009	M	29	USA
102-002	M	29	USA
102-013	F	28	USA
103-007	M	26	USA
201-001	F	28	USA
301-111	M	29	USA
302-122	F	32	USA
401-001	F	31	USA

Recipient:

- Trusted researcher
- Collaboration history
- Data Sharing Agreement signed
- Data to be transferred over sFTP

CONTEXT OF DATA RELEASE

CONTROLLED RELEASE

Relationship
With Requestor

Contract
Present / Enforceable

Data Selection
All v Subset

APPROVED

PUBLIC RELEASE



IDENTIFIERS IN THE DATA



USUBJID	SEX	AGE	COUNTRY
101-002	M	30	USA
101-006	F	31	USA
101-009	M	29	USA
102-002	M	29	USA
102-013	F	28	USA
103-007	M	26	USA
201-001	F	28	USA
301-111	M	29	USA
302-122	F	32	USA
401-001	F	31	USA



CALCULATE...

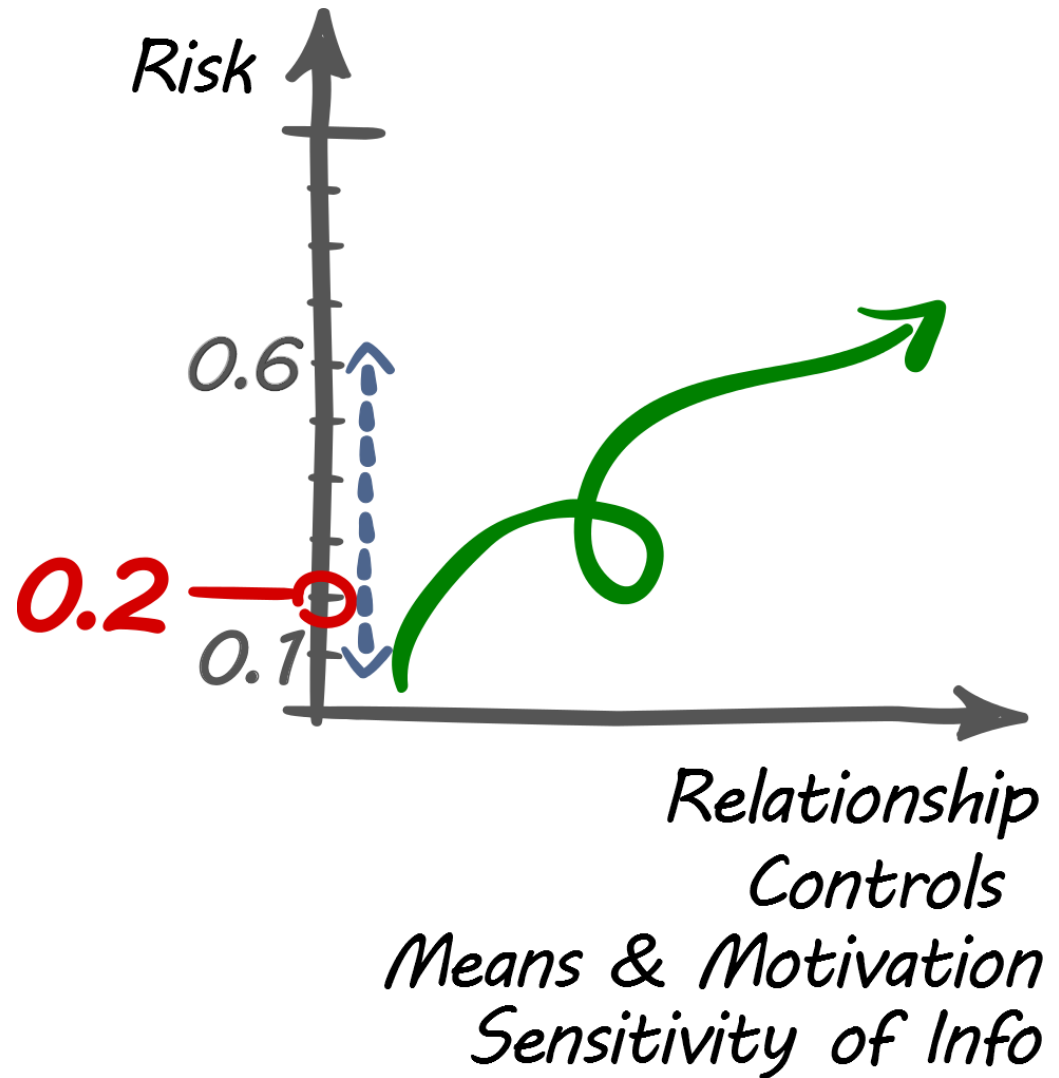
$$\Pr(\text{re-id}) = \Pr(\text{attempt}) \times \Pr(\text{re-id} \mid \text{attempt})$$

$$\hat{p}(\text{unique}) = ?$$



Attempt (I):

Deliberate attack



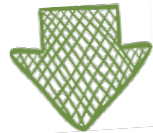
Attempt (II):

Inadvertent Re-Identification

	MS Cases	Overall Population
US	400K	317M
Worldwide	2.3M	7.2B

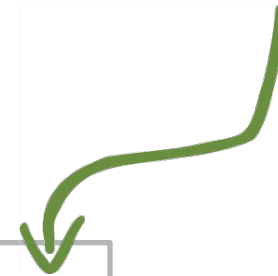


$$Pr(acquaintance) = 1 - \left(1 - \frac{\text{Cases}}{\text{Population}} \right)^{150}$$



	Pr(Acquaintance)
US	0.17
Worldwide	0.05

Dunbar's Number





Attempt (III): Data Breach

himss 0.27

Security measures
Access restrictions



0.14

CALCULATE...

$\text{Max}(0.2, 0.17, 0.27)$



$$\text{Pr}(\text{re-id}) = 0.27 \times \text{Pr}(\text{re-id} \mid \text{attempt})$$

$$\hat{p}(\text{unique}) = ?$$



Probability of Successful Re-Identification

USUBJID	SEX	AGE	COUNTRY
CT1/101	M	26	USA
CT1/102	F	28	USA
CT1/103	F	31	USA
CT1/104	M	29	USA
CT1/105	F	28	USA
CT1/106	M	30	USA
CT1/107	M	29	USA
CT1/108	F	32	USA
CT1/109	M	29	USA
CT1/110	F	31	USA



Probability of Successful Re-Identification

				Study population	
USUBJID	SEX	AGE	COUNTRY	Equiv. Class (Size)	Re-Id risk
CT1/101	M	26	USA	A (1)	1
CT1/102	F	28	USA	B (2)	0.5
CT1/103	F	31	USA	C (2)	0.5
CT1/104	M	29	USA	D (3)	0.33
CT1/105	F	28	USA	B (2)	0.5
CT1/106	M	30	USA	E (1)	1
CT1/107	M	29	USA	D (3)	0.33
CT1/108	F	32	USA	F (1)	1
CT1/109	M	29	USA	D (3)	0.33
CT1/110	F	31	USA	C (2)	0.5
Pr_{max}				1	
Pr_{avg}				0.6	



Probability of Successful Re-Identification

				Study pop	Similar trials	
USUBJID	SEX	AGE	COUNTRY	Re-Id risk	Equiv. Class Size	Re-Id risk
CT1/101	M	26	USA	1	(12)	0.083
CT1/102	F	28	USA	0.5	(32)	0.031
CT1/103	F	31	USA	0.5	(27)	0.037
CT1/104	M	29	USA	0.33	(11)	0.091
CT1/105	F	28	USA	0.5	(32)	0.031
CT1/106	M	30	USA	1	(15)	0.067
CT1/107	M	29	USA	0.33	(11)	0.091
CT1/108	F	32	USA	1	(4)	0.250
CT1/109	M	29	USA	0.33	(11)	0.091
CT1/110	F	31	USA	0.5	(27)	0.037
Pr_{max}				1	0.25	
Pr_{avg}				0.6	0.081	

Similar trials

SEX	AGE	COUNTRY	Equiv. Class Size
M	26	USA	12
F	28	USA	32
...	...		...
...	...		...



CALCULATE...

$$\Pr(\text{re-id}) = 0.27 \times 0.6 = 0.162$$

$$\hat{p}(\text{unique}) = ?$$



Uniqueness and k -anonymity

USUBJID	SEX	AGE	COUNTRY
CT1/101	M	26	USA
CT1/102	F	28	USA
CT1/103	F	31	USA
CT1/104	M	29	USA
CT1/105	F	28	USA
CT1/106	M	30	USA
CT1/107	M	29	USA
CT1/108	F	32	USA
CT1/109	M	29	USA
CT1/110	F	31	USA

Study population

Equiv. Class (Size)	Re-Id risk
A (1)	1
B (2)	0.5
C (2)	0.5
D (3)	0.33
B (2)	0.5
E (1)	1
D (3)	0.33
F (1)	1
D (3)	0.33
C (2)	0.5



$k=2$

% of non k -anonymous records = 30%

CALCULATE...

$$\text{Pr}(\text{re-id}) = 0.27 \times 0.6 = 0.162$$

$$\hat{p}(\text{unique}) = 30\%$$





0.05

0.33

**THRESHOLD
0.09**

+ NO unique records

$$\text{Pr}(\text{re-id}) = 0.27 \times 0.6 = 0.162$$

$$\hat{p}(\text{unique}) = 30\%$$



APPLY DE-ID TECHNIQUES...

USUBJID	SEX	COUNTRY	Study population		
			AGE _{DE-ID}	Equiv. Class (Size)	Re-Id risk
CT1/101	M	USA	21-30	X (5)	0.2
CT1/102	F	USA	21-30	Y (2)	0.5
CT1/103	F	USA	31-40	Z (3)	0.33
CT1/104	M	USA	21-30	X (5)	0.2
CT1/105	F	USA	21-30	Y (2)	0.5
CT1/106	M	USA	21-30	X (5)	0.2
CT1/107	M	USA	21-30	X (5)	0.2
CT1/108	F	USA	31-40	Z (3)	0.33
CT1/109	M	USA	21-30	X (5)	0.2
CT1/110	F	USA	31-40	Z (3)	0.33

$$\Pr_{\max} = 0.5$$

$$\Pr_{\text{avg}} = 0.3$$



... AND RECALCULATE

$$\Pr(\text{re-id}) = 0.27 \times 0.3 = 0.081 \checkmark$$

$$\hat{p}(\text{unique}) = 0\% \checkmark$$



BALANCE

BETWEEN **RISK** AND **DATA UTILITY**





WRAP-UP

- **Context** (relationship , contract, means, sensitivity).
- **Direct** and **Quasi**-identifiers and de-id techniques.
- **Study** pop v **Similar Trials**.
- **Average** v **Maximum** risk.
- **k-anonymity** – select k (and proportion allowed).
- Selection of **Thresholds** and comparison of results.
- **Data Utility** considerations.

DOCUMENT
all assumptions
and justifications

EMA Policy 0070

- Only clinical reports
(no patient-level data in Phase I)
- Narratives within scope
- Public release:
 - Risk of re-id **attempt is 1**
 - **Maximum risk** across data set
 - Suggested threshold is **0.09**



BOOKS & RESOURCES



- El Emam, K. (2013).
Guide to the De-Identification of Personal Health Information.
Boca Raton, FL: CRC Press.
- El Emam, K., Arbuckle, L. (2013).
Anonymizing Health Data Case Studies and Methods to Get You Started.
Sebastopol, CA: O'Reilly Media.
- Institute of Medicine (2015).
Sharing Clinical Trial Data Maximizing Benefits, Minimizing Risk.
Washington, DC: The National Academies Press.
- PhUSE
De-identification standards for CDISC SDTM 3.2.
- European Medicines Agency. (2014).
EMA/240810/2013 - Publication of clinical data for medicinal products for human use.
- European Medicines Agency. (2016).
EMA/90915/2016 - External guidance on the implementation of the European Medicines Agency policy on the publication of clinical data for medicinal products for human use.

Thank You!

lukasz.kniola@biogen.com

