Risk of re-identification of de-identified data

ELSI online workshop bescherming patiëntgegevens in wetenschappelijk onderzoek – 19 februari 2021

Eric Sijbrands, dept. of Internal Medicine, Erasmus MC

Healthcare – as identifiable as possible

Healthcare – as identifiable as possible

identify for high quality of care  keep information confidential

technical solutions to prevent unauthorized access to the hospital information system

SOP's

ISO/IEC 27001	directive to establish an information security management system (ISMS)
ISO 27002 ISO 27799	implementation (PDCA) of ISMS processes
NEN 7510	- availability - integrity - confidentiality
NEN 7512	data exchange
NEN 7513	logging
NEN 7521	access to patient records (for example, only with two-factor authentication)

6 aanhoudingen

Research – de-identified data

personal information

identified

pseudomized
direct identifiers -

de-identified
direct identifiers -
known indirect identifiers -

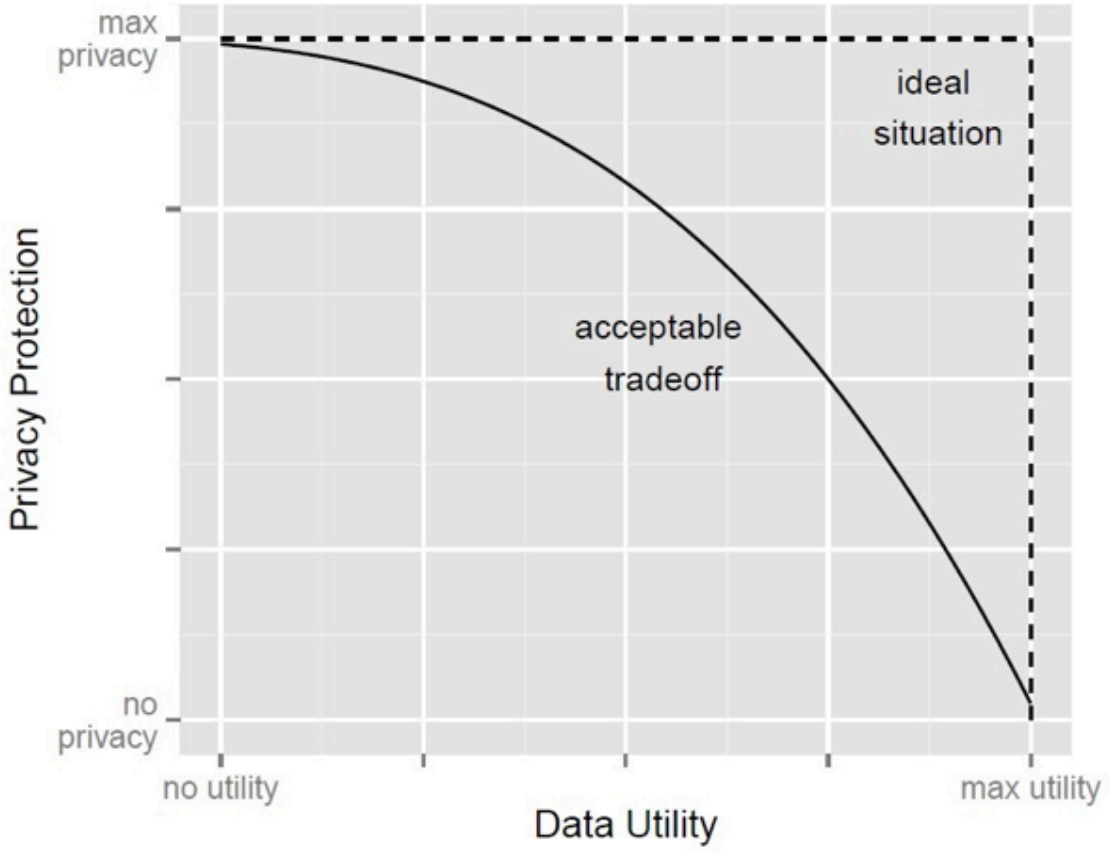
anonymized
direct identifiers -
indirect identifiers -

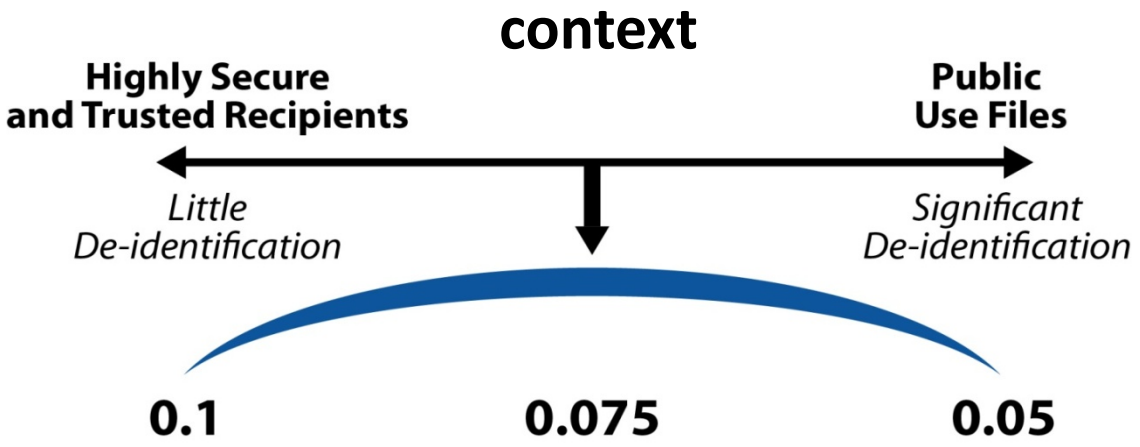
no: - singling out
- linkability
- inference

threshold

eliminate/transform direct identifiers:
name, telephone number, insurance number, medical record number, etc.
eliminate/transform indirect/quasi-identifiers:
sex, date of birth, ethnicity, location (postal codes), event dates, medical codes, income, profession, etc.

Risk of re-identification





El Emam, Arbuckle. Chapter 2 in Anonymizing Health Data. 2013: 22 and 39.

estimate risk **transform data**

Risk of re-identification – example 1

free text column “Occupation”

answer: Mayor of Rotterdam

Risk of re-identification – example 2

Medical Records

ID	Name	Gender	Date of Birth	Hometown	Diagnosis
1	Sister Susie	F	1/1/1953	Lafayette, IN	Myeloid leukemia
2	Jack Sprat	M	3/15/1953	Lafayette, IN	Hypertension
3	Mary Contrary	F	2/28/1953	Washington, IN	Myocardial infarction
4	Boy Blue	M	7/4/1955	Washington, IN	Myocaridal infarction
5	King Cole	M	3/31/1957	Lafayette, IN	Diabetes
6	Jill Hill	F	1/12/1955	Washington, IN	Diabetes
7	Jack Hill	M	1/12/1955	Washington, IN	hypertension

pseudomized

Released Dataset

Gender	Year of Birth	Diagnosis
F	1953	Myeloid leukemia
M	1953	Hypertension
F	1953	Myocardial infarction
M	1955	Myocardial infarction
M	1955	Hypertension

Voter List

Name	Gender	Year of Birth
Sister Susie	F	1953
Jack Sprat	M	1953
Mary Contrary	F	1953
Boy Blue	M	1955
King Cole	M	1957
Jill Hill	F	1955
Jack Hill	M	1955
Jane Goose	F	1957
Jack Nimble	M	1956
Betty Blue	F	1956

Figure 1 Example of de-identification and re-identification using public records.

Benitez. J Am Med Inform Assoc 2010;17:169-77.

Risk of re-identification – example 3

Original Prescription Database

IDENTIFYING VARIABLE	QUASI-IDENTIFIERS		
Name	Gender	Year of Birth	ID
John Smith	Male	1979	2046059
Alan Smith	Male	1982	716839
Hercules Green	Male	1979	2241497
Gill Stringer	Female	1995	2046059
Marie Kirkpatrick	Female	1986	392537
Bill Nash	Male	1995	363766
Albert Blackwell	Male	1998	544981
Alice Smith	Female	1987	293512
Douglas Henry	Male	1979	544981
Freda Shields	Female	1995	596612
Fred Thompson	Male	1987	725765

De-identification

QUASI-IDENTIFIERS		
Gender	Decade of Birth	ID
Male	1970-1979	2046059
Male	1980-1989	716839
Male	1970-1979	2241497
Female	1990-1999	2046059
Female	**1980-1989**	**392537**
Male	1990-1999	363766
Male	1990-1999	544981
Female	**1980-1989**	**293512**
Male	1970-1979	544981
Female	1990-1999	596612
Male	1980-1989	725765

Alice Smith, Female, 1987

El Emam. Can J Hosp Pharm 2009;62:307-19.

Risk of re-identification

overall risk = data risk x context risk

data risk = 1/minimal equivalence class size (= maximum risk)

context risk = P(re-identification|**deliberate** attempt) x P(attempt)
P(re-identification|**inadverted** attempt) x P(acquainted)
P(re-identification|**breach**) x P(breach)

Context risk of re-identification

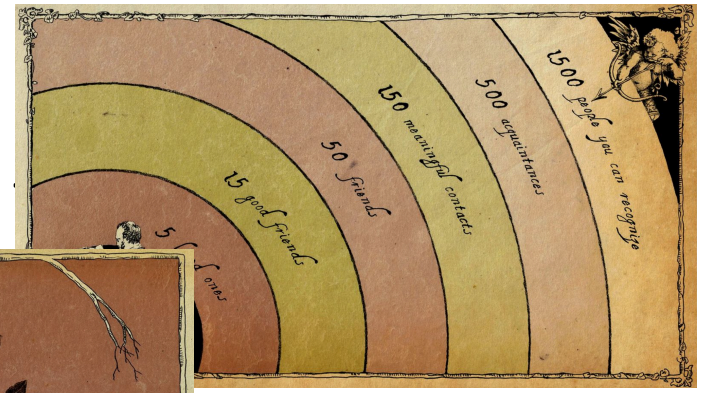
P(re-identification|**deliberate** attempt) x P(attempt)

P(re-identification|**inadverted** attempt) x P(acquainted)

P(re-identification|**breach**) x P(breach)

security \ trust	high	medium	low
high	0.05	0.10	0.20
medium	0.20	0.30	0.40
low	0.40	0.50	0.60

Dunbar's number: 148

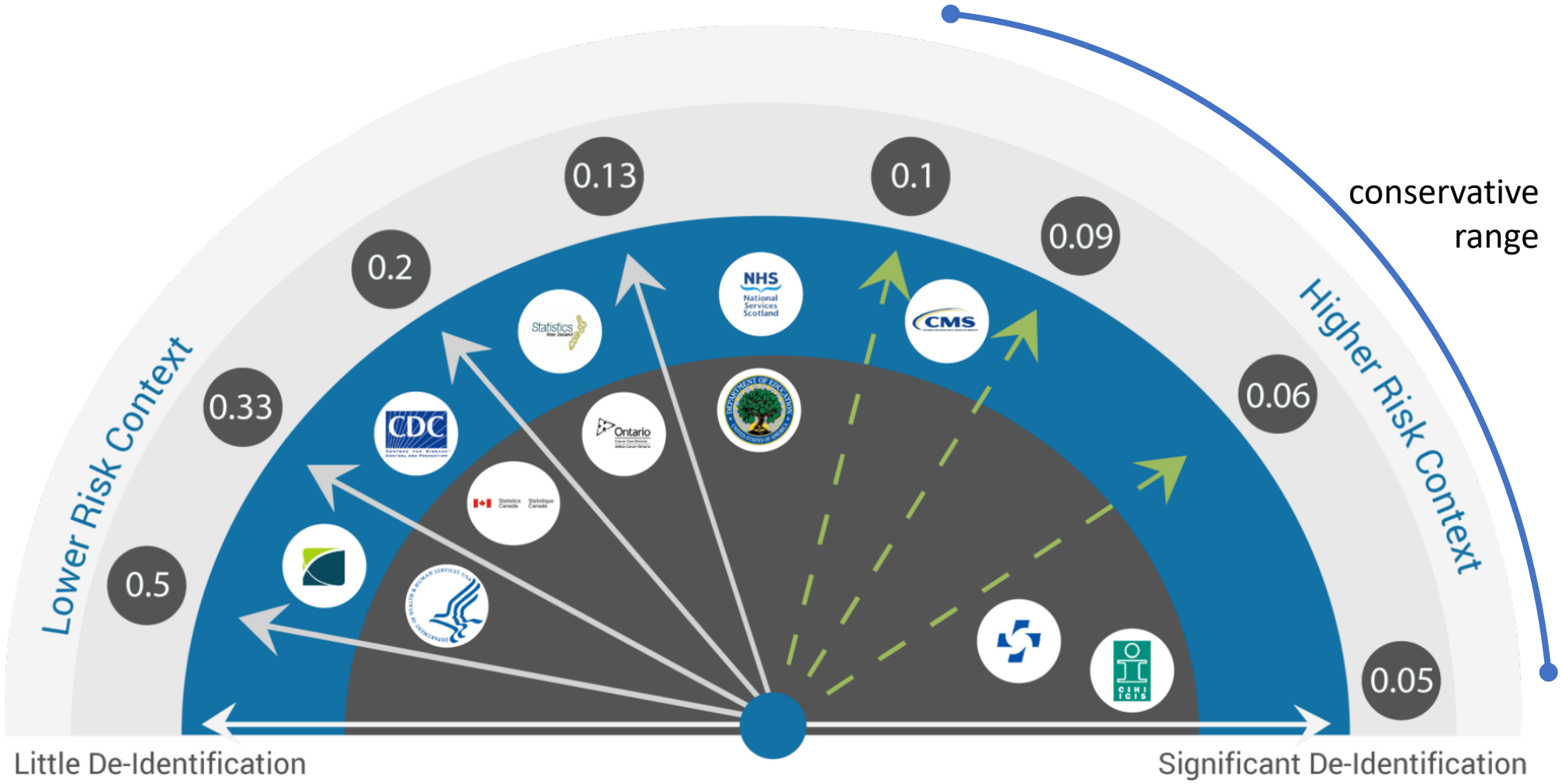


Risk of re-identification

overall risk = data risk x context risk

context risk =	P(re-identification|deliberate attempt) x P(attempt)	0.3-0.4
	P(re-identification|inadverted attempt) x P(acquainted)	$1-(1-P)^{150}$
	P(re-identification|breach) x P(breach)	0.27 or 0.14

Risk threshold



EU ICU database – example 4

AmsterdamUMCdb - intro

- European ICU database (MIMIC/eICU lookalike)
- Single university medical center
- N=23500
- Yearly n=20000
- Microsoft SQL server web edition: 7 tables
- On request access

AmsterdamUMCdb – de-identification

- UI and admission numbers — with software changed in random numbers
- dates — miliseconds after admission to ICU (t=0)
- age — 5-year categories
- admission year — 3-year categories
- body weight — rounded to integer values
- all free text notes — eliminated
- all images — eliminated
- usernames — replaced by userID's
- notes, pictures, comments, part of the free text fields were eliminated
- all remaining free text: other — searched for all names of patients and staff and identifiable items
- names of units/departments — codes
- server names, database names and personal computer names were de-identified
- identifyer key is destroyed after each update

AmsterdamUMCdb – in addition

- opt-out and right to be forgotten procedures available
- controlled access protocol
- delegation log:

  privileges, tasks, access rights during production
- education log:

  date of instruction on how to work with the AmsterdamUMCdb, including ethical standards, respecting privacy, prohibition of reidentification attempts by those having access to source data and safety of the Patient Data Management System and Hospital Information System.

Discussion

- Anonymous by design
- Big data ⟶ all rows are unique
- Risk of re-identification
  - large bulk of measurements produces individual codes: quasi-identifiers
  - measurements are not replicable, distinguishable, and knowable
  - re-identification risk = 0 without unauthorized access to HIS
  - in case of unauthorized access to HIS re-identification is not relevant, as quick and easy direct identification is available