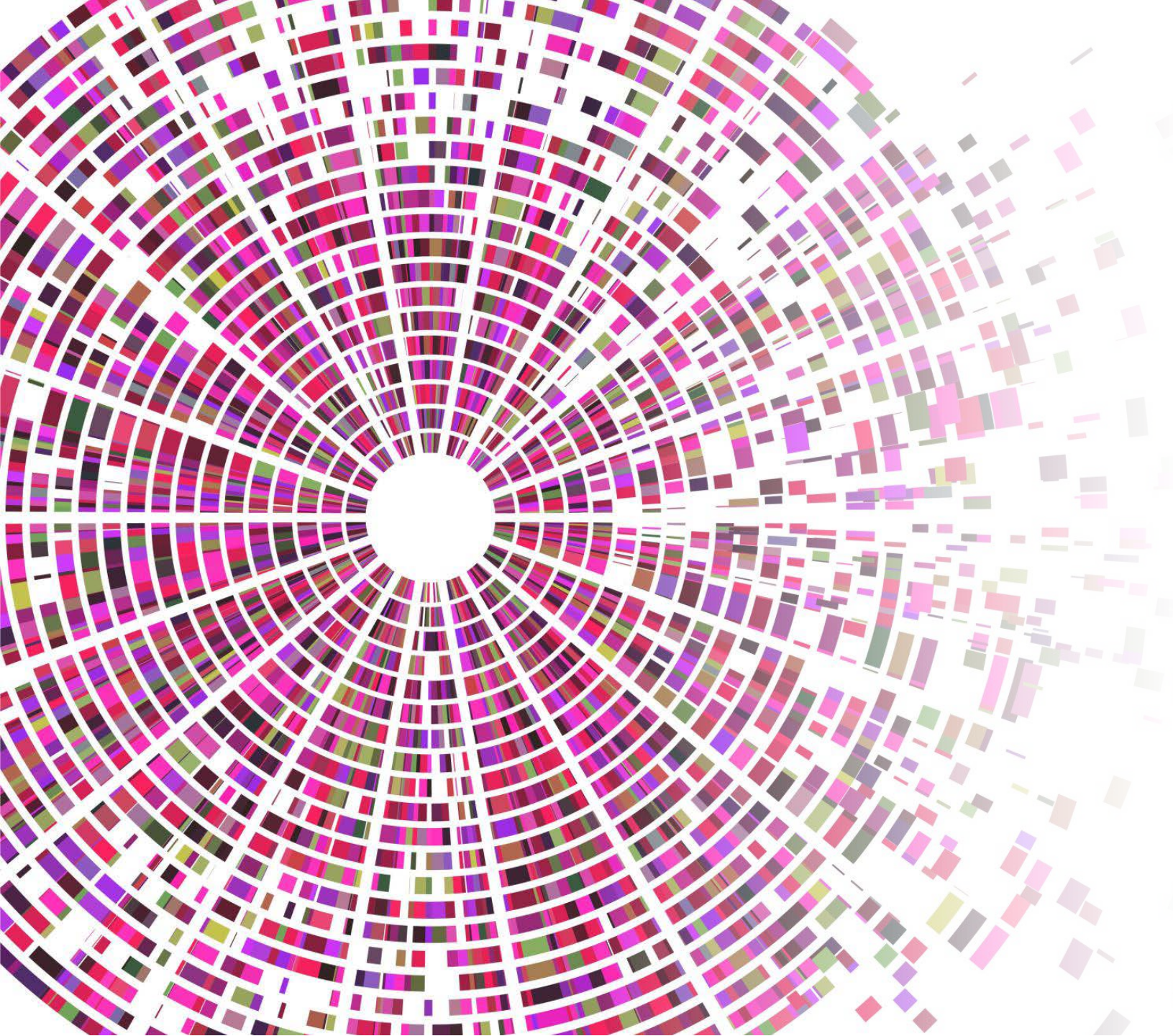


Anonymization and re-identification risk analysis at the Cancer Registry of Norway

INF5130

7th October, 2021

SAGAR SEN, Senior Research Scientist, SINTEF

Outline

- **Attacks on privacy and their consequences**
- Anonymization criteria (k-anonymity, l-diversity, t-closeness)
- Data curation at the Cancer Registry of Norway (CRN)
- Sharing data for research @ CRN
- Data fuzzification at the Cancer Registry of Norway
- Re-identification Risk Analysis with ARX
- Conclusion

Prosecutor/background knowledge attack

Employer is trying to find the test result of **28 year old male doctor**

ID	Sex	Age	Profession	Drug test
1	Male	37	Doctor	Negative
2	Female	28	Doctor	Positive
3	Male	37	Doctor	Negative
4	Male	28	Doctor	Positive
5	Male	28	Doctor	Negative
6	Male	37	Doctor	Negative

In the *prosecutor* model the attacker targets a specific individual and it is assumed that she already knows (e.g. employer in a company) that data about the individual is contained in the dataset.

Journalist attack

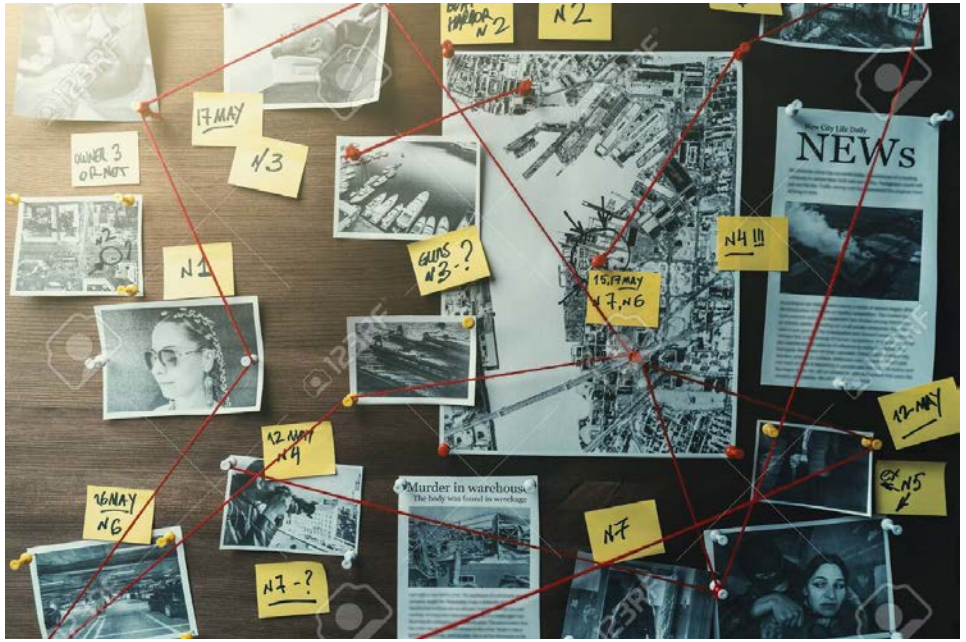
Published Data

	Non-Sensitive Data			Sensitive Data
#	Zip	Age	Nationality	Condition
1	13053	28	Indian	Heart Disease
2	13067	29	American	Heart Disease
3	13053	35	Canadian	Viral Infection
4	13067	36	Japanese	Cancer

Data leak!

#	Name	Zip	Age	Nationality
1	John	13053	28	American
2	Bob	13067	29	American
3	Chris	13053	23	American

Voter List



In the *journalist* model the attacker targets a specific individual but it is not expected that she possesses background knowledge about membership. However, the journalist has access to a public database.

Marketer attack

With a certain property – e.g. heart disease

An equivalence class or a group of an anonymized table is **a set of records with the same values for the quasi-identifier attributes**

Equivalence class		Anonymized table		Public database		Probability of match
Gender	Age	Count	Record number	Count	Record number	
Male	1950-1959	3	1,4,12	4	1,4,12,27	3/4
Male	1960-1969	2	2,14	5	2,14,15,22,26	2/5
Male	1970-1979	2	9,10	5	9,10,16,20,23	2/5
Female	1960-1969	2	7,11	5	7,11,18,19,21	2/5
Female	1970-1979	2	6,13	5	6,13,17,24,25	2/5
Expected number of identified records						2.35

How many people in the target group to sell a product?

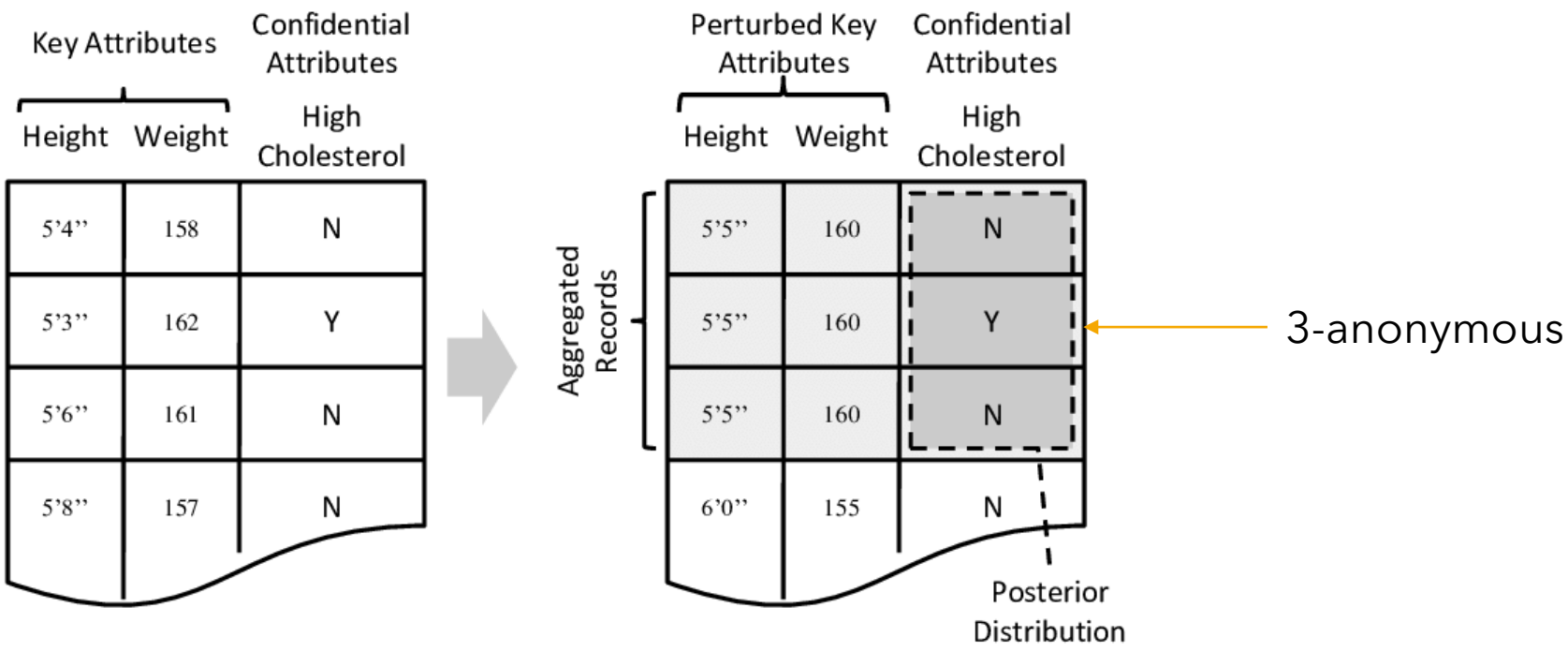
In the *marketer* model the attacker does not target a specific individual but she aims **at re-identifying a high number of individuals**. An attack can therefore only be considered successful if a larger fraction of the records could be re-identified.

Outline

- Attacks on privacy and their consequences
- **Anonymization criteria (k-anonymity, l-diversity, t-closeness)**
- Data curation at the Cancer Registry of Norway (CRN)
- Sharing data for research @ CRN
- Data fuzzification at the Cancer Registry of Norway
- Re-identification Risk Analysis with ARX
- Conclusion

K-anonymity

Sweeney, Latanya. "k-anonymity: A model for protecting privacy." *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems* 10.05 (2002): 557-570.)



A dataset is k-anonymous if each record cannot be distinguished from at least k-1 other records **regarding the quasi-identifiers.**

L-diversity

Machanavajjhala, Ashwin, et al. "l-diversity: Privacy beyond k-anonymity." *ACM Transactions on Knowledge Discovery from Data (TKDD)* 1.1 (2007): 3-es.

***l*-Diversity for sensitive attribute values**

Lname	Diagnosis
Smith	Cancer
Smith	Cancer
Johns	HIV
James	HIV
Peter	Diabetic
Green	Cancer
Peter	HIV
Green	Diabetic
James	Cancer
Johns	HIV

Problem: Inference - anyone named Smith has Cancer in this database.

Lname	Diagnosis
Smith	
Smith	Cancer
Johns	HIV
James	HIV
Peter	Diabetic
Green	Cancer
Peter	HIV
Green	Diabetic
James	Cancer
Johns	

Solution: Diversify sensitive attribute values for every $k>l$ of the same quasi attributes values.

This privacy model can be used to protect data against attribute disclosure by ensuring that each sensitive attribute has at least **ℓ "well represented" values in each equivalence class.**

t-closeness

Li, Ninghui, Tiancheng Li, and Suresh Venkatasubramanian. "t-closeness: Privacy beyond k-anonymity and l-diversity." *2007 IEEE 23rd International Conference on Data Engineering*. IEEE, 2007.

Caucas	787XX	Flu
Caucas	787XX	Shingles
Caucas	787XX	Acne
Caucas	787XX	Flu
Caucas	787XX	Acne
Caucas	787XX	Flu
Asian/AfrAm	78XXX	Flu
Asian/AfrAm	78XXX	Flu
Asian/AfrAm	78XXX	Acne
Asian/AfrAm	78XXX	Shingles
Asian/AfrAm	78XXX	Acne
Asian/AfrAm	78XXX	Flu

Equivalence class is the set of records that have the same values of quasi-identifiers.

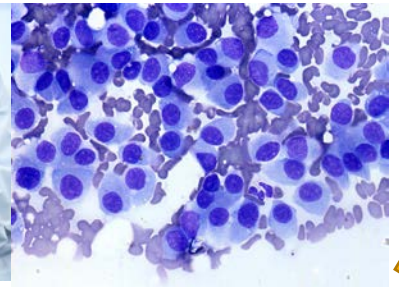
Distribution of sensitive attributes within each quasi-identifier group should be "close" to their distribution in the entire original database

It requires that the distributions of values of a sensitive attribute **within each equivalence class** must have **a distance of not more than *t* to the distribution of the attribute values in the input dataset**. For this purpose, it **bounds** the cumulative absolute difference between the frequency distributions, also measured using the **Earth Mover Distance** or **Wasserstein metric.**

Outline

- Attacks on privacy and their consequences
- Anonymization criteria (k-anonymity, l-diversity, t-closeness)
- **Data curation at the Cancer Registry of Norway (CRN)**
- Sharing data for research @ CRN
- Data fuzzification at the Cancer Registry of Norway
- Re-identification Risk Analysis with ARX
- Conclusion

Data curation at the Cancer Registry of Norway



Pathology reports

Clinical notifications

Cancer Registry of Norway

Statistisk sentralbyrå
Statistics Norway

Death Certificates

Hospital Patient Administration

Cervical Cancer Screening Program

Cervical smear - a test that saves lives

Important information for those turning 25 in 2020

1 Book an appointment for a smear test with your GP

2 Take a smear

3 Reduce your chances of developing cervical cancer

i You will find all the information you need enclosed, or at: https://www.kreftregisteret.no/en/cervix

It is important, even for young, healthy women who have been vaccinated against HPV, to have cervical smears (cervical screening tests).

Having regular smear tests may detect cell changes before they develop into cervical cancer. Cell changes are not the same as cervical cancer and usually do not give symptoms. Cell changes can be easily treated.

The start of a good habit

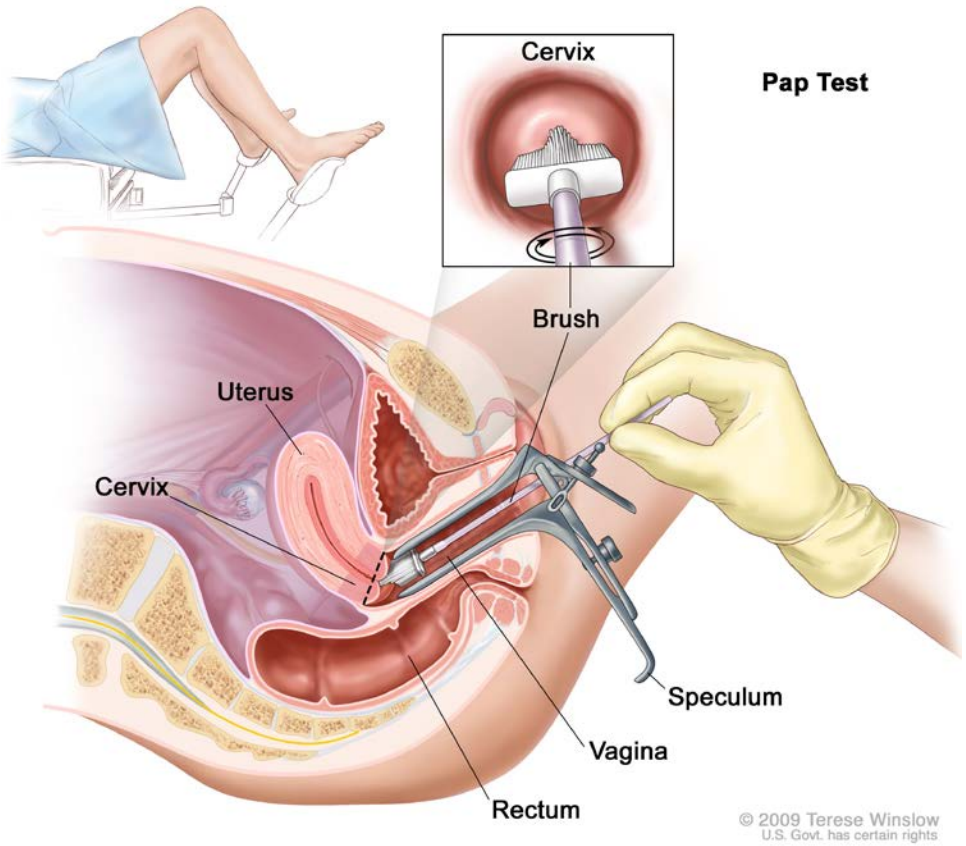
It is recommended that women between the ages of 25 and 69 have cervical screening tests regularly, to prevent cervical cancer. Having regular cervical screening tests significantly reduces the chances of developing cervical cancer. Once you have had a screening test, you will receive a reminder from CervicalScreen Norway when it is time for the next one.

Best regards,

Ameli Tropé
Head of CervicalScreen Norway

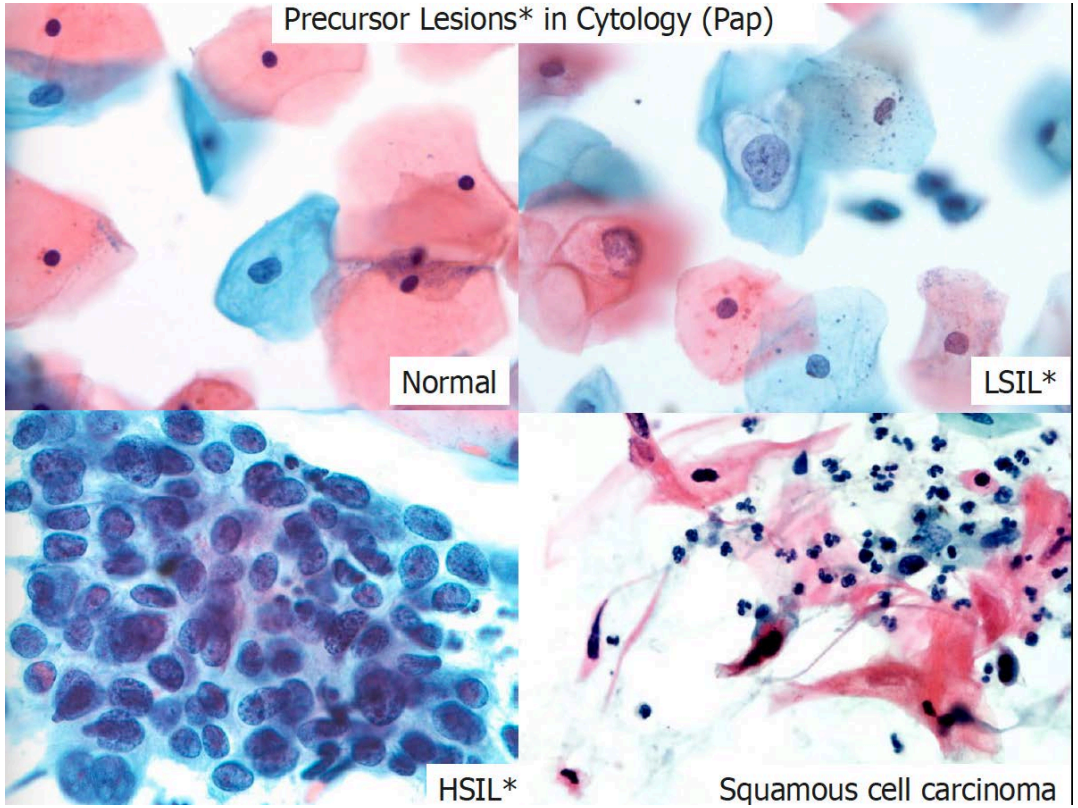
For further information, please see reverse →

Invitation letter

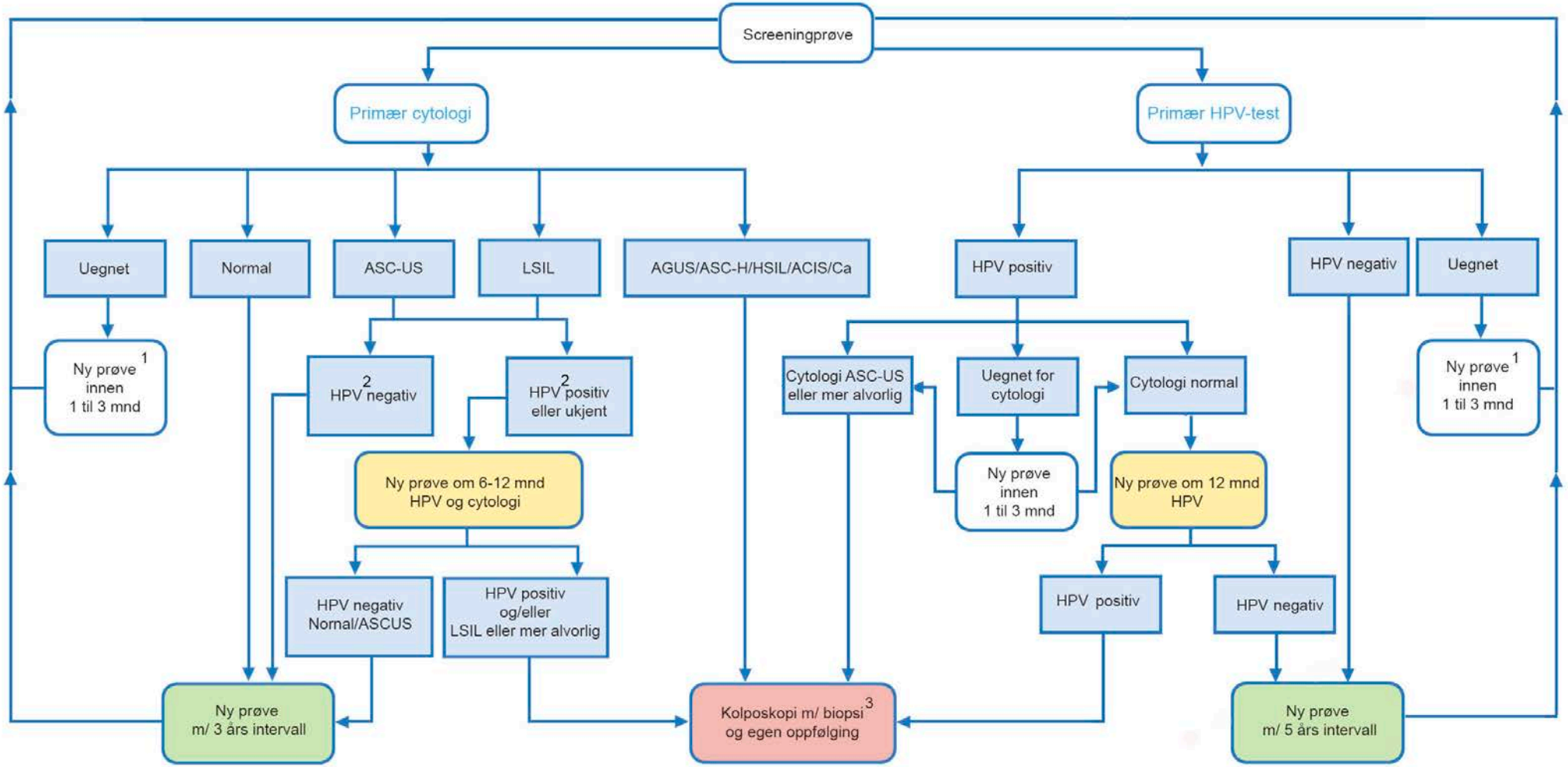


Pap smear

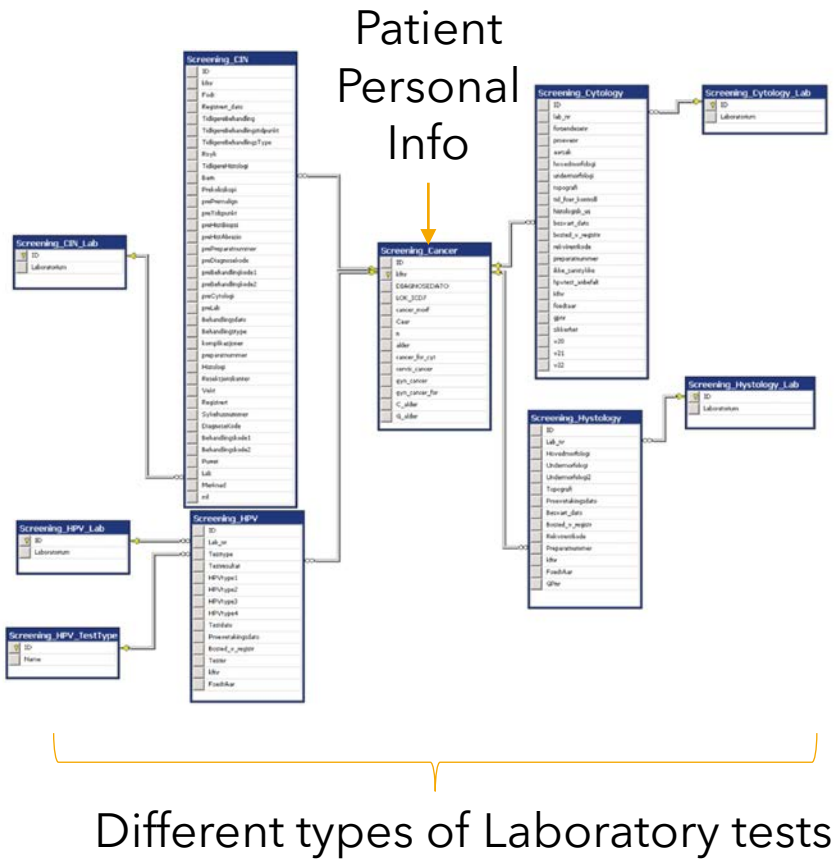
Pap smear


Precursor Lesions* in Cytology (Pap)
Normal
LSIL*
HSIL*
Squamous cell carcinoma

Cervical Cancer Screening Algorithm



Database schema for cervical cancer screening



Key Variables extracted

- Patient ID
- Birth date
- Diagnosis date
- Type (of test)
- Diagnosis1
- Diagnosis2
- Stage
- Lab number
- Region
- Censor date

Sample records in screening dataset

1	ID	birthdate	diagnosisdate	type	diagnosis1	diagnosis2	stage	lab_nr	reg	censordate
2	1	15.08.1960	15.05.1992	cyt	13	76700	999	19	2	
3	1	15.08.1960	15.09.1992	cyt	12	69000	999	19	2	
4	1	15.08.1960	15.11.1992	cyt	13	76700	999	19	2	
5	1	15.08.1960	15.02.1994	cyt	11	100	999	19	2	
6	1	15.08.1960	15.04.1995	cyt	11	100	999	19	2	
7	1	15.08.1960	15.03.1997	cyt	11	100	999	19	2	
8	1	15.08.1960	15.05.1998	cyt	11	100	999	19	2	
9	1	15.08.1960	15.08.2000	cyt	11	100	999	19	2	
10	1	15.08.1960	15.07.2002	cyt	11	100	999	19	2	
11	1	15.08.1960	15.09.2004	cyt	11	100	999	19	2	
12	1	15.08.1960	15.01.2006	hist	20	100	999	19	2	
13	1	15.08.1960	15.01.2006	cyt	11	100	999	19	2	
14	1	15.08.1960	15.02.2007	cyt	11	100	999	19	2	
15	1	15.08.1960	15.05.2007	hist	20	100	999	19	2	
16	1	15.08.1960	15.12.2008	cyt	11	100	999	19	2	
17	1	15.08.1960	15.01.2012	cyt	11	110	999	19	2	
18	1	15.08.1960	15.11.2014	cyt	11	100	999	19	2	
19	2	15.02.1927	15.12.1991	cyt	11	100	999	8	3	15.06.2010
20	2	15.02.1927	15.08.1993	cyt	11	100	999	8	3	15.06.2010
21	3	15.11.1959	15.08.1993	cyt	13	76700	999	11	9	

Types of variables/attributes

- ***Identifying attributes*** are associated with a high risk of re-identification. They will be removed from the dataset. Typical examples are names **or D-number/Personnumer**.

- ***Quasi-identifying attributes*** can in combination be used for re-identification attacks. They will be transformed. Typical examples are gender**, data of diagnosis**, **date of birth** and **postal codes**.

- ***Sensitive attributes*** encode properties with which individuals are not willing to be linked with. As such, they might be of interest to an attacker and, if disclosed, could cause harm to data subjects. They will be kept unmodified but may be subject to further constraints, such as t-closeness or l-diversity. Typical examples are diagnoses such as **HPV+, or Cancer**

- ***Insensitive attributes*** are not associated with privacy risks. They will be kept unmodified.

Sample records in de-identified dataset

Identifying Quasi-identifying Sensitive

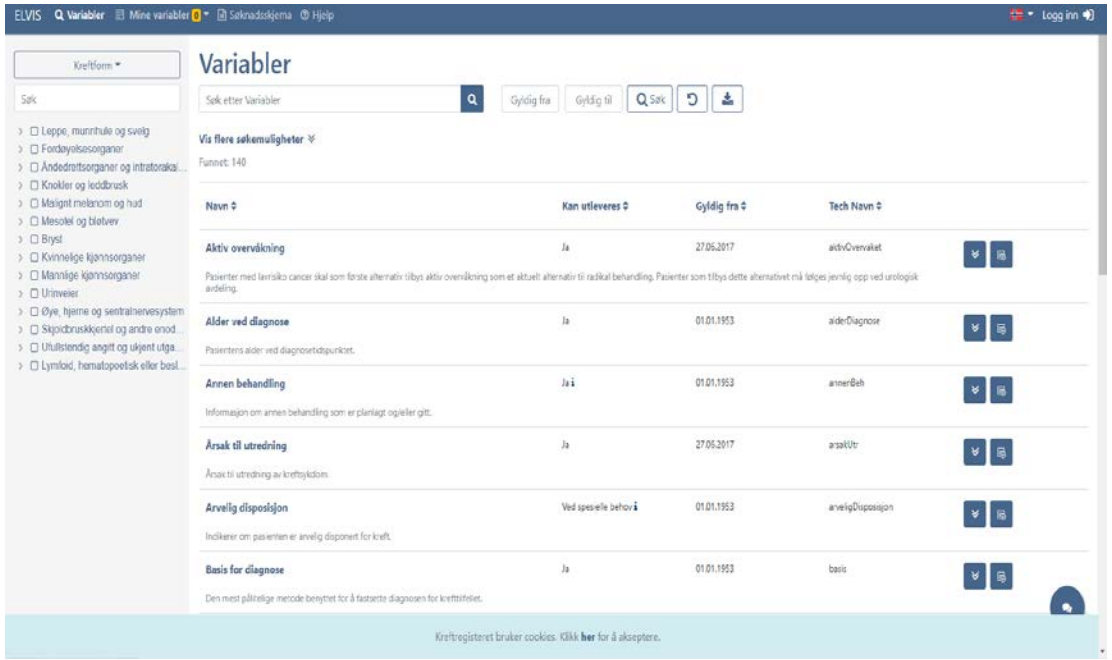
ID	birthdate	diagnosisdate	type	diagnosis1	diagnosis2	stage	lab_nr	reg	censordate
1	15.08.1960	15.05.1992	cyt	13	76700	999	19	2	
1	15.08.1960	15.09.1992	cyt	12	69000	999	19	2	
1	15.08.1960	15.11.1992	cyt	13	76700	999	19	2	
1	15.08.1960	15.02.1994	cyt	11	100	999	19	2	
1	15.08.1960	15.04.1995	cyt	11	100	999	19	2	
1	15.08.1960	15.03.1997	cyt	11	100	999	19	2	
1	15.08.1960	15.05.1998	cyt	11	100	999	19	2	
1	15.08.1960	15.08.2000	cyt	11	100	999	19	2	
1	15.08.1960	15.07.2002	cyt	11	100	999	19	2	
1	15.08.1960	15.09.2004	cyt	11	100	999	19	2	
1	15.08.1960	15.01.2006	hist	20	100	999	19	2	
1	15.08.1960	15.01.2006	cyt	11	100	999	19	2	
1	15.08.1960	15.02.2007	cyt	11	100	999	19	2	
1	15.08.1960	15.05.2007	hist	20	100	999	19	2	
1	15.08.1960	15.12.2008	cyt	11	100	999	19	2	
1	15.08.1960	15.01.2012	cyt	11	110	999	19	2	
1	15.08.1960	15.11.2014	cyt	11	100	999	19	2	
2	15.02.1927	15.12.1991	cyt	11	100	999	8	3	15.06.2010
2	15.02.1927	15.08.1993	cyt	11	100	999	8	3	15.06.2010
3	15.11.1959	15.08.1993	cyt	13	76700	999	11	9	

Outline

- Attacks on privacy and their consequences
- Anonymization criteria (k-anonymity, l-diversity, t-closeness)
- Data curation at the Cancer Registry of Norway (CRN)
- **Sharing data for research @ CRN**
- Data fuzzification at the Cancer Registry of Norway
- Re-identification Risk Analysis with ARX
- Conclusion

Sharing data for research

ELVIS metadata bank



https://metadata.kreftregisteret.no/variables/search?selection=cancer_sites

Types of data extraction

- Anonymous aggregated data
- Personally identifiable individual data
- De-identified individual data

Requirements

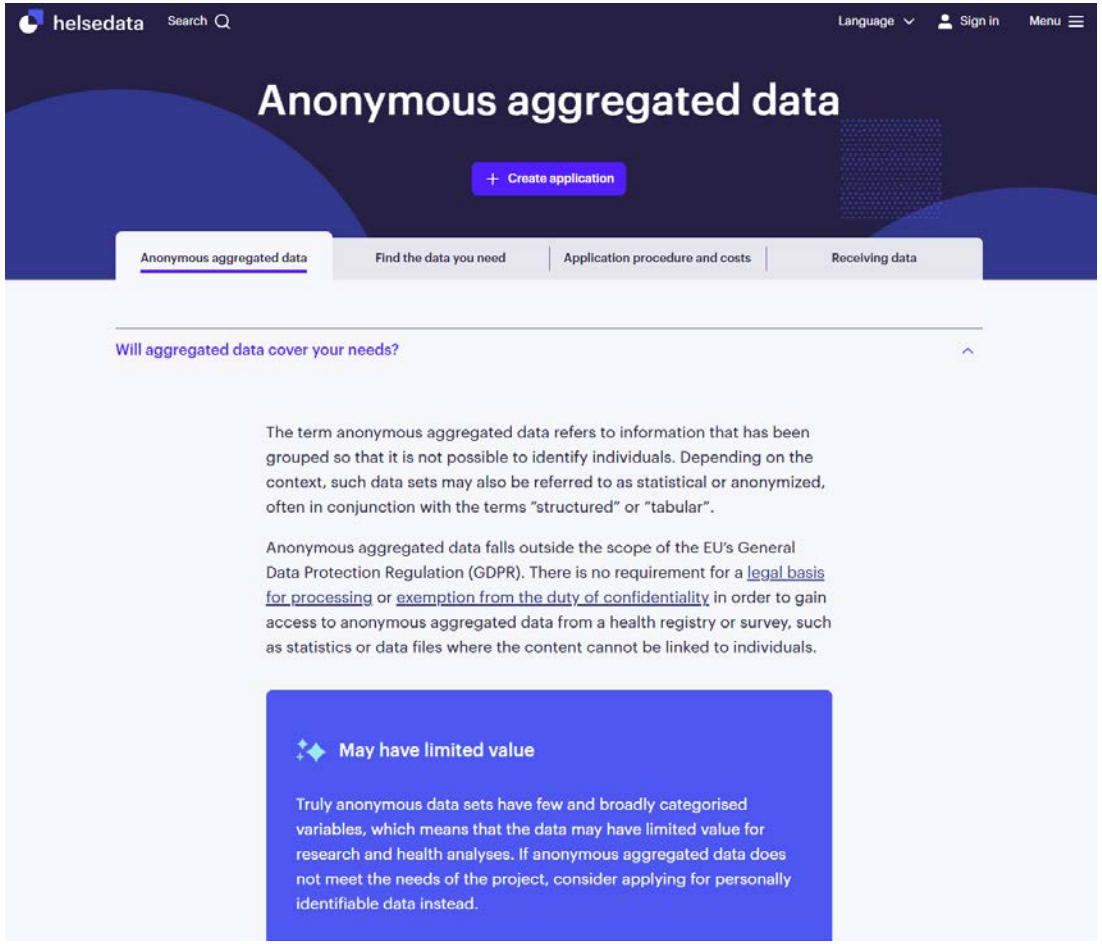
- Documentation of legal basis for processing of personal data
- Consent or exemption from the duty of confidentiality

Delivery time

- De-identified Cancer Registry data: 30 days
- Cancer Registry data linked to other sources: 60 days

GDPR
The General Data Protection Regulation

Sharing data for research – aggregated data



Sharing data for research – personally identifiable data

Outline

- Attacks on privacy and their consequences
- Anonymization criteria (k-anonymity, l-diversity, t-closeness)
- Data curation at the Cancer Registry of Norway (CRN)
- Sharing data for research @ CRN
- **Data fuzzification at the Cancer Registry of Norway**
- Re-identification Risk Analysis with ARX
- Conclusion

Fuzzy algorithm for the cervical cancer screening program

Ursin, G., Sen, S., Mottu, J. M., & Nygård, M. (2017). Protecting privacy in large datasets—first we assess the risk; then we fuzzy the data. *Cancer Epidemiology and Prevention Biomarkers*, *26*(8), 1219-1224.

Step 1*—setting all dates to the* ***15th of the month.***

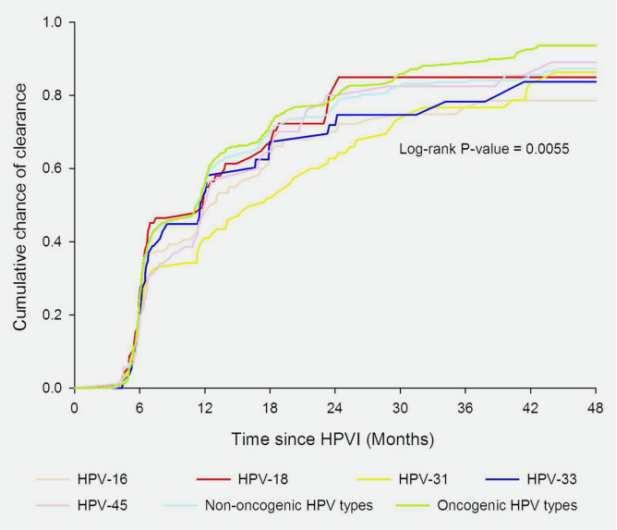
- *Affected birthdate, diagnosis date, and censor date*
- *Did not affect the age of a person*
- *Did not affect the diagnosis as HPV infection clearance takes 6 months*

Step 2*—adding noise or "****fuzziness****" to dates.*

- Based on the type of study and the scientific objectives
- All dates were perturbed by a random integer between **+4 and -4** (not 0) as disease clearance probability does not change under 6 months

Step 3*—All original IDs were assigned a random ID*

- De-identified IDs



Output of fuzzification

1	ID	birthdate	diagnosisdate	type	diagnosis1	diagnosis2	stage	lab_nr	reg	censordate
2	1	15.08.1960	15.05.1992	cyt	13	76700	999	19	2	
3	1	15.08.1960	15.09.1992	cyt	12	69000	999	19	2	
4	1	15.08.1960	15.11.1992	cyt	13	76700	999	19	2	
5	1	15.08.1960	15.02.1994	cyt	11	100	999	19	2	
6	1	15.08.1960	15.04.1995	cyt	11	100	999	19	2	
7	1	15.08.1960	15.03.1997	cyt	11	100	999	19	2	
8	1	15.08.1960	15.05.1998	cyt	11	100	999	19	2	
9	1	15.08.1960	15.08.2000	cyt	11	100	999	19	2	
10	1	15.08.1960	15.07.2002	cyt	11	100	999	19	2	
11	1	15.08.1960	15.09.2004	cyt	11	100	999	19	2	
12	1	15.08.1960	15.01.2006	hist	20	100	999	19	2	
13	1	15.08.1960	15.01.2006	cyt	11	100	999	19	2	
14	1	15.08.1960	15.02.2007	cyt	11	100	999	19	2	
15	1	15.08.1960	15.05.2007	hist	20	100	999	19	2	
16	1	15.08.1960	15.12.2008	cyt	11	100	999	19	2	
17	1	15.08.1960	15.01.2012	cyt	11	110	999	19	2	
18	1	15.08.1960	15.11.2014	cyt	11	100	999	19	2	
19	2	15.02.1927	15.12.1991	cyt	11	100	999	8	3	15.06.2010
20	2	15.02.1927	15.08.1993	cyt	11	100	999	8	3	15.06.2010
21	3	15.11.1959	15.08.1993	cyt	13	76700	999	11	9	

Outline

- Attacks on privacy and their consequences
- Anonymization criteria (k-anonymity, l-diversity, t-closeness)
- Data curation at the Cancer Registry of Norway (CRN)
- Sharing data for research @ CRN
- Data fuzzification at the Cancer Registry of Norway
- **Re-identification Risk Analysis with ARX**
- Questions

Re-identification Risk Analysis in the Norwegian Cervical Screening Program

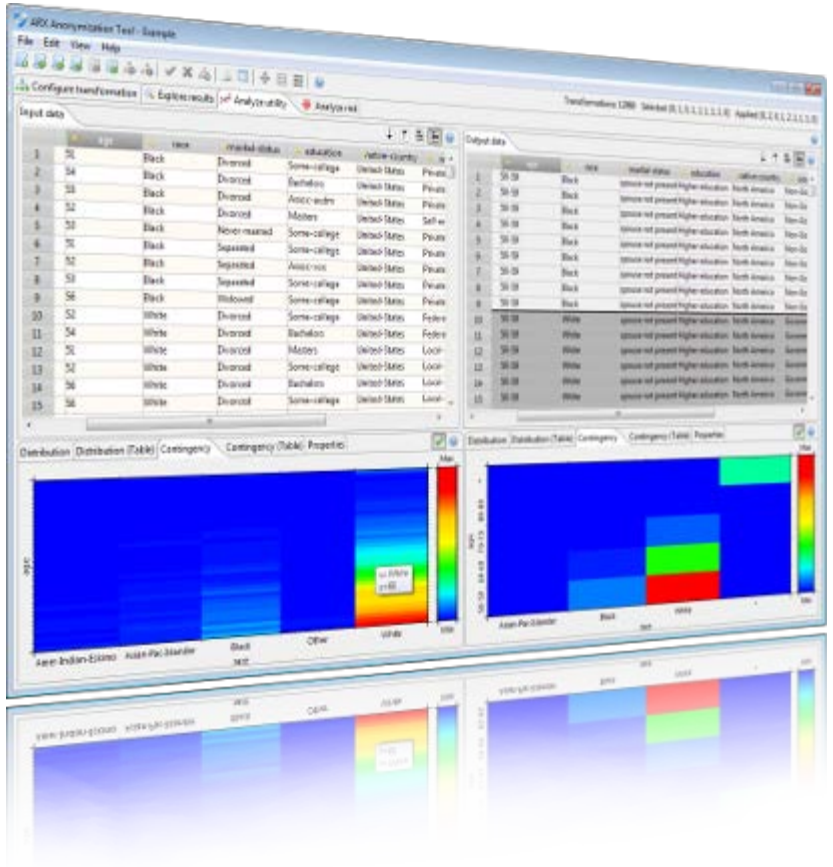
- The dataset contained 5,693,582 records of screening related examinations taken by 911,510 distinct women. The birth dates of the women ranged from March 1905 to February 1996.

The risk of reidentification was assessed for the following datasets:

- D1. **Realistic dataset** of women attending cervical cancer screening in Norway.
- D2. **k-Anonymization** of the dataset D1 by changing all dates in the dataset to 15th of the month.
- D3. **Fuzzifying the month** in D2 by adding a random factor between −4 and +4 months to each month as described above.

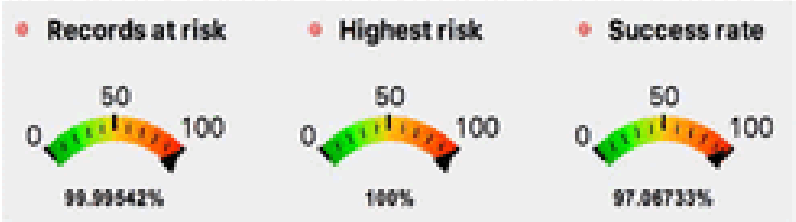
ARX - Data Anonymization and Risk Analysis Tool

Prasser, Fabian, et al. "Arx-a comprehensive tool for anonymizing biomedical data." *AMIA Annual Symposium Proceedings*. Vol. 2014. American Medical Informatics Association, 2014.



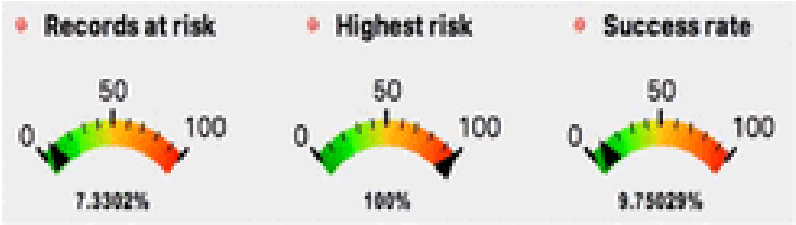
Re-identification Risk Analysis in the Norwegian Cervical Screening Program

A Realistic dataset



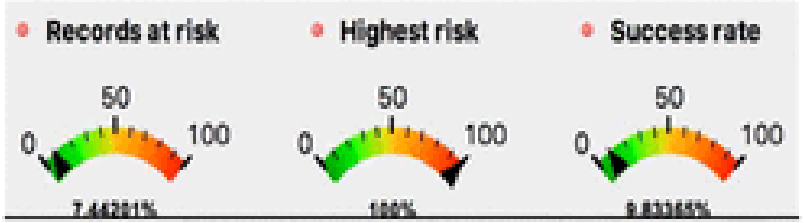
Average prosecutor risk:	97.06733%
Lowest prosecutor risk:	20%
Highest prosecutor risk:	100%
Records affected by lowest risk:	0.00009%
Records affected by highest risk:	94.198%

B All days to 15th of month



Average prosecutor risk:	9.7%
Lowest prosecutor risk:	1.176%
Highest prosecutor risk:	100%
Records affected by lowest risk:	0.00149%
Records affected by highest risk:	6.0%

C Fuzzy factor ± 4 months



Average prosecutor risk:	9.8%
Lowest prosecutor risk:	1.265%
Highest prosecutor risk:	100%
Records affected by lowest risk:	0.00416%
Records affected by highest risk:	6.1%

Outline

- Data curation at the Cancer Registry of Norway (CRN)
- Sharing data for research @ CRN
- Attacks on privacy and their consequences
- Anonymization criteria (k-anonymity, l-diversity, t-closeness)
- Data fuzzification at the Cancer Registry of Norway
- Re-identification Risk Analysis with ARX
- **Conclusion**

Conclusion

- We learn about the types of attacks to privacy : prosector, journalist, and marketer attacks
- The basic approaches to anonymization need to satisfy: k-anonymity, l-diversty, and t-closeness
- The cervical cancer screening program is a good example of where anonymization and re-identification risk analysis has been useful.
- The Cancer Registry today uses various forms of fuzzification in all its databases (beyond cervical cancer) to share aggregate data after verifying its validity using ARX

How did the concept of privacy origniate?

The first man who, having enclosed a piece of ground, bethought himself of saying **"This is mine,"** and found people simple enough to believe him, was the real founder of civil society.

–Jean-jacques Rousseau, Discourse on the origin of inequality

Hirshleifer, Jack. "Privacy: Its origin, function, and future." *The Journal of Legal Studies* 9.4 (1980): 649-664.

Privacy and its origins…

The etymology of the word privacy stems from **privus**, the original archaic meaning being **single**.

The implied context is *not a solitary human being* but rather the individual facing the potential claims of other persons

How is privacy different from seclusion?

Seclusion is *withdrawal from society*

Privacy is a way to organize civil society

How is privacy different from secrecy?

Privacy

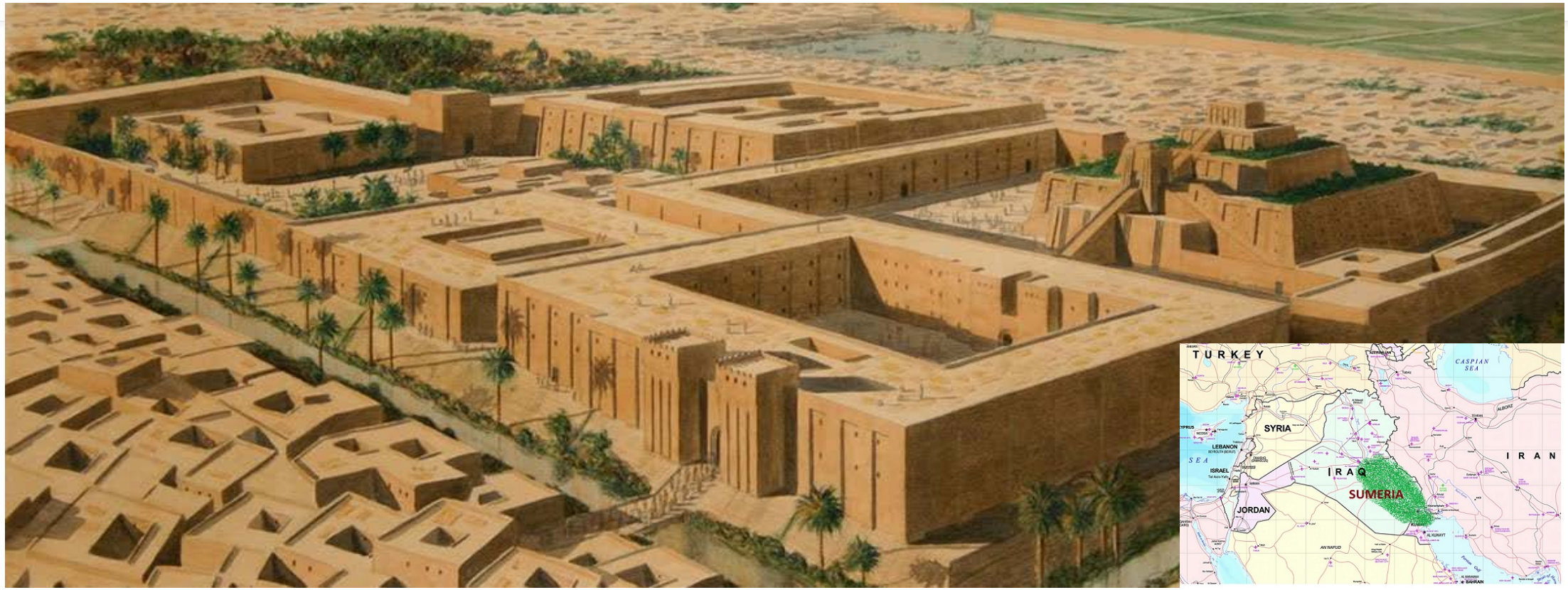
an aspect of

Secrecy

A sales call is an "Invasion of privacy" but no secret information is "usually" revealed

Secrecy is the ability to control dissemination and use of information (or possessed) by oneself

Privacy, an age old concept, is about autonomy in society



Sumerian Civilization, 3000 BC – a civiliation with a social structure with a supporting social ethic

What is invasion of privacy and its consequences?

Vasco de gama
1498

Merchants of Calicut, India held hostage

Dis-possession of private property

The Great Explorers of the Colonial Era

What are the consequences of invasion of privacy of our data in 100 years?