Understanding privacy risk in electronic health records: a case study on Discharge Summaries

Arlene Casey (1,2), Amy Tilbrook (1), Stuart Dunbar (1), Atul Anand (1), Chloe Brook (1), Pamela Linksted (1), Katherine O'Sullivan (3), Charlie Mayor (4), Jacqueline Caldwell (5), Elizabeth Ford (6), Kathy Harrison (1), Nicholas Mills (1)

1 DataLoch, Usher Institute, University of Edinburgh; 2 ACRC, Usher Institute, University of Edinburgh; 3 Grampian Data Safe Haven (DasH), University of Aberdeen; 4 West of Scotland Safe Haven, NHS Glasgow; 5 eDRIS, Public Health Scotland; 6 Brighton and Sussex Medical School, University of Sussex

DataLoch

ACRC Advanced Care Research Centre

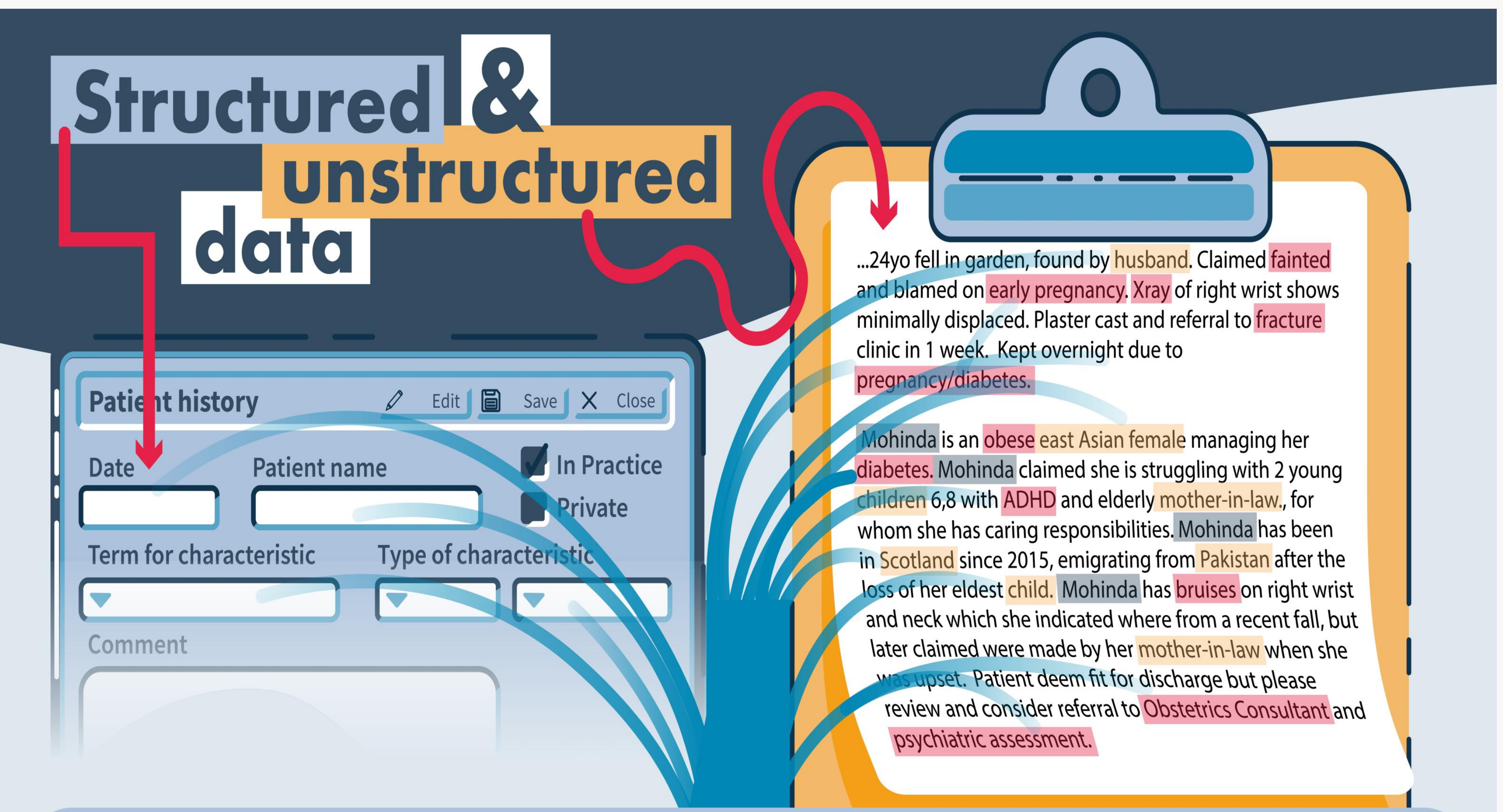
DARE UK Phase 1 Driver Project

Privacy risk assessment is key when releasing health data to ensure released data does not contain identifiable patient information

CHALLENGE: Risk assessments for data release are often manual, time consuming and can prohibit data release, ultimately limiting new health and social care innovation. Most health research is conducted using structured data. Although unstructured data makes up 70-80% of health data, due to both its unstructured format as well as privacy risks for patients its use is limited. Risk assessment frameworks are not well defined for unstructured data and more work is needed to understand privacy risks, the risk of revealing patient identity, from this type of data to enable its release.

AIM: We apply Natural Language Processing (NLP) techniques to explore a year's worth of hospital Discharge Summaries looking at privacy risks and how these accumulate across reports, building a privacy risk map.

Using the privacy risk map, we are building a dashboard that helps gain a better insight into privacy risks and enables assessment ahead of data release.

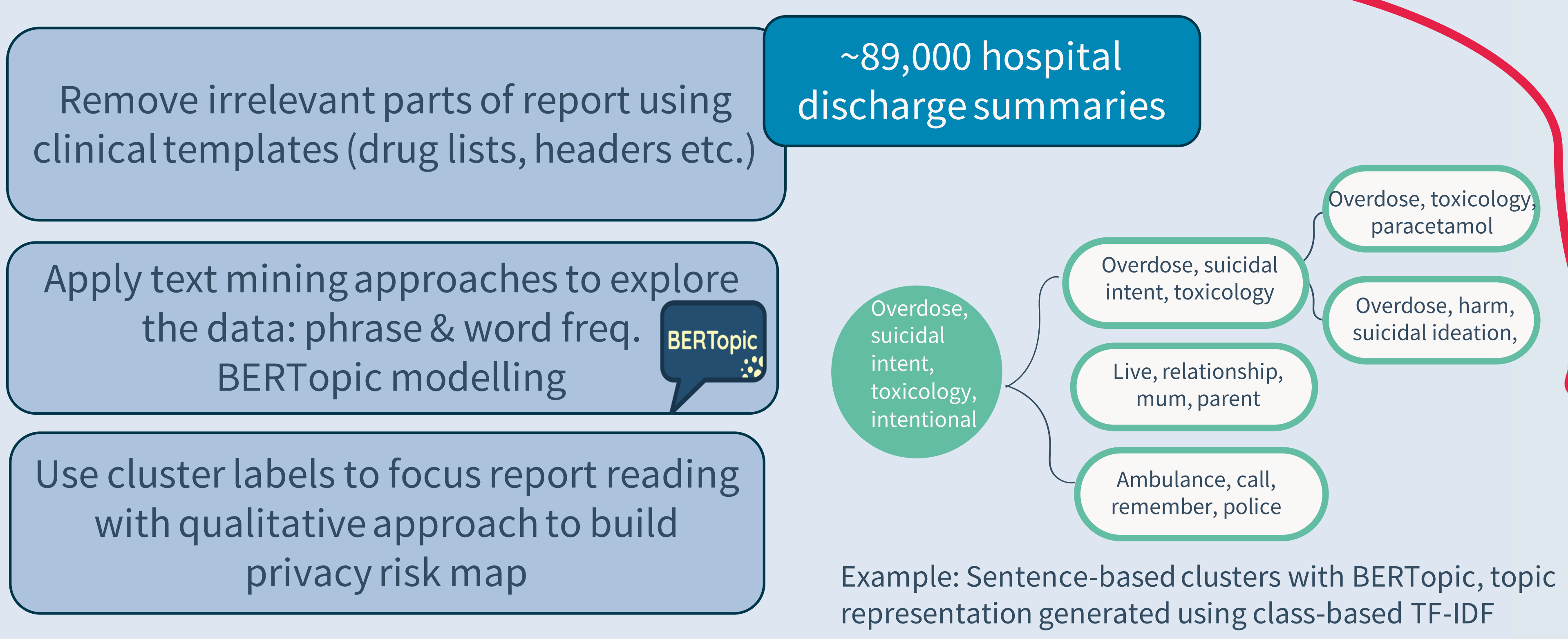


...24yo fell in garden, found by husband. Claimed fainted and blamed on early pregnancy. Xray of right wrist shows minimally displaced. Plaster cast and referral to fracture clinic in 1 week. Kept overnight due to pregnancy/diabetes.

Mohinda is an obese east Asian female managing her diabetes. Mohinda claimed she is struggling with 2 young children 6,8 with ADHD and elderly mother-in-law., for whom she has caring responsibilities. Mohinda has been in Scotland since 2015, emigrating from Pakistan after the loss of her eldest child. Mohinda has bruises on right wrist and neck which she indicated where from a recent fall, but later claimed were made by her mother-in-law when she was upset. Patient deem fit for discharge but please review and consider referral to Obstetrics Consultant and psychiatric assessment.

70-80% of health data is unstructured but its use in research is limited due to privacy risk concerns. We define privacy risks as direct or indirect.

Finding Indirect Privacy Risks



~89,000 hospital discharge summaries

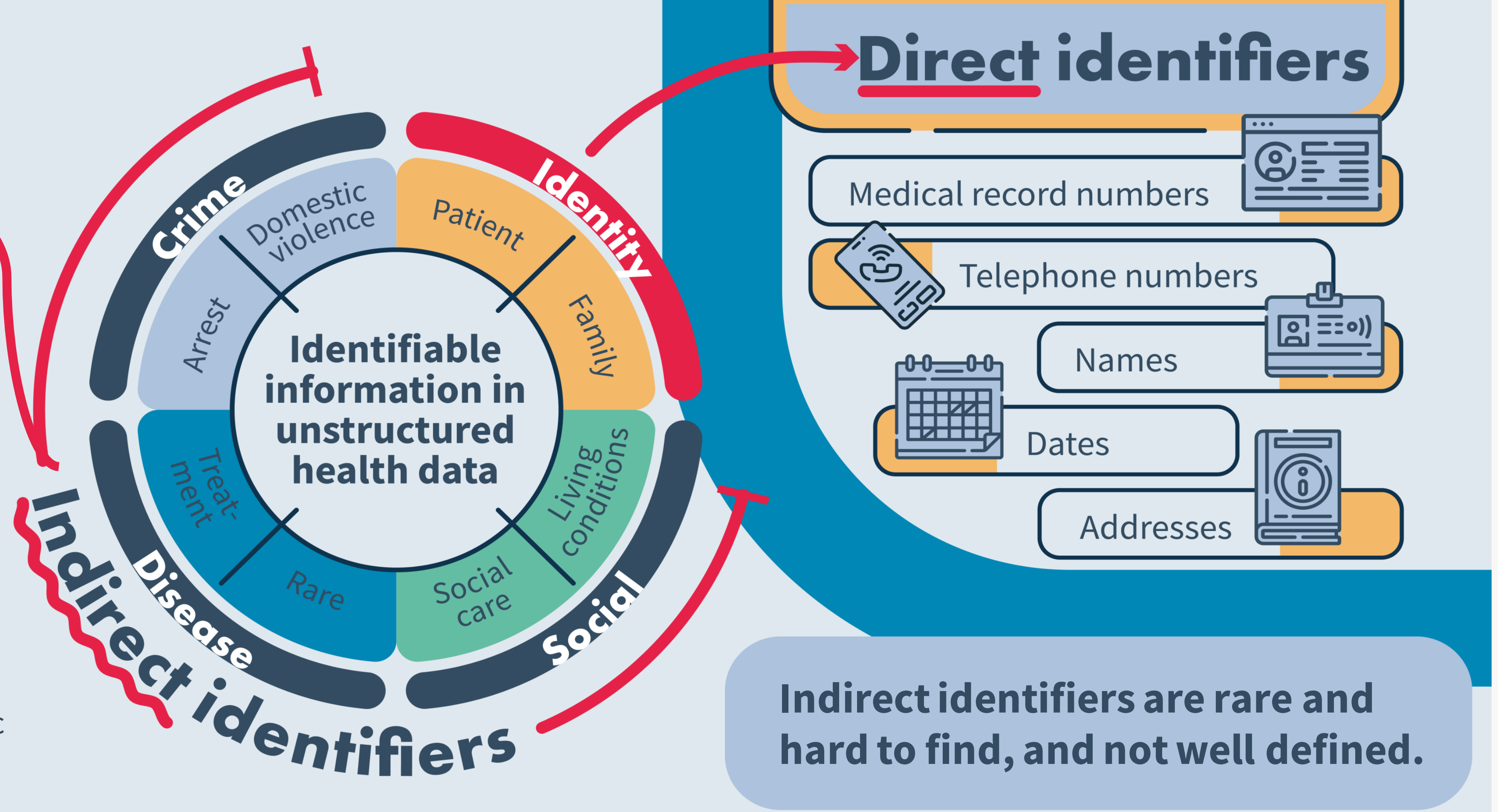
Remove irrelevant parts of report using clinical templates (drug lists, headers etc.)

Apply text mining approaches to explore the data: phrase & word freq. BERTopic modelling

Use cluster labels to focus report reading with qualitative approach to build privacy risk map

Overdose, suicidal intent, toxicology, intentional
Overdose, suicidal intent, toxicology
Live, relationship, mum, parent
Ambulance, call, remember, police
Overdose, toxicology, paracetamol
Overdose, harm, suicidal ideation,

Example: Sentence-based clusters with BERTopic, topic representation generated using class-based TF-IDF



Medical record numbers

Telephone numbers

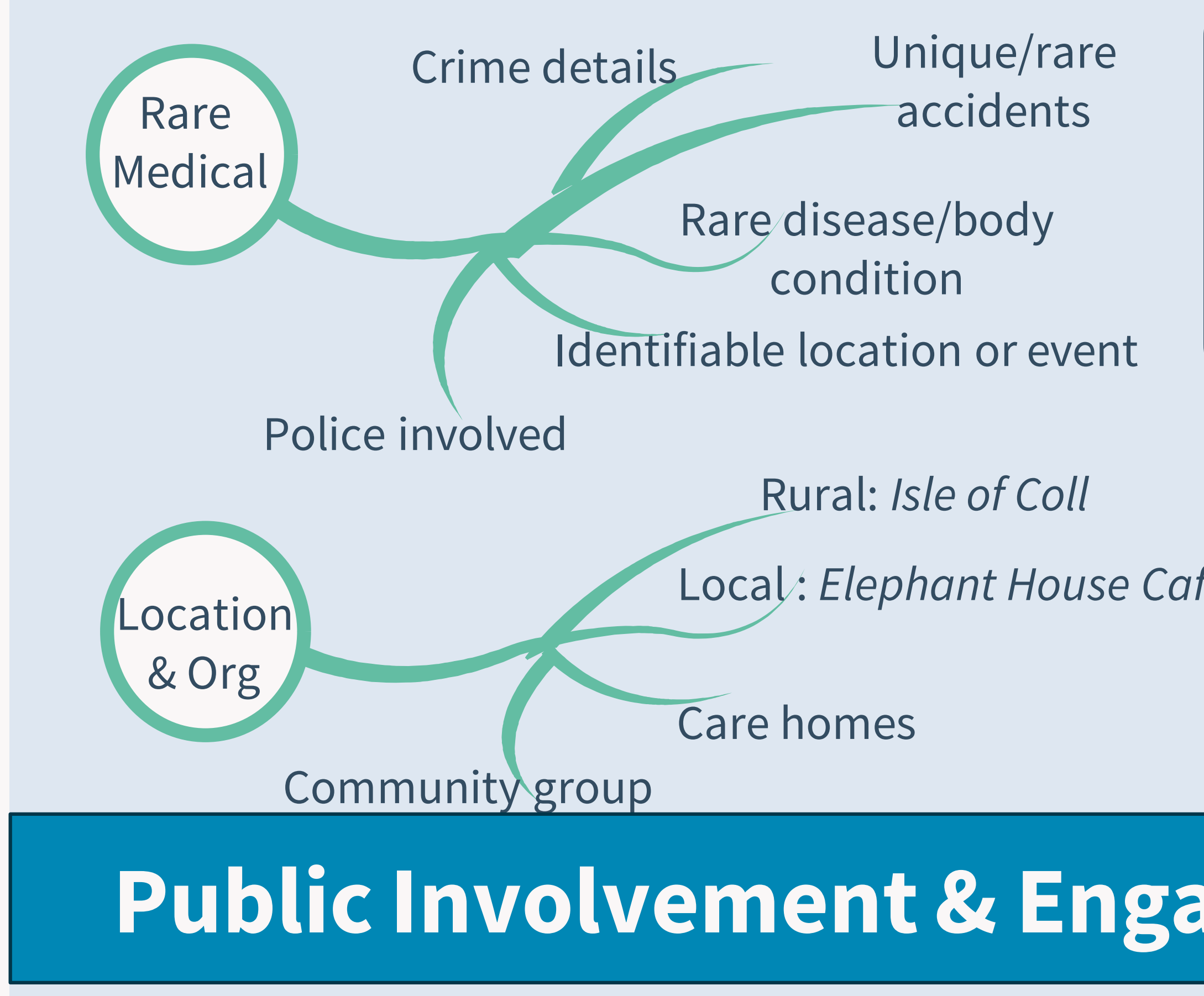
Names

Dates

Addresses

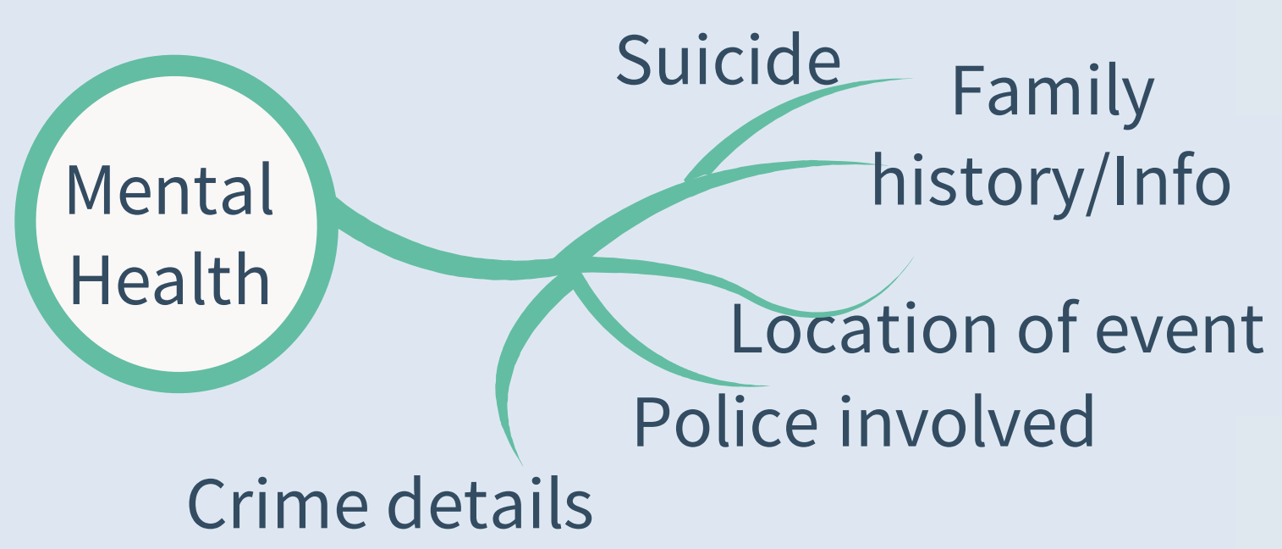
Indirect identifiers are rare and hard to find, and not well defined.

Most prevalent indirect privacy risks. Rare and unique events give most concerns particularly in younger ages as does cumulative information of frequent patients.

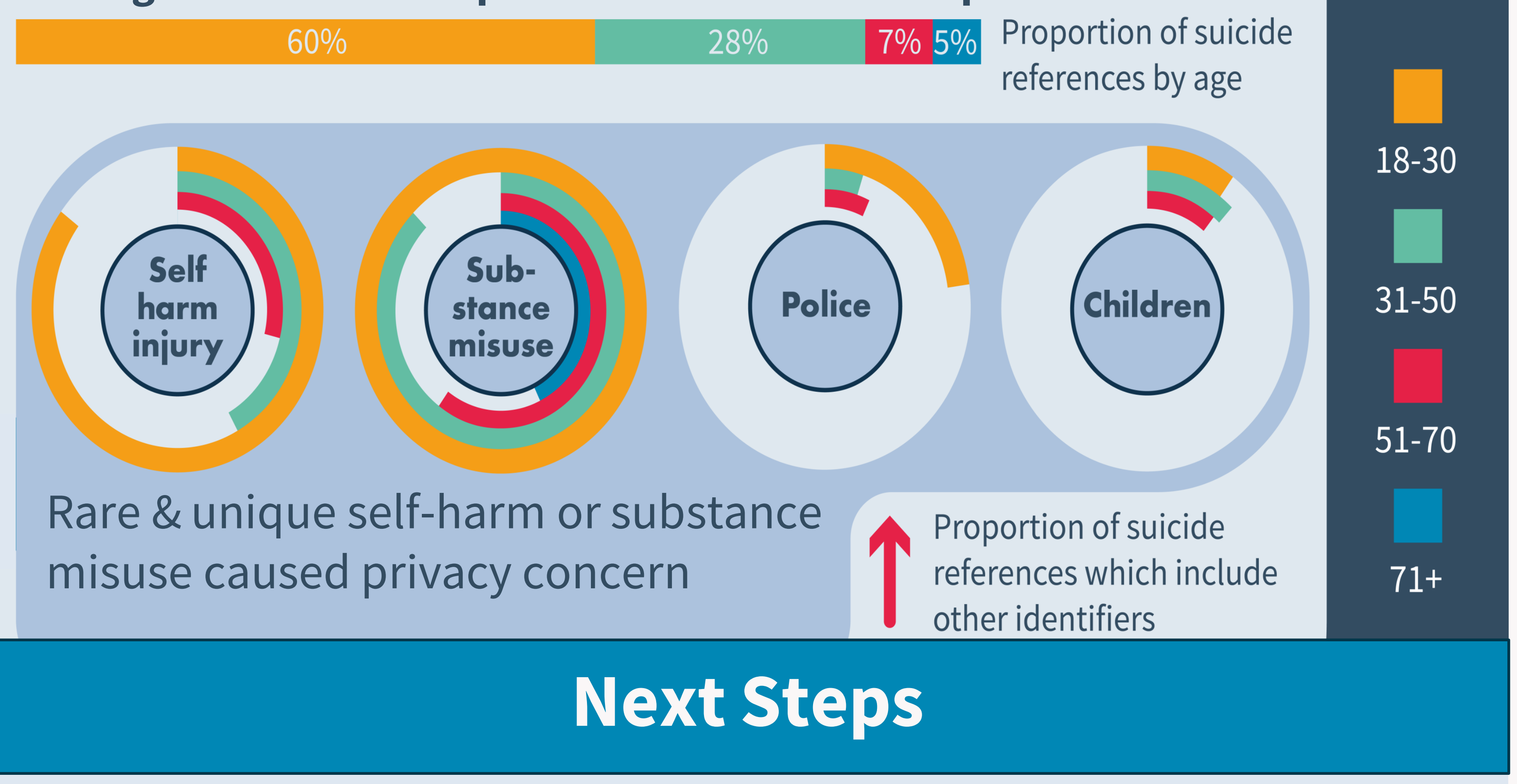


Examples of Rare accidents

'fell concrete pit at work'
'wood-chipper accident'
'died in kiteboard accident'
'jump from window in burning apartment'
'loss of hand/foot or a specific digit'



Discharge summaries for patients who have attempted suicide



Rare & unique self-harm or substance misuse caused privacy concern

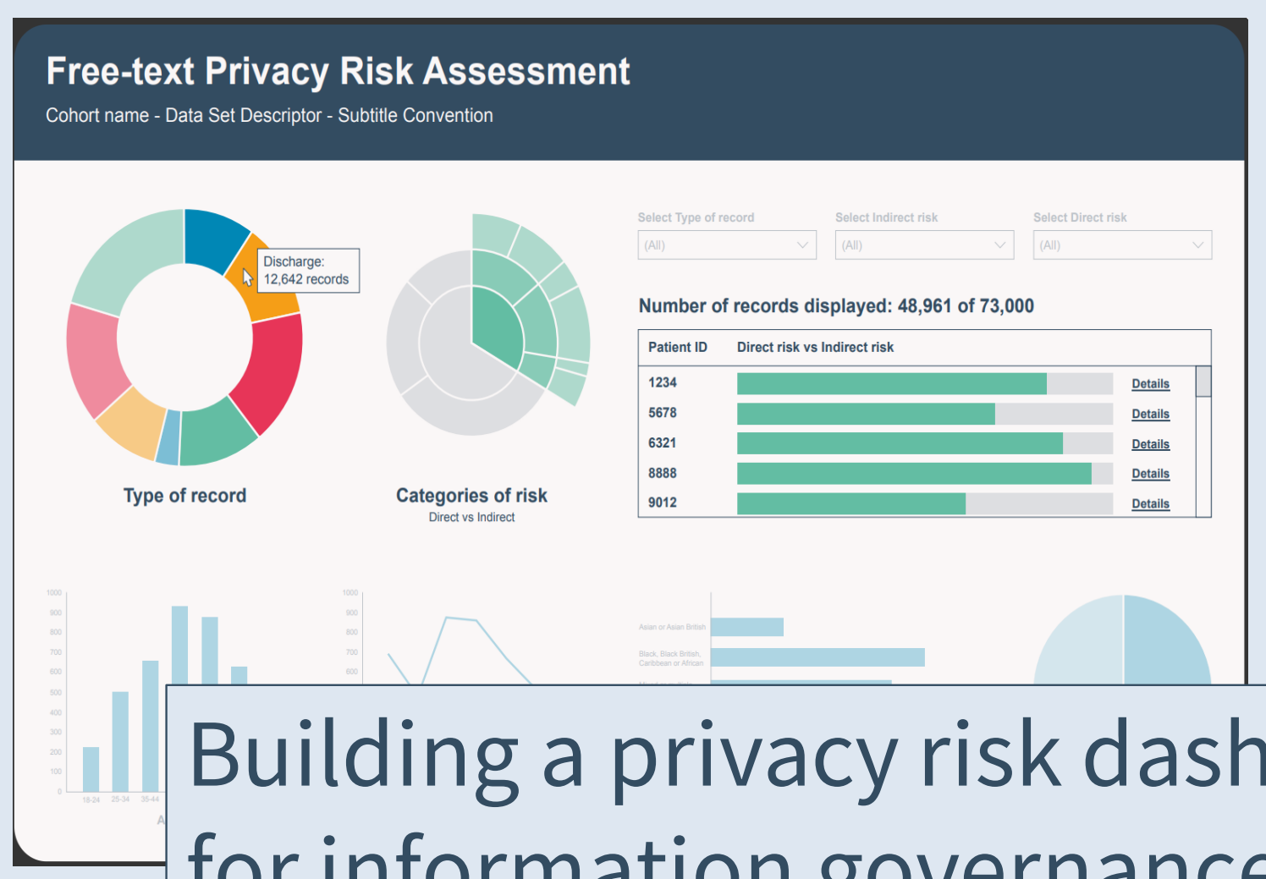
Public Involvement & Engagement Work

PPIE: explored public opinion of identifiers and how semi-automation can be used to address risks.

The public are broadly supportive of the use of semi-automation to address privacy risks but believe:

- Some data should be coded to ensure valuable details for research are not lost while preserving confidentiality. The process should involve discussion with specialists.
- Audit trails and human assessment are necessary to ensure processes run correctly

Next Steps



Free-text Privacy Risk Assessment

Cohort name - Data Set Descriptor - Subtitle Convention

Number of records displayed: 48,961 of 73,000

Patient ID	Direct risk vs Indirect risk	
1234		Details
5678		Details
6321		Details
8888		Details
9012		Details

Building a privacy risk dashboard for information governance to assess a cohort's privacy risk, based on known direct and our map of indirect identifiers

Evaluating and developing tools for semi-automation of finding identifiers in collaboration with the Scottish Safe Haven Network and partners

For more information: Arlene.Casey@ed.ac.uk

Image panels 2,4,6 on this poster were designed by Dr Lana Woolford at Cloud Chamber Studios (lana@cloud-chamber-studios.com)

DataLoch · DaSH · Public Health Scotland · University of Aberdeen · NHS Greater Glasgow and Clyde West of Scotland Data Safe Haven · Health Informatics Centre · Research Data Scotland · University of Sussex · The University of Edinburgh · Usher Institute

The University of Edinburgh · Data-Driven Innovation · UK Government · Scottish Government · City Region Deal · NHS Lothian · The University of Edinburgh · Usher Institute