DARE UK

Semi-Automated Risk Assessment (SARA)

17 Jan 2024

Stuart Dunbar, dataloch@ed.ac.uk

Engagement Manager, DataLoch

University of Edinburgh

Semi-automated risk assessment at the data access stage

DARE UK

- Focus on two parts of the process:
  - Enhancing record-keeping on how data are brought together for research purposes (data provenance)
  - Identifying privacy risks within excerpts of unstructured free-text (e.g. hospital discharge summaries)

1. Understand public and stakeholder perceptions of appropriate risk levels around data provenance and privacy in clinical free-text

2. Framework and prototype for partial automation of risk assessment of clinical free-text

3. Framework for semi-automation of data provenance creation and auditing to improve risk assessment

Framework for semi-automation of data provenance creation and auditing to improve risk assessment

DARE UK

The Context

Data can be sourced from a number of systems that support direct patient care

- Cardiac Surgeries and Procedures
- Critical Care treatment
- Maternity services
- GP events: codes for symptoms, observations, diagnoses, etc
- GP registrations
- Emergency Department attendances
- Hospital Admissions and Discharges
- Emergency Department attendances
- Laboratory results: e.g. blood tests
- Mental Health services

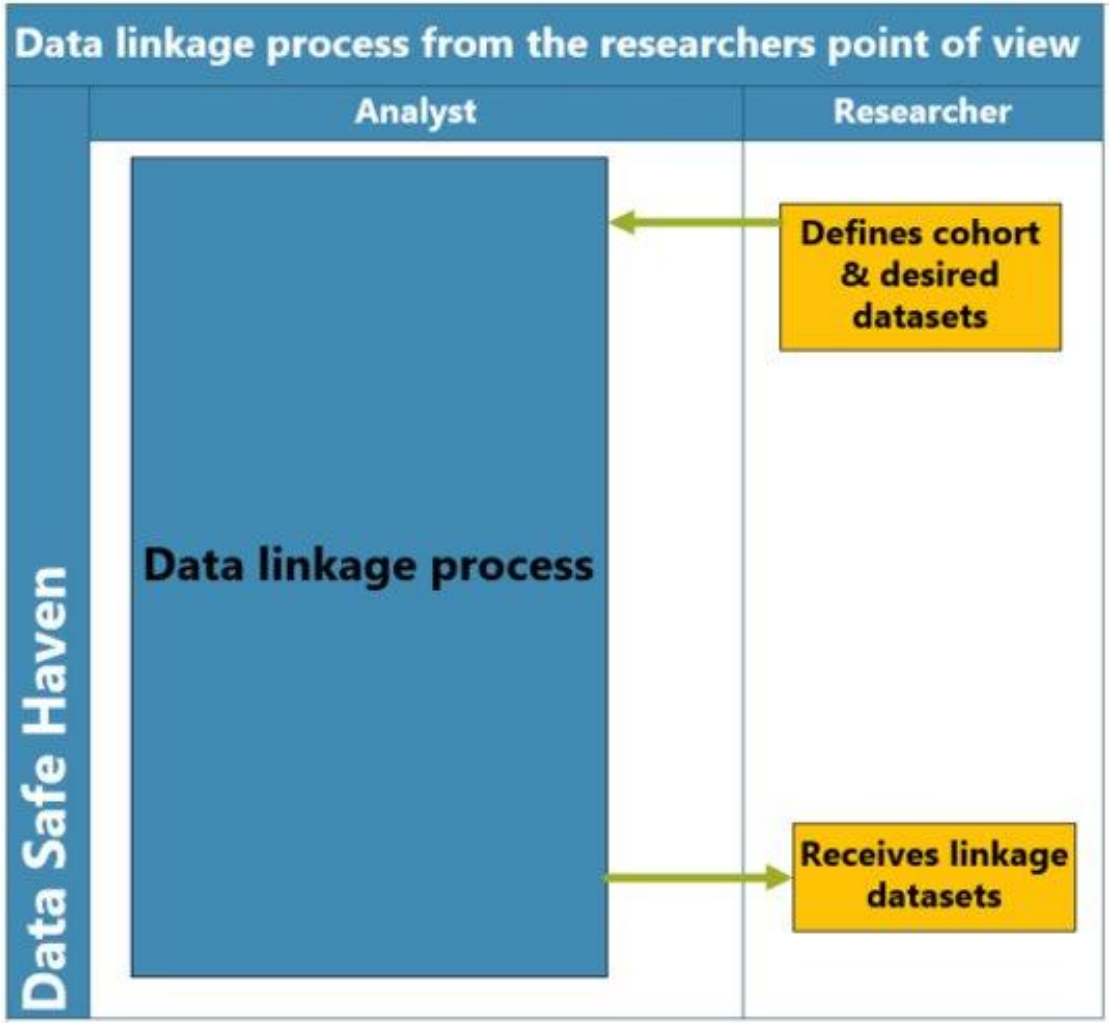
Framework for semi-automation of data provenance creation and auditing to improve risk assessment

DARE UK

The Context

Data linkage within trusted research environments is perceived by researchers to be 'a black box'. Better understanding has potential to enhance research.

Provenance (information about entities, activities, and people involved in producing data) can be used to form assessments about its quality, reliability or trustworthiness and enhance knowledge/usefulness of research output.



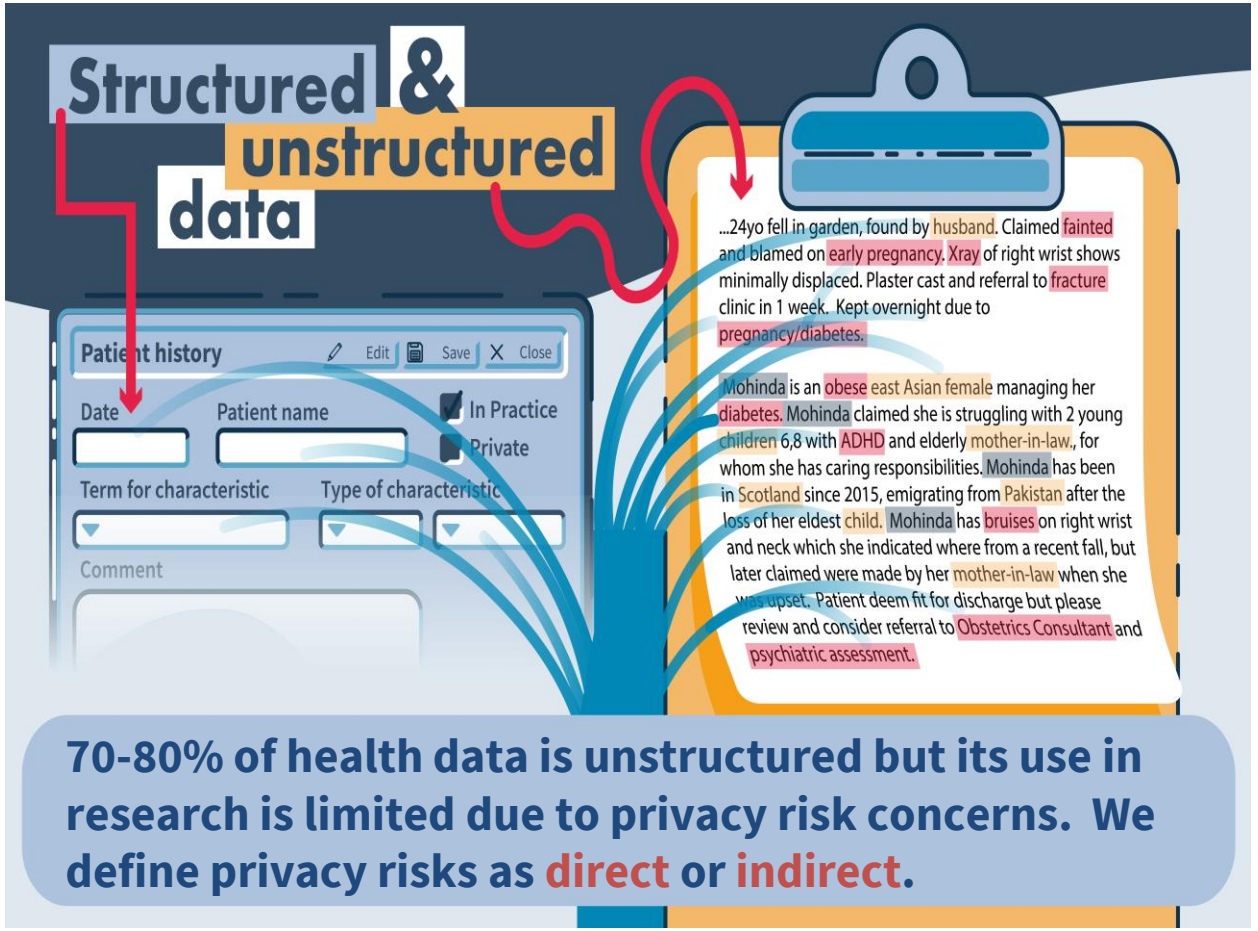
Framework and prototype for partial automation of risk assessment of clinical free-text

DARE UK

The Context

The processing of structured data for research is well understood.

Structured data includes dates, names, codes for different conditions, etc

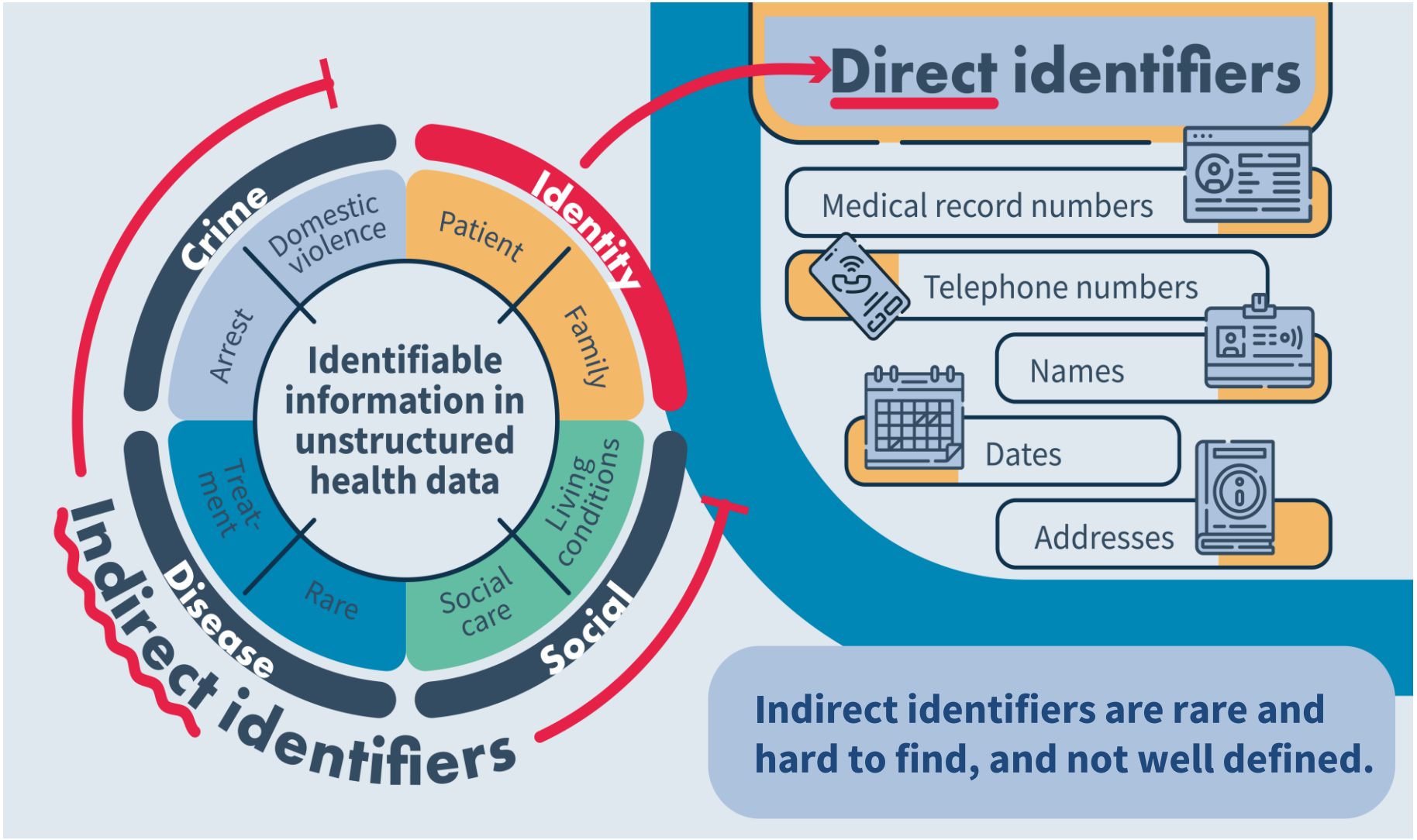


Unstructured data – like GP notes and hospital discharge summaries - is more challenging to make available for research.

The privacy risks need closer scrutiny.

Framework and prototype for partial automation of risk assessment of clinical free-text

DARE UK



DARE UK

Approach to Public Involvement and Engagement

Understand public and stakeholder perceptions of appropriate risk levels around data provenance and privacy in clinical free-text

DARE UK

Approach to Public Involvement and Engagement (PIE)

Deliberative workshops

- Introductory online workshop for all participants
- In-person workshops – Aberdeen / Edinburgh
- 39 participants in all, targeted recruitment

Online survey

- Questions informed by workshop findings
- Used KnowledgePanel UK – existing Ipsos panel
- 1,030 adult respondents

DARE UK

Approach to Data Provenance

DARE UK

Framework for semi-automation of data provenance creation and auditing to improve risk assessment

Principle Activities

- Interviews with analysts, information governance specialists, research coordinators to understand what information is beneficial for each role
- Development of background processes
- Production of dashboard to provide information in meaningful way

dash:projectX/import/DaSH123_SMR01_Release_v1.csv

Dataset Details

Description: In-patient hospital data capturing information throughout patient's hospital stay, diagnosis and operations performed.

Path: file:///c:/projectX/import/DaSH123_SMR01_Release_v1.csv

Row Count: 231

Unique CHIs: 211

Male/Female ratio: 49

Validations

Variables Violating Min Constraint (0): PASS

Variables Violating Max Constraint (0): PASS

Sensitive Variables Found (0): PASS

Dataset Contains Variables Not Present In Linkage Plan (0): PASS

Dataset Variables Statistics

	Data Type	Min Value	Max Value	Smallest Distinct Nu...	Complete	Complete (%)
ADMISSION DATE	date	2015-08-01	2018-12-31	1	231	100
ADMISSION REASON	string	1A		1	231	100
ADMISSION TYPE	numeric	11	99	1	231	100
DATE OF MAIN OP...	date	2015-08-01	2018-12-31	1	231	100
DATE OF MAIN OP...	date	2015-08-08	2018-12-31	1	13	6
DATE OF MAIN OP...	date	2015-09-02	2018-12-31	1	2	1
DISCHARGE DATE	date	2015-08-06	2018-12-31	1	231	100
DISCHARGE TRAN...	string	00	99	1	231	100
DISCHARGE TYPE	string	10	11	1	231	100
DaSH123 StudyNu...	numeric	2222000001	2222000658	1	231	100
MAIN CONDITION	string	D123	Z987	1	231	100
MAIN OPERATION	string	XZ98	DC321	1	231	100
MAIN OPERATION 1	string	XZ98	DC321	1	13	6
MAIN OPERATION 2	string	XZ98	XZ98	1	2	1
OTHER CONDITIO...	string	D123	D123	1	210	91
OTHER CONDITIO...	string	D123	D123	1	183	80

dash:projectX/import/DaSH123_SMR01_Release_v1.csv

DARE UK

Dataset Details

Description: In-patient hospital data capturing information throughout patient's hospital stay, diagnosis and operations performed.

Path: file:///c:/projectX/import/DaSH123_SMR01_Release_v1.csv

Row Count: 231

Unique CHIs: 211

Male/Female ratio: 49

Validations

Variables Violating Min Constraint (0): **PASS**

Variables Violating Max Constraint (0): **PASS**

Sensitive Variables Found (0): **PASS**

Dataset Contains Variables Not Present In Linkage Plan (0): **PASS**

Dataset Variables Statistics

	Data Type	Min Value	Max Value	Smallest Distinct Nu...	Complete	Complete (%)
ADMISSION DATE	date	2015-08-01	2018-12-31	1	231	100
ADMISSION REASON	string	1A		1	231	100
ADMISSION TYPE	numeric	11	99	1	231	100
DATE OF MAIN OP...	date	2015-08-01	2018-12-31	1	231	100
DATE OF MAIN OP...	date	2015-08-08	2018-12-31	1	13	6
DATE OF MAIN OP...	date	2015-09-02	2018-12-31	1	2	1
DISCHARGE DATE	date	2015-08-06	2018-12-31	1	231	100
DISCHARGE TRAN...	string	00	99	1	231	100
DISCHARGE TYPE	string	10	11	1	231	100
DaSH123 StudyNu...	numeric	2222000001	2222000658	1	231	100
MAIN CONDITION	string	D123	Z987	1	231	100
MAIN OPERATION	string	XZ98	DC321	1	231	100
MAIN OPERATION 1	string	XZ98	DC321	1	13	6
MAIN OPERATION 2	string	XZ98	XZ98	1	2	1
OTHER CONDITIO...	string	D123	D123	1	210	91
OTHER CONDITIO...	string	D123	D123	1	183	80

DARE UK

Framework for semi-automation of data provenance creation and auditing to improve risk assessment

Public Feedback incorporated into Dashboards

dash:projectX/import/DaSH123_SMR01_Release_v1.csv

Dataset Details

Description: In-patient hospital data capturing information throughout patient's hospital stay, diagnosis and operations performed.

Path: file:///c:/projectX/import/DaSH123_SMR01_Release_v1.csv

Row Count: 231

Unique CHIs: 211

Male/Female ratio: 49

Validations

Variables Violating Min Constraint (0): PASS

Variables Violating Max Constraint (0): PASS

Sensitive Variables Found (0): PASS

Dataset Contains Variables Not Present In Linkage Plan (0): PASS

Dataset Variables Statistics

	Data Type	Min Value	Max Value	Smallest Distinct Nu...	Complete	Complete (%)
ADMISSION DATE	date	2015-08-01	2018-12-31	1	231	100
ADMISSION REASON	string	1A		1	231	100
ADMISSION TYPE	numeric	11	99	1	231	100
DATE OF MAIN OP...	date	2015-08-01	2018-12-31	1	231	100
DATE OF MAIN OP...	date	2015-08-08	2018-12-31	1	13	6
DATE OF MAIN OP...	date	2015-09-02	2018-12-31	1	2	1
DISCHARGE DATE	date	2015-08-06	2018-12-31	1	231	100
DISCHARGE TRAN...	string	00	99	1	231	100
DISCHARGE TYPE	string	10	11	1	231	100
DaSH123 StudyNu...	numeric	2222000001	2222000658	1	231	100
MAIN CONDITION	string	D123	Z987	1	231	100
MAIN OPERATION	string	XZ98	DC321	1	231	100
MAIN OPERATION 1	string	XZ98	DC321	1	13	6
MAIN OPERATION 2	string	XZ98	XZ98	1	2	1
OTHER CONDITIO...	string	D123	D123	1	210	91
OTHER CONDITIO...	string	D123	D123	1	183	80

App

Project Details: This is an example project with dummy data for testing.

Variable Specification: Inspect

Released Files (2)

DaSH123_PIS_Release_v1.csv

DaSH123_SMR01_Release_v1.csv

Activities

(08/15/23 13:00:14) - NHS Data Extraction

(08/18/23 13:00:14) - Data Linkage

(08/20/23 13:56:14) - Validation Check

(08/21/23 14:56:14) - Sign Off

(08/21/23 15:56:14) - Data Release

Activity Details

Activity: Analysts joined, filtered, deidentified with pseudo-UIDs and exported datasets from NHS/external sources.

Completed by: dash:staff/JD

Inputs (3)

SMR01

PIS

dlp

Outputs (2)

DaSH123_PIS_Release_v1.csv

DaSH123_SMR01_Release_v1.csv

Validations

No validation checks set for this activity

Comments

JD 2023-08-15 14:32:17

PIS source data missing June and July 2017 records. Researchers ok'd.

DARE UK

Approach to Privacy Risk

Framework and prototype for partial automation of risk assessment of clinical free-text

DARE UK

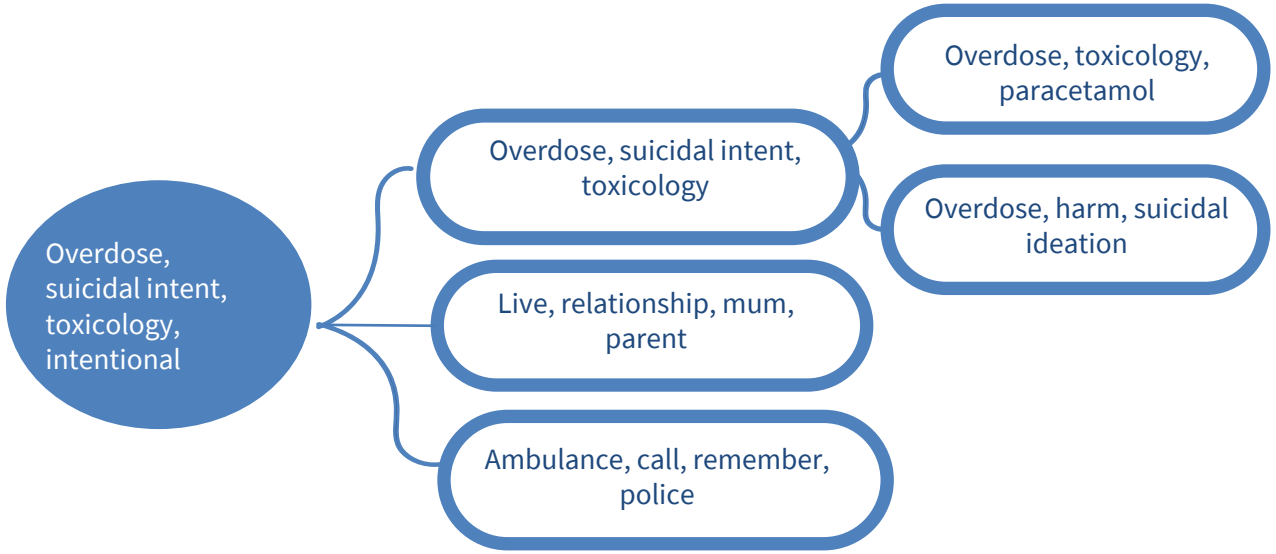
Principle Activities

~89,000 discharge summaries (age range, SIMD, Ethnicity, Sex)

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

BERTopic



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

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

Framework and prototype for partial automation of risk assessment of clinical free-text

What did we learn during our public consultation?

DARE UK

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

Participants were broadly supportive of the use of semi-automation to address privacy risks but believed:

- 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.

Framework and prototype for partial automation of risk assessment of clinical free-text

Privacy Risk Dashboard

DARE UK

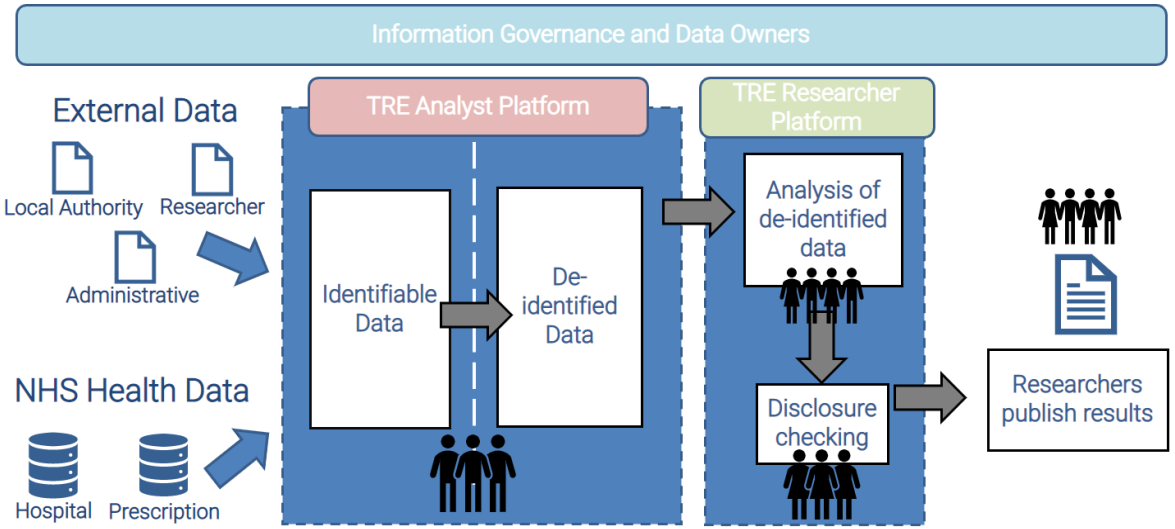
"Understanding the risks in new free text data or data within cohorts will help us make proportionate risk-based decisions faster."

DARE UK

Next Steps for Data Provenance and Privacy Risk

Next Steps

- Trial the dashboards within other Trusted Research Environments
- Create different dashboards tailored for different users
- Can the privacy risk approach be applied to other record types beyond hospital discharge summaries?
- Further public consultation to provide feedback on future iterations and ensure approach remains appropriate
- Combining the two elements into a single, overarching risk identification and assessment dashboard



DARE UK

Thank you

Stuart Dunbar, dataloch@ed.ac.uk

Engagement Manager, DataLoch

University of Edinburgh

UK Research and Innovation

HDRUK Health Data Research UK

ADRUK

THE UNIVERSITY of EDINBURGH | usher institute

DataLoch

1495 UNIVERSITY OF ABERDEEN

DaSH Grampian Data Safe Haven
University of Aberdeen • NHS Grampian

US UNIVERSITY OF SUSSEX

NHS Greater Glasgow and Clyde
West of Scotland Data Safe Haven

Public Health Scotland