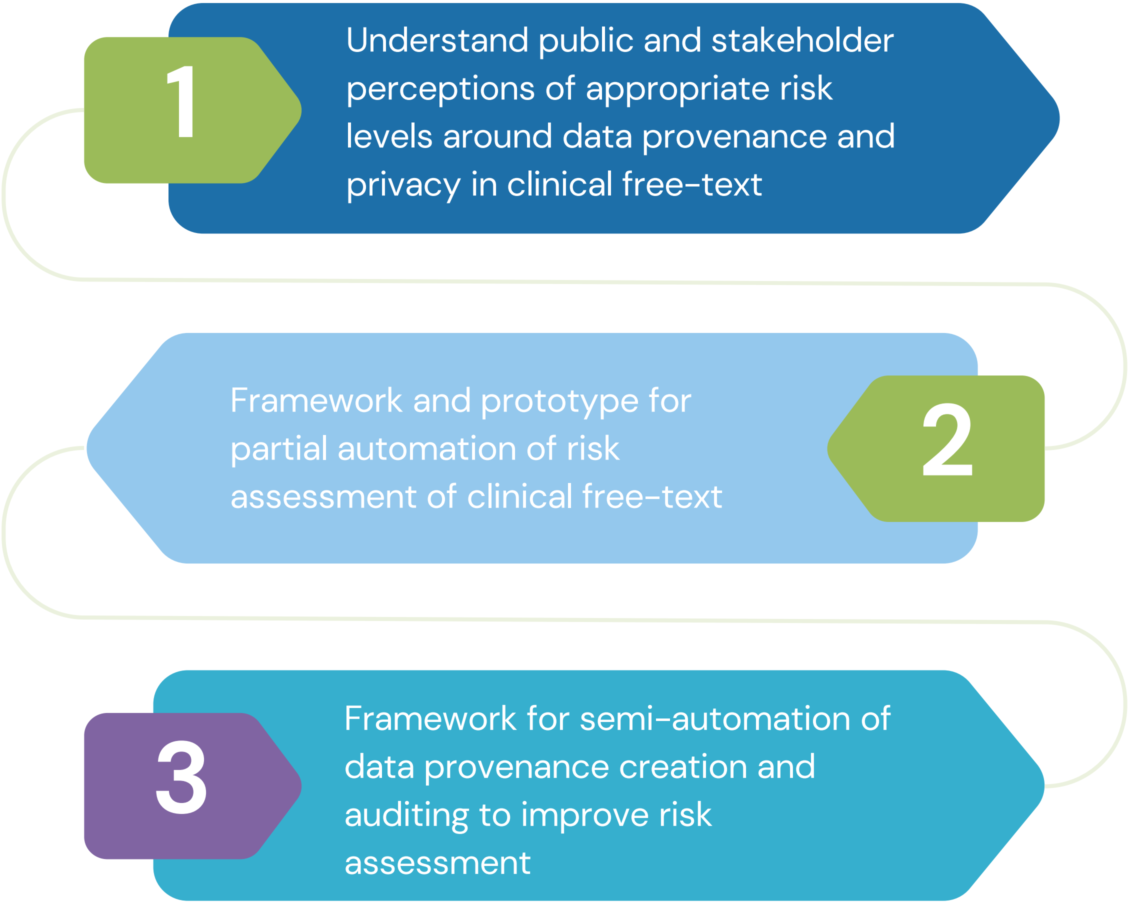
DARE UK

SARA – Mid-sprint Review Project Updates and Progress

Dr Arlene Casey

University of Edinburgh

Work Package Overview

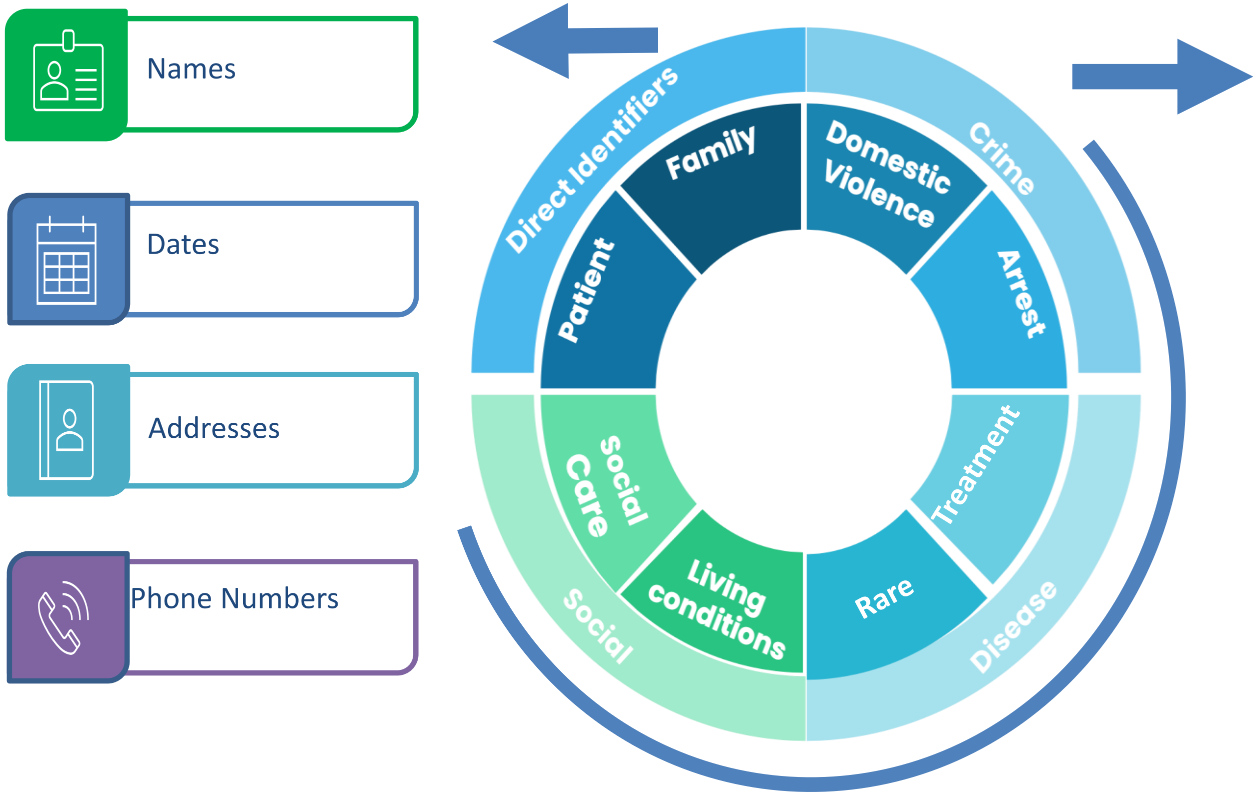


1 Understand public and stakeholder perceptions of appropriate risk levels around data provenance and privacy in clinical free-text

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

WP2: Framework and prototype for partial automation of risk assessment in clinical free-text



Names

Dates

Addresses

Phone Numbers

Direct Identifiers

Family

Patient

Crime

Domestic Violence

Arrest

Disease

Treatment

Rare

Social

Social Care

Living conditions

Indirect Identifiers: information that increases the risk of patient identification

Goal: Map & understand the risk categories

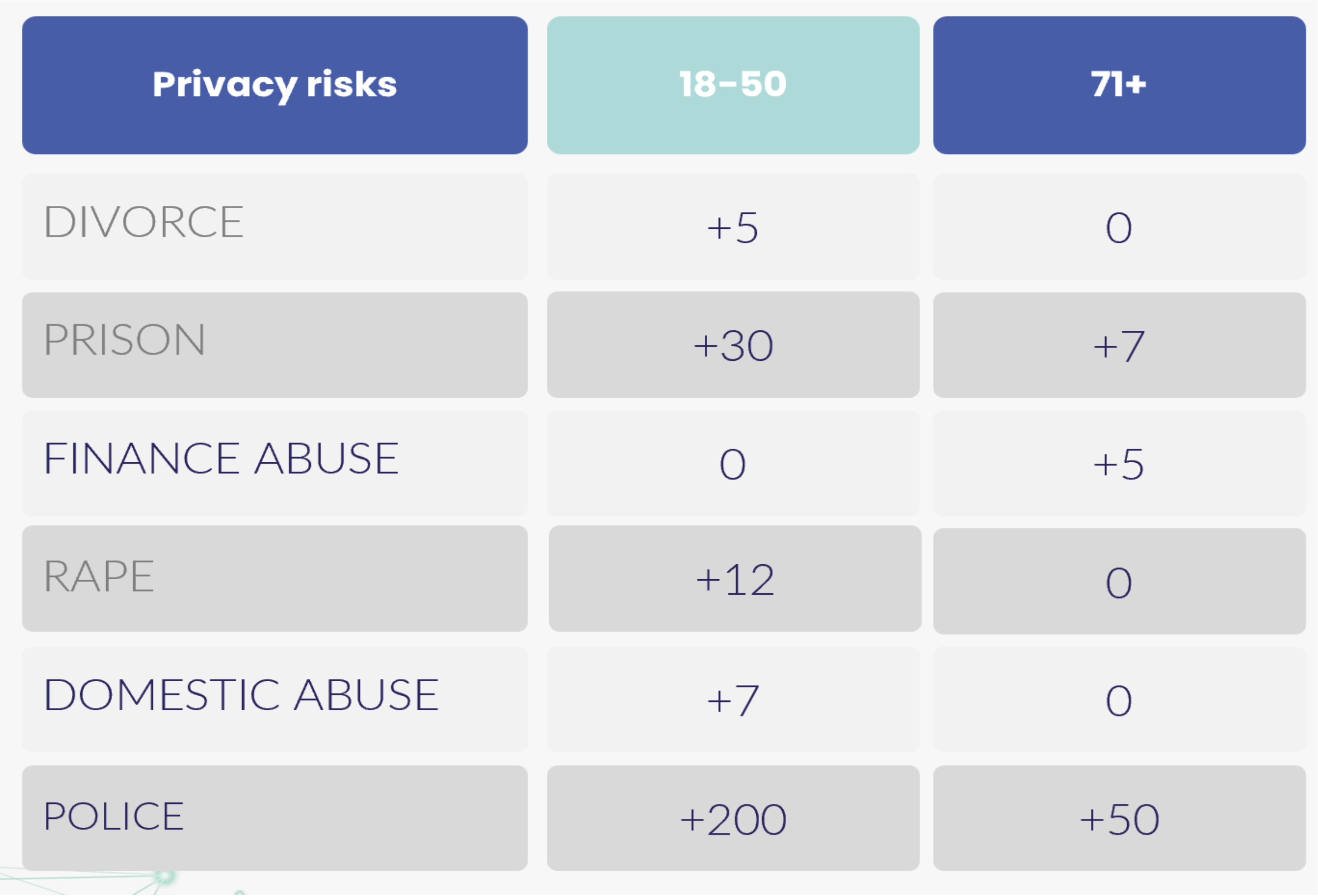
1 Year Discharge summaries + 18years 3 major NHS Lothian hospitals (age bands, SIMD, ethnicity)

Standard NLP clean-up (tokenisation, zoning, sticky keyboards)

Preliminary analysis

WP2: Framework and prototype for partial automation of risk assessment in clinical free-text

DARE UK



Privacy risks	18-50	71+
DIVORCE	+5	0
PRISON	+30	+7
FINANCE ABUSE	0	+5
RAPE	+12	0
DOMESTIC ABUSE	+7	0
POLICE	+200	+50

Used preliminary analysis to create material for PPI sessions

Examples – not a definitive list

WP2: Framework and prototype for partial automation of risk assessment in clinical free-text

- Complete our analysis and mapping of the risk categories

- Free-text Risk Dashboard
  - Appointed ‘We are rationale’
  - Design workshop 21st June, mock design, we develop prototype (august/early sept)

- Considering Publication options

WP3: Framework for semi-automation of data provenance creation and management

WP3: Interviews with TRE Analysts, Researchers, and IG

DARE UK

TRE Analysts

Have I followed correct procedures when processing data?

Have I removed all identifiable information?

Have I linked the data together per the researchers' project permissions and data specifications?

Researchers

Does the TRE Analyst understand my project, the patients I want to study and how I need the data provided to me so I can do my research?

I have not been able to see any of the identifiable data – how do I know that the data provided to me was correctly extracted and linked?

Information Governance and Data Owners

Has all data provided to researchers been correctly de-identified so the patients' confidentiality are maintained?

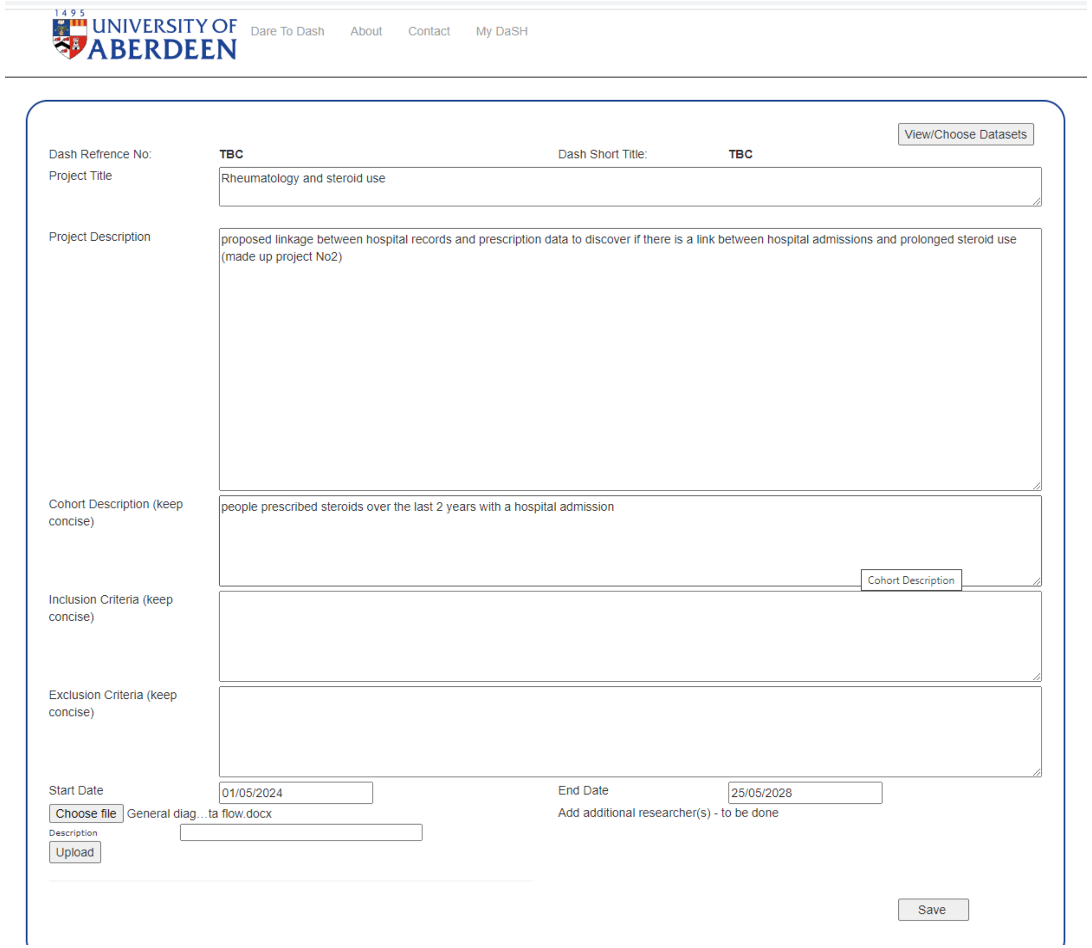
Have the TRE analysts only provided the specific data required for the research project and correctly linked it according to the project permissions?

Is there an audit trail of all steps in the data workflow so that we have proof that the data was processed correctly?

WP3: Co-design workshops

DARE UK

Provenance data capture at project start



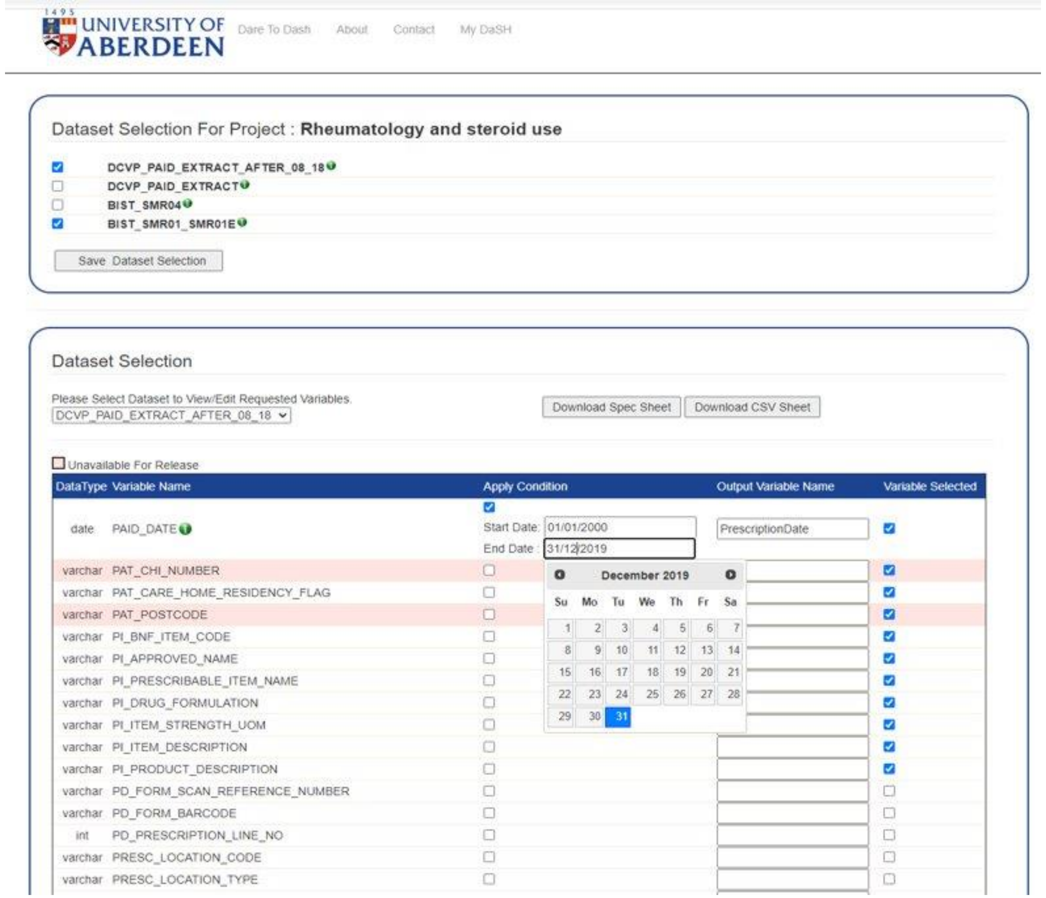
Dash Refrence No:	TBC	Dash Short Title:	TBC
Project Title	Rheumatology and steroid use		
Project Description	proposed linkage between hospital records and prescription data to discover if there is a link between hospital admissions and prolonged steroid use (made up project No2)		
Cohort Description (keep concise)	people prescribed steroids over the last 2 years with a hospital admission		
Inclusion Criteria (keep concise)			
Exclusion Criteria (keep concise)			
Start Date	01/05/2024	End Date	25/05/2028

General diag…ta flow.docx

Description

Add additional researcher(s) - to be done

Researcher creates project application via online form



Dataset Selection For Project : **Rheumatology and steroid use**

- ☑ DCVP_PAID_EXTRACT_AFTER_08_18
- ☐ DCVP_PAID_EXTRACT
- ☐ BIST_SMR04
- ☑ BIST_SMR01_SMR01E

Dataset Selection

Please Select Dataset to View/Edit Requested Variables.
DCVP_PAID_EXTRACT_AFTER_08_18

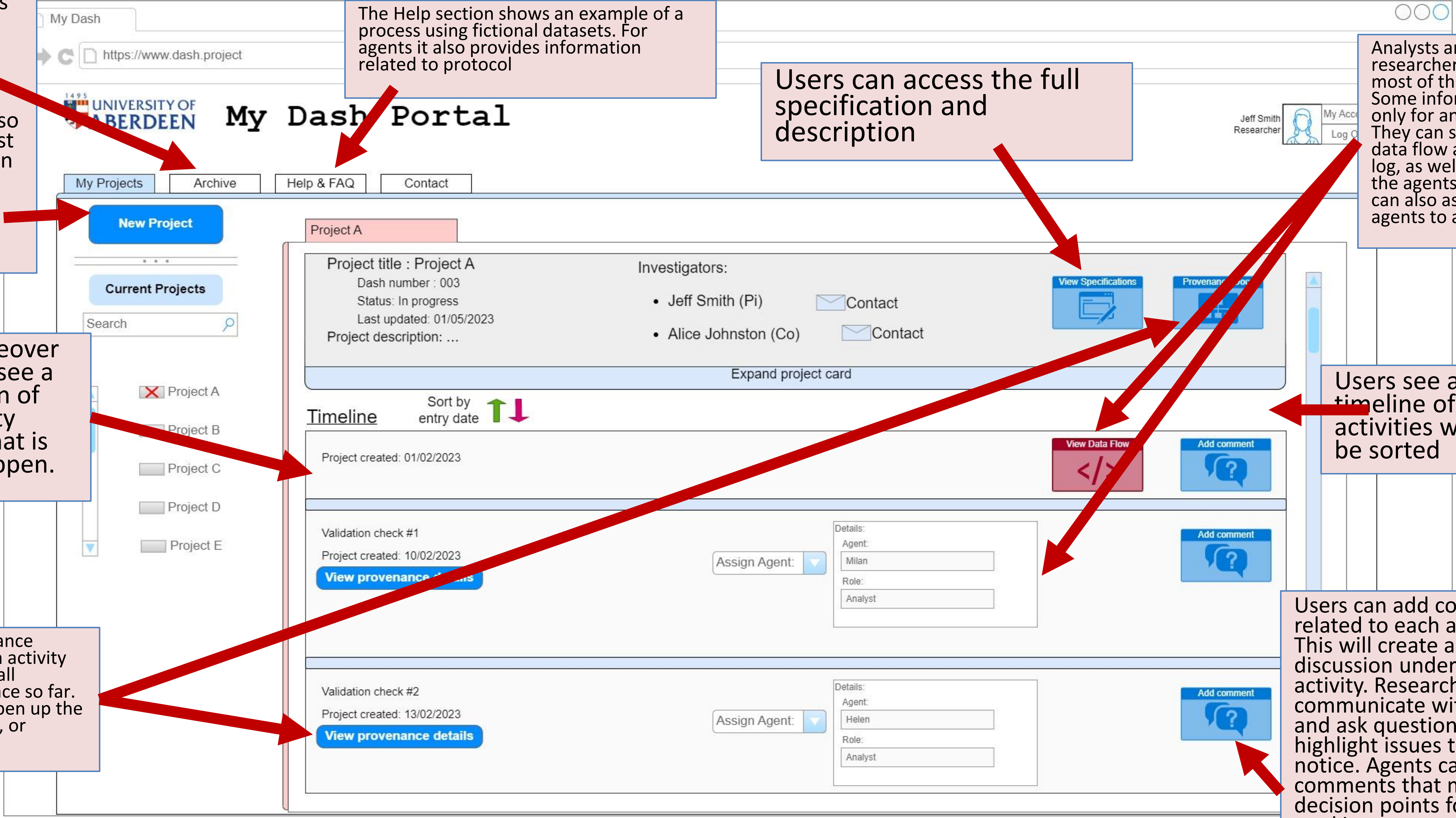
☐ Unavailable For Release

DataType	Variable Name	Apply Condition	Output Variable Name	Variable Selected
date	PAID_DATE	☑ Start Date: 01/01/2000 End Date : 31/12/2019	PrescriptionDate	☑
varchar	PAT_CHI_NUMBER	☐		☑
varchar	PAT_CARE_HOME_RESIDENCY_FLAG	☐		☑
varchar	PAT_POSTCODE	☐		☑
varchar	PI_BNF_ITEM_CODE	☐		☑
varchar	PI_APPROVED_NAME	☐		☑
varchar	PI_PRESCRIBABLE_ITEM_NAME	☐		☑
varchar	PI_DRUG_FORMULATION	☐		☑
varchar	PI_ITEM_STRENGTH_UOM	☐		☑
varchar	PI_ITEM_DESCRIPTION	☐		☑
varchar	PI_PRODUCT_DESCRIPTION	☐		☑
varchar	PD_FORM_SCAN_REFERENCE_NUMBER	☐		☐
varchar	PD_FORM_BARCODE	☐		☐
int	PD_PRESCRIPTION_LINE_NO	☐		☐
varchar	PRESC_LOCATION_CODE	☐		☐
varchar	PRESC_LOCATION_TYPE	☐		☐

Researcher selects variables and defines applicable constraints (e.g., date ranges, min/max values or string values)

WP3: Co-designed prototype dashboard

DARE UK



Project A

Project title : Project A

Dash number : 003

Status: In progress

Last updated: 01/05/2023

Project description: ...

Investigators:

- Jeff Smith (Pi) Contact
- Alice Johnston (Co) Contact

Timeline

Sort by entry date

Project created: 01/02/2023

Validation check #1

Project created: 10/02/2023

Assign Agent:

Details:

Agent: Milan

Role: Analyst

Validation check #2

Project created: 13/02/2023

Assign Agent:

Details:

Agent: Helen

Role: Analyst

Completed projects go in the Archive. Open projects can be accessed from the navigation.

Researchers can also make a new request (either integrated in the portal, or redirected to the request platform)

The Help section shows an example of a process using fictional datasets. For agents it also provides information related to protocol

Users can access the full specification and description

Analysts and researchers share most of the view. Some information is only for analysts. They can see the full data flow and record log, as well as who the agents are. They can also assign other agents to an activity.

Users can mouseover each activity to see a short description of what that activity refers to and what is supposed to happen.

Users see a timeline of all activities which can be sorted

Users can see provenance details related to each activity or a summary file for all activities that took place so far. This allows them to open up the data provenance card, or download it.

Users can add comments related to each activity. This will create a discussion under each activity. Researchers can communicate with agents and ask questions or highlight issues they notice. Agents can add comments that motivate decision points for tracking.

WP3: Detailed project-specific dashboard prototype

DARE UK

Project information

Project title: Project A

DaSH number: 003

PI: Jeff Smith

Last update: 01/05/2023

Current activity

Data Selection #1

01/05/2023

Agent: Adrian

Role: Lead Analyst

No potential issues identified during this activity

A short summary of the provenance highlighting the list of datasets, row counts, variables, number of records, cohort specification used and comparison to the provided specification.

View specification

View code

Dataset (version)	Last updated	Extracted Variables	Number of records extracted	Cohort specification
Dataset 001 (v1)	03/04/2022	AGE, X, Y, Z	30000	25 year old patients born in March
Dataset 002 (v3)	02/01/2021	AGE, X, Y, Z	30000	25 year old patients born in March
Dataset 003 (v1)	02/01/2021	AGE, X, Y, Z	30000	25 year old patients born in March
Dataset 004 (v2)	03/05/2009	AGE, X, Y, X	30000	25 year old patients born in March

Note: Provenance details are shown depending on the current activity. Other activities will require other information, a mapping of which information is used during which activity needs to be done.

Examples of other information:

Flagging of identifiable information fields

Basic statistics

Aggregate breakdown

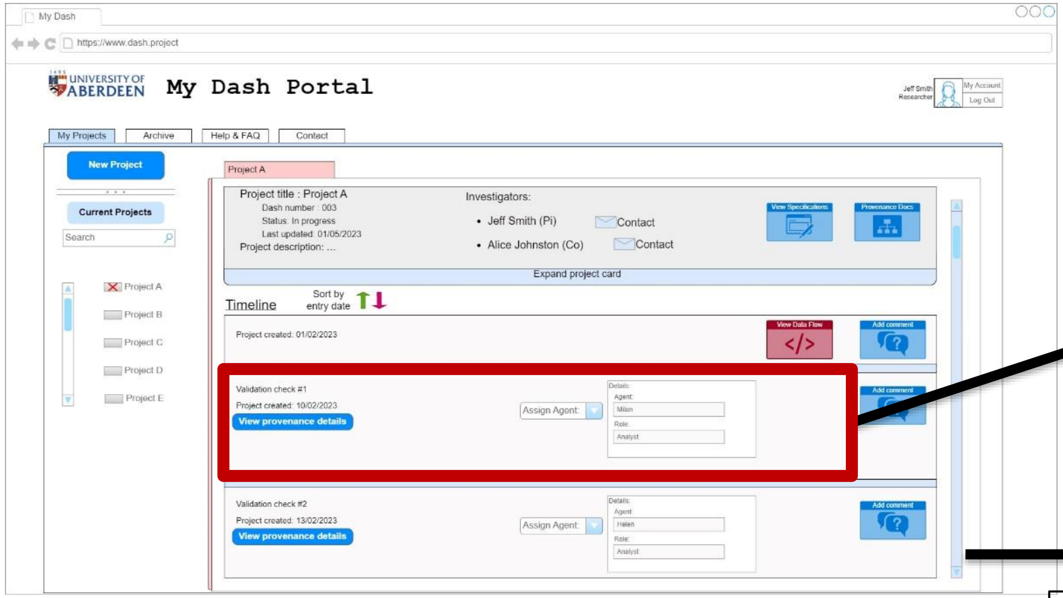
List of released datasets, file locations, dates released

Dataset last update

How many people linked/not linked/invalid or missing linkage variable

WP3: Detailed project-specific dashboard prototype

DARE UK



My Dash Portal

Project title : Project A

Dash number : 003

Status: In progress

Last updated: 01/05/2023

Project description: ...

Investigators:

- Jeff Smith (Pi) Contact
- Alice Johnston (Co) Contact

Timeline

Project created: 01/02/2023

Validation check #1

Project created: 10/02/2023

Details:

Agent: Milan

Role: Analyst

Validation check #2

Project created: 13/02/2023

Details:

Agent: Helen

Role: Analyst

```
{
    "@id": "file//:ProjectA/validationCheck_1",
    "@type": ["CreateAction", shp:ValidationCheck],
    "agent": {
      "@id": "https://www.abdn.ac.uk/people/katherine.osullivan/"
    },
    "object": {
      "@id": "file//:ProjectA/data.csv"
    },
    "result": [
      {
        "@id": "file//:ProjectA/ValidationCheckReport.csv"
      }   ]
  }
 {
    "@id": "file//:ProjectA/ValidationCheckReport.csv",
    "@type": ["File", shp:ValidationCheckReport],
    "description": "This report contains...",
    ……. }
```

Dataset (version)	Last updated	Extracted Variables	Number of records extracted	Cohort specification
Dataset 001 (v1)	03/04/2022	AGE, X, Y, Z	30000	25 year old patients born in March
Dataset 002 (v3)	02/01/2021	AGE, X, Y, Z	30000	25 year old patients born in March
Dataset 003 (v1)	02/01/2021	AGE, X, Y, Z	30000	25 year old patients born in March
Dataset 004 (v2)	03/05/2009	AGE, X, Y, X	30000	25 year old patients born in March

```
{
    "@id": "file//:ProjectA/dataset001_v1.csv",
    "@type": ["File", shp:DatasetRelease],
    shp:hasHash: "dfdec888b72151965a34b4b59031290a",
    "description": "Results from the PIS dataset matching the cohort criteria for 25 year old patients born in May between 1980 - 2020",
    "encodingFormat": "csv",
    prov:wasDerivedFrom: [file//:ProjectA/DisclosureCohortSpec.csv, PIS-Dataset#2eb]
    "exifData": [
     {
      "@id": "file//:ProjectA/dataset001_v1.csv#2eb90b09"
     },
     ……     ]
}
{
    "@id": "file//:ProjectA/data.csv#2eb90b09",
    "@type": ["PropertyValue", shp:ExtractedVariables],
    prov:hadMember: [ https://https://www.abdn.ac.uk/iahs/facilities/grampian-data-safe-haven/catalogue/variables/AGE,
https://https://www.abdn.ac.uk/iahs/facilities/grampian-data-safe-haven/catalogue/variables/X, …]
    }
```

WP3: Framework for semi-automation of data provenance creation and management

DARE UK

Next Steps

- Incorporate PPIE workshop feedback into low-fi designs and formalise
- Develop GUI for Safe Haven dashboards and deploy in Safe Haven.
- Update SHP Ontology
- Report detailing interviews/co-design workshops
- Deploy any user feedback / formal evaluation incorporated into final reports
- If possible, deploy in NHS environment (dependent on NHS Research passport being granted)
- Considering publication options

WP1: Understand public and stakeholder perceptions of appropriate levels of risk around data provenance and privacy in clinical free-text.

WP1: Understand public and stakeholder perceptions of appropriate levels of risk around data provenance and privacy in clinical free-text.

DARE UK

- Working with Ipsos Scotland on design and delivery
- Learning session (online) held informing 40 participants (from Edinburgh and Aberdeen regions) about risk assessment of clinical free text and risk mitigation using data provenance:
    - Intro to health care data and opportunities/challenges for research
    - What data provenance is and why it matters
    - The challenge of indirect identifiers in unstructured data

WP1: Overarching questions for the deliberative workshops

1. What type of record-keeping should Trusted Research Environments provide to ensure a transparent process, while also keeping data confidential? (WP2)

2. When providing researchers with access to free-text patient data, how should Trusted Research Environments maintain confidentiality to ensure trustworthiness? (WP3)

3. How can semi-automating processes help make record-keeping and the maintenance of confidentiality more robust yet still trustworthy? (Across WPs)

WP1: Examples shared with participants

❑ 5 Discharge Summaries

❑ 3 Case Studies with Example Dashboards

WP1 example: Case Study 1

Daisy is a Data Analyst working at the University of Aberdeen’s Grampian Data Safe Haven. Her work requires her to extract, pseudonymise and link routinely collected but unconsented health and social care data on behalf of researchers, who cannot access patient-identifiable data to protect patient confidentiality and privacy.

Daisy’s current work is supporting a researcher, Tom, on his project that involves looking at children’s mental health and whether children receive specialist support when they have been referred by their GP or whether they have visited a hospital to receive acute treatment, and whether children have received any psychiatric prescriptions either by their GP or via the hospital. Tom’s cohort are children 5-18 in the last 10 years that meet these conditions.

Daisy is aware that this is a particularly sensitive project because it involves children and a mental health diagnosis, and requires a data provenance output that will provide her with assurances that she has extracted and linked the data according to the legal and ethical permissions of the project.

Dashboard 1 – For TRE Analysts during Extraction and Linkage

Dataset	Field Name	Total cohort	% of cohort	Minimum Value	Maximum Value
GP Referral	Age	18,000	100%	4	18
A&E	Main Condition 1	11,000	61%	Patient Injury - Road Traffic Accident (RTA)	Patient Injury - Self Inflicted (Injury or Poisoning)
A&E	Main Condition 2	7,000	39%	Patient Injury - Self Inflicted (Injury or Poisoning)	Patient Injury - Self Inflicted (Injury or Poisoning)
Prescribing Information	Drug	9,000	50%	Abilify	zolpidem
ALL	Patient ID	18,000	100%	1580346223	15000180001

WP1: Dashboard 1 – For TRE Analysts during Extraction and Linkage

Dataset	Field Name	Total cohort	% of cohort	Minimum Value	Maximum Value	Error
GP Referral	Age	18,000	100%	**4**	18	**Yes**
A&E	Main Condition 1	11,000	61%	**Patient Injury - Road Traffic Accident (RTA)**	Patient Injury - Self Inflicted (Injury or Poisoning)	**Yes**
A&E	Main Condition 2	7,000	39%	Patient Injury - Self Inflicted (Injury or Poisoning)	Patient Injury - Self Inflicted (Injury or Poisoning)	OK
Prescribing Information	Drug	9,000	50%	Abilify	zolpidem	OK
ALL	Patient ID	18,000	100%	**1580346223**	15000179999	**Yes**

Age is outside of range 5-18

Main condition is not Mental Health related

Patient ID has not been anonymised to an 11-digit number

WP1: What we've learnt so far (subject to further analysis ahead of full reporting)

DARE UK

- Data provenance
  - Central dashboards considered sensible approach to record-keeping
  - Level of detail debated due to speeding-up processes (rather than identifiability concerns)
  - Practical suggestions for dashboards:
    - Sort so that error rows appear first
    - Include explanations of errors in interface (i.e. the yellow boxes)
    - Avoid no error green text to avoid complacency
    - IG manager should still spot-check for errors throughout

WP1: What we've learnt so far (subject to further analysis ahead of full reporting)

DARE UK

- Accessing free-text data
  - Hard to apply one-size-fits-all – depends on research purpose/interest which is variable
  - TREs and researchers to collaborate more (e.g. earlier involvement of researchers) to ensure things working 'properly' (further analysis to interrogate meaning)
  - Coding/rewording data to make less specific but still research-useful (e.g. age-bands rather than age; location type rather than location)
  - Standardise processes across TREs (transparency / trustworthiness)

WP1: What we've learnt so far (subject to further analysis ahead of full reporting)

- Semi-automation
  - Participants generally comfortable with idea
    - Speed-up process and assist with volume
    - Ensure humans remain part of the decision-making around risks
  - Ensure different languages are handled (e.g. Gaelic)
  - Consider how to improve consistency of original notes
    - Work with practitioners to limit inclusion of indirect identifiers within free-text: "Semi-automation is only as good as the person putting the information in and the person taking the information out."

WP1: Next PIE steps

DARE UK

- Public survey (target: 1000 respondents) in development
- Questions are being informed by learning from the workshops:
  - What are the gaps that remain?
  - Ranking of options suggested by workshop participants
- Final report to be complete in August
  - Online publication (e.g. through the DataLoch website)

- Development of full workshop report
  - Draft to be received in early July
  - Period of refinement
  - Online publication

DARE UK

Questions / comments?

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

UK Research and Innovation

HDRUK Health Data Research UK

ADRUK