Plan Overview

A Data Management Plan created using DMPonline

Title: Efficient Screening in Medical Systematic Reviews

Creator: Aaron Fletcher

Principal Investigator: Aaron Fletcher

Data Manager: Aaron Fletcher

Project Administrator: Aaron Fletcher

Affiliation: The University of Sheffield

Funder: UKRI Future Leaders Fellowships

Template: Postgraduate Research DMP (The University of Sheffield)

ORCID iD: 0000-0002-4776-066X

Project abstract:

Continuous Active Learning (CAL) has emerged as a promising technique to enhance the efficiency and accuracy of systematic reviews in medicine. This PhD proposal investigates the application of CAL, specifically focusing on the title and abstract screening substage of systematic reviews. The primary goal is to minimise expert intervention while maintaining high accuracy in document classification, thereby addressing the increasing volume of research and limited resources in healthcare. The PhD research will focus on screening prioritisation and stopping methods and utilising datasets such as CLEF-TAR and the Synergy Dataset, which provide real-world scenarios and imbalances typical of systematic reviews.

ID: 170395

Start date: 18-03-2025

End date: 18-09-2027

Last modified: 11-02-2025

Grant number / URL: EP/S023062/1

Efficient Screening in Medical Systematic Reviews

Defining your data

- What digital data (and physical data if applicable) will you collect or create during the project?
- How will the data be collected or created, and over what time period?
- What formats will your digital data be in? (E.g. .docx, .txt, .jpeg)
- Approximately how much digital data (in GB, MB, etc) will be generated during the project?
- Are you using pre-existing datasets? Give details if possible, including conditions of use.

What digital data (and physical data, if applicable) will you collect or create during the project?

The data used for this PhD project will be digital.

- **PubMed ID:** a unique string that refers to a piece of research published on PubMed.
- **Relevancy Judgement:** A binary string (0,1) representing a decision as to whether that research paper is relevant to a search question.
- **Recombined Abstract:** A string which contains information reformed from an abstract inverted index sourced fromOpenAlex.
- **Paper Metadata**: Metadata regarding the PubMed ID (i.e. Author, Cititing Works, Journal Publication Source). Source from OpenAlex (an open-source bibliographic catalogue).

How will the data be collected or created, and over what time period?

OpenAlex will be used to collect metadata and recombine abstracts for a given PubMed ID using API interface provided by OpenAlex. As the number of PubMed IDs to mine is finite, this will be done over 24 hours.

What formats will your digital data be in? (E.g. .docx, .txt, .jpeg)

Txt Files, with each text file corresponding to a PubMed ID.

Approximately how much digital data (in GB, MB, etc) will be generated during the project?

Anticipated to be > 1 gig in size (estimated based on a CLEF post-processed dataset (containing only title and abstract but not metadata) used by other researchers is 207.3 MB.

Are you using pre-existing datasets? Give details if possible, including conditions of use

Pubmed ID and relevancy judgments are sourced from the CLEF (under an MIT license) and Synergy datasets (under a CC0-1.0 license). These are standard datasets used within the field of information retrieval.

Looking after data during your research

- Where will you store digital data during the project to ensure it is secure and backed up regularly? (University research storage)
- How will you name and organise your data files? (An example filename can help to illustrate this)
- If you collect or create physical data, where will you store these securely?
- How will you make data easier to understand and use? (E.g. include file structure and methodology in a README file)
- Will you use extra security precautions for any of your digital or physical data? (E.g. for sensitive and/or personal data)

Data Storage

- Data will be stored in the university research storage and automatically mirrored in up to two locations.

No physical data will be collected during this study.

Data Organisation

Each dataset (CLEF/Synergy) in the study will have a folder. The data will be laid out in the following format:

```
|-- Dataset
|   |-- Year
|   |   |-- [PubMed ID]
|   |   |   |-- Title and Abstract.txt
|   |   |   |-- Metadata.txt
| README.txt
```

Explanations of the data creation process will be placed into the README.txt file.

As a PubMed ID can appear in different years within this dataset, it is not a valid unique identifier for each article. However, a unique identifier for each article is assured through the layout: dataset-year-PubMed ID

Will you use extra security precautions for any of your digital or physical data? (E.g. for sensitive and/or personal

data)
No sensitive or personal data is collected for this research.

Storing data after your research

- Which parts of your data will be stored on a long-term basis after the end of the project?
- Where will the data be stored after the project? (E.g. University of Sheffield repositoryORDA, or a subject-specific repository)
- How long will the data be stored for? (E.g. standard TUoS retention period of minimum 10 years after the project)
- Who will place the data in a repository or other long-term storage? (E.g. you, or your supervisor)
- If you plan to use long-term data storage other than a repository, who will be responsible for the data?

Which parts of your data will be stored on a long-term basis after the end of the project?
Code generated for this project and results will be stored on GitHub, and the University repositoryORDA.

Sharing data after your research

- How will you make data available outside of the research group after the project? (E.g. openly available through a repository, or on request through your department)
- Will you make all of your data available, or are there reasons you can’t do this? (E.g. personal data, commercial or legal restrictions, very large datasets)
- If there are reasons you can’t share all of your data, how might you make as much of it available as possible? (E.g. anonymisation, participant consent, sharing analysed data only)
- How will you make your data as widely accessible as possible? (E.g. include a data availability statement in publications, ensure published data has a DOI)
- What licence will you apply to your data to say how it can be reused and shared? (E.g. one of theCreative Commons licences)

How will you make data available outside of the research group after the project? (E.g. openly available through a repository, or on request through your department)
Not applicable, all data generated for this project can be recreated. The methods for recreating this data will be made available publically via GitHub, along with the source code for any projects. This data will also be placed in ORDA. Additionally, sharing abstracts in their reconstituted form might breach copyright laws, depending on the source (and therefore, it would be easier to share the approach to generating the dataset rather than the actual data).

Putting your plan into practice

- Who is responsible for making sure your data management plan is followed? (E.g. you with the support of your supervisor)
- How often will your data management plan be reviewed and updated? (E.g. yearly and if the project changes)
- Are there any actions you need to take in order to put your data management plan into practice? (E.g. requestingUniversity research storage via your supervisor.)

Ensuring the project's compliance with this DMP is the responsibility of the Data manager. Reviews of the DMP will occur every 3 months during the project until the project is completed.