WEATHERED:

Data Management Plan

**Title:** **W**eather **E**ffects on **A**mbulance and **T**reatment in **H**ip **E**mergency **R**esponse and **E**mergency **D**epartments [WEATHERED]

**Researcher:** Paul Harford

**Supervisors:** Dr Alison O’Connor & Dr Meghana Kshirsagar

This data management plan (DMP) outlines how the WEATHERED project will maintain and control data during the lifecycle of the project. **The DMP is considered a living document, as such, this document is subject to modification.** The key purpose of the WEATHERED project is to analyse how hip fracture patient demand is informed by population demographics and local weather patterns. In this research we will integrate multiple datasets including National Ambulance Service (NAS) patient records, Irish Hip Fracture Database (IHFD) patient records, population demographics (central statistics office, CSO) and weather reports (Met Éireann).

# Introduction

This research plans to use historical data from multiple sources (see Section 2). The objective of this analysis is to integrate these data sources and devise a machine learning pipeline with predictive capabilities. Programming in this research will be conducted using Python programming language. Packages used in analyses will be managed via Anaconda (Anaconda, 2016). using self-contained isolated environments. Reproducibility of coding exercises will be ensured by exporting the environment configuration.

This project plans to use existing historical data form multiple sources specific guidance on each source is provided in Section 2. Some datasets cannot be published publicly to ensure individual privacy and compliance with ethical policies, further information is provided in Section 6. Where possible aggregated/anonymised datasets will be made available. Data manipulation and processing will be conducted using modern, acceptable programming packages such as: Pandas, SQL and/or Polars to facilitate the application of machine learning.

To facilitate increased understanding jupyter notebooks (file extension ‘.ipynb’), which enable both text markup and code-based analysis in a single location, will be used where possible. To demonstrate data insights and support understanding of statistical inferences visualisations will comprise a significant output of this work. Images created in this project are likely to be file types: \*.png, \*.jpg, \*.eps.

# Data sources

The timeframe of interest to this research is 2013 – 2023. However, given the use of multiple datasets revisions to the timeframe may be required if, for example, a particular dataset contains gaps in the requested timeframe. Specifics related to reduction/removal of information in datasets will be documented as needed. In the following sections we address specific information related to each individual data source.

## National Ambulance Service (NAS) database

NAS employs the Advanced Medical Priority Dispatch System (AMPDS) to triage and prioritise emergency calls. As this research is focussed on hip fracture patients, the NAS dataset will be limited to data related to AMPDS code 17 which represents patients with ‘fall’ related injuries. The database is currently known to include items as listed in Table 1.

Table 1 NAS dataset identifiers, data type and explanatory comments

|  |  |  |
| --- | --- | --- |
| Header | Data Type | Comments |
| Incident Location | String | Postcode/county/address? |
| Date | Date | Date of incident in for DD/MM/YYY |
| 999 time | Time | Time of emergency services call in HH:MM:SS |
| Destination ED | String | Destination emergency department |
| Patient age | Integer | Patient age in years |
|  |  |  |

Analysing the NAS data will require specific types of data output. The expected types of data to be generated from the dataset are listed in Table 2.

Table 2 List of NAS dataset file extensions and types generated during analysis.

|  |  |
| --- | --- |
| File extension | Comments |
| .csv | Pre-processed data will be stored in .csv |
| .log | Log file containing start and end times of analysis |
| png/jpeg | Outputs of analysis in visual format |
| .ipynb | Jupyter notebook created to execute data processing and machine learning. |

## Irish Hip Fracture Database (IHFD)

The Irish Hip Fracture Database (IHFD) is a clinically led, audit that measures care provision patients aged 60+ years. The IHFD operates across sixteen HSE public hospitals in the Republic of Ireland (ROI).

The database is currently known to include items as listed in Table 3. This will expand on access to the data as required.

Table 3 IHFD dataset identifiers, data type and explanatory comments

|  |  |  |
| --- | --- | --- |
| Header | Data Type | Comments |
| Date | Date | Date of incident in for DD/MM/YYY |
| Destination ED | String | Destination emergency department |
| Patient age | Integer | Patient age in years |
|  |  |  |

Analysing the IHFD data will require specific types of data output. The expected types of data to be generated from the dataset are listed in Table 4.

Table 4 List of IHFD dataset file extensions and types generated during analysis

|  |  |
| --- | --- |
| File extension | Comments |
| .csv | Pre-processed data will be stored in .csv |
| .log | Log file containing start and end times of analysis |
| png/jpeg | Outputs of analysis in visual format |
| .ipynb | Files created to run machine learning and data processing tasks. |
|  |  |
|  |  |

## Central Statistics Office (CSO) database

Census data is a detailed count of every person living in Ireland on a particular date. The census takes place every 5 years by the Central Statistics Office (CSO). This data is made publicly available [REF].

The database header , data type and explanatory comments will be added when the data has been analysed

Table 5 Census dataset identifiers, data type and explanatory comments

|  |  |  |
| --- | --- | --- |
| Header | Data Type | Comments |
| Date | Date |  |
|  | String |  |
|  | Integer |  |
|  |  |  |

Analysing the Census data will require specific types of data output. The expected types of data to be generated from the dataset are listed in Table 6.

Table 6 List of Census dataset file extensions and types generated during analysis

|  |  |
| --- | --- |
| File extension | Comments |
| .csv | Pre-processed data will be stored in .csv |
| .log | Log file containing start and end times of analysis |
| png/jpeg | Outputs of analysis in visual format |
| .ipynb | Files created to run machine learning and data processing tasks. |

## Met Éireann database

The Met Eireann weather event data is publicly available data which provides information on adverse weather events in Ireland in the CAP (Common Alerting Protocol) format.

The database is currently known to include items as listed in Table 7. This will expand on access to the data as required.

Table 7 IHFD dataset identifiers, data type and explanatory comments

|  |  |  |
| --- | --- | --- |
| Header | Data Type | Comments |
| sender | object | sender |
| sent | datetime | sent |
| status | object | status |
| msgType | object | msgType |
| scope | object | scope |
| language | object | language |
| category | object | category |
| event | object | event |
| responsetype | object | responsetype |
| urgency | object | urgency |
| severity | object | severity |
| certainty | object | certainty |
| effective | datetime | effective |
| onset | datetime | onset |
| expires | datetime | expires |
| senderNme | object | senderNme |
| headline | object | headline |
| description | object | description |
| instruction | object | instruction |
| param\_awareness\_level | object | param\_awareness\_level |
| areaDesc | object | areaDesc |
| counties | object | counties |
| param\_awareness\_type | object | param\_awareness\_type |
| param\_situation | object | param\_situation |
| standardized\_level | object | standardized\_level |
| standardized\_severity | object | standardized\_severity |
| standardized\_type | object | standardized\_type |
| sender | object | sender |
| sent | datetime | sent |
| status | object | status |
| msgType | object | msgType |
| scope | object | scope |
| language | object | language |
| category | object | category |
| event | object | event |
| responsetype | object | responsetype |
| urgency | object | urgency |
| severity | object | severity |
| certainty | object | certainty |
| effective | datetime | effective |
| onset | datetime | onset |
| expires | datetime | expires |
| senderNme | object | senderNme |
| headline | object | headline |
| description | object | description |
| instruction | object | instruction |
| param\_awareness\_level | object | param\_awareness\_level |
| areaDesc | object | areaDesc |
| counties | object | counties |
| param\_awareness\_type | object | param\_awareness\_type |
| param\_situation | object | param\_situation |
| standardized\_level | object | standardized\_level |
| standardized\_severity | object | standardized\_severity |
| standardized\_type | object | standardized\_type |

Analysing the IHFD data will require specific types of data output. The expected types of data to be generated from the dataset are listed in Table 8.

Table 8 List of Met Eireann dataset file extensions and types generated during analysis

|  |  |
| --- | --- |
| File extension | Comments |
| .csv | Pre-processed data will be stored in .csv |
| .log | Log file containing start and end times of analysis |
| png/jpeg | Outputs of analysis in visual format |
| .ipynb | Files created to run machine learning and data processing tasks. |

# Data security

For datasets containing sensitive information the following protocols are in place.

## Ethical aspects

The IHFD database requires ethical approval prior to data access. Ethical approval provided by the University of Limerick’s Science & Engineering (UL S&E) faculty. The application *‘An assessment of hip-fractures in Ireland’* was submitted to UL S&E ON 28/06/2024 and was approved by committee on 26/07/2024.

## Data pseudonymisation and anonymity

The data pseudonymisation process for the IHFD is currently unknown. The standard approach is to have a team responsible for ingesting, processing, transforming, pseudonymising and storing records in a format accessible for statistical analysis. In the pseudonymisation process, all personal identifiers are removed. This is done by converting the identifier numbers in each dataset to a Protected Identifier Key (PIK). PIKs are a unique and non-identifiable number. Using PIKs enables the data owners and researchers to protect and secure the confidentiality of the data.

Specifically, The IHFD database is pseudonymised and does not contain any identifying patient information. The NAS dataset will also be provided pseudonymised and will not contain any identifying patient information.

For this research the University of Limerick will receive pseudonymised data only. Keys to deanonymise the data will be held by the data controller.

## Data security

Data security is ensured through cloud storage. Changes to the sensitivity of data will be addressed on an ongoing basis through the life of the project. Text documents (e.g. \*.tex, \*.txt, \*.doc, \*docx, \*.pdf) will be stored on the researchers hard-drive (HD) and backed up via cloud storage (it is expected that OneDrive and other Microsoft cloud products will be used).

Code files (e.g. \*.py, \*.ipynb) will be stored on the researchers HD and backed up via GitHub. All data is stored and managed in accordance with the University of Limerick’s data protection policy.

Published data, uploaded to GitHub on the *ul\_project\_Msc\_AI* repository, will be stored for long term preservation and curation. Restricted or unpublished data will also be uploaded to the repository with relevant access restrictions as described previously.

# Data Processing

This document will be modified on an ongoing basis to update information related to data management as needed.

## Data integration

Figure 1 shows the expected initial data integration.



Figure 1 Initial data integration to create combined dataset.

## Data Processing

### NAS database

The NAS data will be processed into a panda’s data frame for initial analysis.

### IFHD database

The IFHD data will be processed into a panda’s data frame for initial analysis.

### CENSUS database

The CENSUS data will be processed into a panda’s data frame for initial analysis. The analysis will be done using pandans, matplotlib, seaborn and python. The exploratory data analysis will determine the relevant data needed to use with the other defined datasets.

### MET Eireann database

The Met Eireann data comes in xml format with individual files for each weather event. These 10k + files were parsed using various python modules (xml.etree, glob, os, codecs, tqdm). This parsed data was added to a panda’s data frame and some additional columns were added to standardize on the weather event type and level and severity. Further analysis will be performed to determine the most relevant data. Null or Unknown values will be investigated in the actual xml file to determine the reason for the null or unknown value.

## Data analysis

The initial algorithms used for data analysis will be regression based like Linear, Poisson, Binomial. These are currently being tested and under review. When the review has been completed the algorithms, details will be updated in this section.

# FAIR Data

At the end of the project (or earlier to support publications), data will be uploaded to the online repository under the University of Limerick Research Community. All uploads will contain the project name ‘WEATHERED’ in the title and will be accompanied by clear descriptions of the contents.

Individual files will be named consistently and comprise of the following information:

1. The name of the file in letters with underscores used in place of spaces (e.g. some\_file\_name)
2. The file extension which can be used to identify appropriate software programmes for reading file data.

Efforts will be made to ensure filenames do not exceed 30 characters but are descriptive enough to provide insight into the data contained within the file. Special characters (e.g. ampersand, question mark etc.) will be avoided.

Data sets will be versioned upon upload into the repository.

Metadata will be created for each sub-directory as detailed in Figure 1. Metadata will be created in a non-proprietary plain text file format (see Appendix A).

## Making data findable, including provisions for metadata

A hierarchical file structure, such as that shown in Figure 2, will be incorporated to separate data into meaningful sub-directories. Each sub-directory will contain metadata in a ‘readme’ text file, the ‘readme’ text files will be version controlled in an appropriate manner and contain metadata.



Figure 2 Example of hierarchical directory control and version-controlled metadata files.

### Making data openly accessible

All metadata produced during this project will be made openly available on various repositories (e.g., GitHub) using the appropriate creative-commons license. Publications will contain a data access statement where DOI links directly to the data and/or metadata records can be found. Metadata can be downloaded from the repository and accessed with any plain text editor or utility (e.g. sublime text, notepad etc. Publications are likely to be written using Microsoft Word (.doc, .docx etc.) or Latex (\*.tex) editors.

Software used to generate project data will be explicitly outlined in any publications and recorded in the relevant project metadata files. It is not possible to include all relevant software that will be used in this project as several programmes (e.g. Microsoft suite) are proprietary in nature. Additional software, e.g. jupyter notebooks, generated during the course of the project will be made accessible via GitHub and will be linked to the associated publication\s using DOI’s.

Data access will be managed by the researcher and the primary investigator. Ascertaining the identity of the person accessing project data will be dealt with on a case-by-case basis. An email contact point will be associated with any repositories. Users may request access to project data via the email contact point.

### Making data interoperable

Most data produced in this project will be interoperable via dataset specific metadata files. Datasets will be generated using comma separated values (‘.csv’) where possible. Additional repository platforms (e.g., GitHub) will be used to make analysis code open-access and/or restricted access as needed.

### Increase data re-use (through clarifying licences)

Data that is made openly available will be licensed using the creative commons (CC) licences. Data that is openly shared at the end of the project will remain accessible indefinitely via the appropriate repository.

# Other

All data management procedures issued by the research institute (University of Limerick) and the project funder (Science Foundation Ireland) will be adhered too.

# References

Anaconda, 2016. Anaconda Software Distribution.

1. Metadata template

The following template was modified from Cornell University’s website (<https://data.research.cornell.edu/content/writing-metadata>) and will be used. This template may be modified include additional information were appropriate.

This SUBDIRECTORY.txt file was generated on YYYYMMDD by NAME

GENERAL INFORMATION

1. Title of Dataset:

2. Author Information

A. Principal Investigator Contact Information

Name:

Institution:

Address:

Email:

B. Associate or Co-investigator Contact Information

Name:

Institution:

Address:

Email:

C. Alternate Contact Information

Name:

Institution:

Address:

Email:

3. Date of data collection (single date, range, approximate date) <suggested format YYYYMMDD>:

4. Geographic location of data collection <latitude, longitude, or city/region, State, Country, as appropriate>:

5. Information about funding sources that supported the collection of the data:

SHARING/ACCESS INFORMATION

1. Licenses/restrictions placed on the data:

2. Links to publications that cite or use the data:

3. Links to other publicly accessible locations of the data:

4. Links/relationships to ancillary data sets:

5. Was data derived from another source? yes/no

A. If yes, list source(s):

6. Recommended citation for this dataset:

DATA & FILE OVERVIEW

1. File List:

<list all files (or folders, as appropriate for dataset organization) contained in the dataset, with a brief description>

2. Relationship between files, if important:

3. Additional related data collected that was not included in the current data package:

4. Are there multiple versions of the dataset? yes/no

A. If yes, name of file(s) that was updated:

i. Why was the file updated?

ii. When was the file updated?

METHODOLOGICAL INFORMATION

1. Description of methods used for collection/generation of data:

<Include links or references to publications or other documentation containing experimental design or protocols used in data collection>

2. Methods for processing the data:

<describe how the submitted data were generated from the raw or collected data>

3. Instrument- or software-specific information needed to interpret the data:

<include full name and version of software, and any necessary packages or libraries needed to run scripts>

4. Standards and calibration information, if appropriate:

5. Environmental/experimental conditions:

6. Describe any quality-assurance procedures performed on the data:

7. People involved with sample collection, processing, analysis and/or submission:

DATA-SPECIFIC INFORMATION FOR: [FILENAME]

<repeat this section for each dataset, folder or file, as appropriate>

1. Number of variables:

2. Number of cases/rows:

3. Variable List:

<list variable name(s), description(s), unit(s)and value labels as appropriate for each>

4. Missing data codes:

<list code/symbol and definition>n

5. Specialized formats or other abbreviations used: