Data Management Plan

Data Collection

This project will use historic data from the **ONS Deaths registered monthly in England and Wales** Database (ONS, 2022) accessed from NHSRDatasets (Turner, 2020), a dataset with merged years in long format.

Primary data will also be collected on a ongoing basis. Administrators will complete a form with data recorded when deaths are certified and registered. A data collection tool has been created using the Jupyter Notebook. Data characteristics will be validated at source with required fields to maintain data quality. It's anticipated that this tool will collect data for at least 5-years[1].

New data will be created, including an index. The dataset will be divided into test and training sets. Scripts (.R and .ipynb), visuals (.png), reports (.Rmd, .pdf and .html) and data sets (.csv) will be saved in the project Github repository to make the project accessible and transparent. A licence will be chosen that requires attribution and limits commercial use.

Metadata

A data dictionary will provide information about the data collected. The data dictionary is saved as both a csv file (collected_data_dictionary.csv) and attached to the R dataset (complete_collected_data.rds). This helps any future data users to understand what data has been collected and why. A screenshot of the data dictionary can be seen in the appendix.

Ethics

This project will only deal with aggregated data that is routinely published publicly. This reduces concerns about sensitivity and security.

Storage and backup

Data management, data quality, and data security will be the responsibility of the project team. The data will be backed-up to a secure cloud-based environment at University of Edinburgh. It will be managed in line with the University's data privacy and security measures. The data will be maintained for the duration of the project and access will be limited to the project team. It is not anticipated it will exceed the allowable data storage limits.

Selection, preservation and sharing

The processed data set, analysis and outputs will be updated annually and preserved for at least 5 years following publication. The results of the analysis will be visualised and made available for NHS resource planners via interactive widget to enable the end users to engage with and extract further insights for operational decisions where possible.

---

[1] This approach has been proposed only to make a data collection tool a valid part of the project narrative. In a real world situation I would continue to use the ONS data as it is high quality (timeliness and completeness).