Plan Overview

A Data Management Plan created using DMPonline

Title: Understanding and Forecasting Socioeconomic Inequalities in Mortality in England with Interpretable AI

Creator: Rui Zhu

Principal Investigator: Rui Zhu

Affiliation: City St George's, University of London

Funder: Economic and Social Research Council (ESRC)

Template: ESRC Template

ORCID iD: 0000-0002-9944-0369

Project abstract:

The project develops an interpretable, deprivation-aware framework that treats socioeconomic deprivation as a core model dimension, alongside age, period and cause of death, in order to better understand the drivers of mortality inequality and their likely future evolution. Methodologically, it develops a deep tensor decomposition framework in which the underlying latent factors are estimated by neural networks, allowing flexible estimation with regularisation while retaining interpretability. The project also includes forecasting and scenario-based analysis to identify causes of death that may offer the greatest leverage for reducing future inequality.

ID: 203406

Start date: 01-09-2026

End date: 31-08-2028

Last modified: 01-05-2026

Understanding and Forecasting Socioeconomic Inequalities in Mortality in England with Interpretable AI

Assessment of existing data

Provide an explanation of the existing data sources that will be used by the research project, with references

The analysis will use two **publicly available** datasets from the **Office for National Statistics**, combined into a single modelling dataset for England. Together, they provide mortality, population and deprivation information stratified by age group, calendar year, cause of death and socioeconomic class measured by Index of Multiple Deprivation (IMD), covering 2001–2024.

Provide an analysis of the gaps identified between the currently available and required data for the research

No major data gaps have been identified. Existing publicly available datasets provide the information required for the research. The main work involves combining these sources into a single modelling dataset and carrying out the preprocessing needed for analysis, including alignment of stratifications across datasets and limited aggregation of sparse cause-of-death categories where necessary.

Information on new data

Provide information on the data that will be produced or accessed by the research project

The project will access publicly available Office for National Statistics datasets containing mortality, population and deprivation information. It will produce a combined modelling dataset in .csv formats derived from these sources through data cleaning, alignment by age group, calendar year, cause of death and deprivation group, and limited preprocessing to support analysis. No new primary data will be collected.

Quality assurance of data

Describe the procedures for quality assurance that will be carried out on the data collected at the time of data collection, data entry, digitisation and data checking.

As the project uses existing publicly available datasets, quality assurance will focus on data integration and preprocessing rather than primary data collection. This will include checking that downloaded files match the official source versions, verifying consistency and completeness of variables across datasets, and ensuring correct alignment by age group, calendar year, cause of death

and deprivation group. Additional checks will be carried out for missing values, duplicate records, inconsistent coding and mismatches in category definitions. Derived variables and any aggregated cause-of-death categories will be checked against the original source data, and descriptive summaries will be used to identify unexpected discrepancies or implausible values before analysis. All preprocessing and checking steps will be documented in reproducible code.

Backup and security of data

Describe the data security and backup procedures you will adopt to ensure the data and metadata are securely stored during the lifetime of the project.

The project uses publicly available data and does not involve personal or sensitive information. Data and metadata will be stored within City St George’s secure institutional research data infrastructure for the lifetime of the project, with access restricted to the project team through password-protected University systems. Derived datasets, code and documentation will be stored on secure institutional drives with routine backup provided through standard University procedures. Version-controlled code and documentation will also be maintained to support reproducibility and recovery. No additional enhanced security arrangements are required beyond standard institutional data security and backup procedures.

Management and curation of data

Outline your plans for preparing, organising and documenting data.

Data will be prepared by cleaning publicly available source files, aligning variables and stratifications across datasets, and creating any derived variables required for modelling. The combined dataset will be organised using clear file structures, naming conventions and version control. Documentation will record the original data sources, preprocessing steps, variable definitions, and any transformations or aggregation of sparse cause-of-death categories. Reproducible code will be used throughout to ensure that the workflow from source data to analytical dataset is transparent and reproducible.

Difficulties in data sharing and measures to overcome these

Identify any potential obstacles to sharing your data, explain which and the possible measures you can apply to overcome these.

No major obstacles to data sharing are anticipated, as the project uses publicly available data and does not involve personal or sensitive information.

Consent, anonymisation and strategies to enable further re-use of data

Make explicit mention of the planned procedures to handle consent for data sharing for data obtained from human participants, and/or how to anonymise data, to make sure that data can be made available and accessible for future scientific research.

This project does not involve the collection of data from human participants and will not use personal, confidential or otherwise sensitive information. The research relies entirely on publicly available data. As a result, no consent procedures for data sharing are required, and no anonymisation of participant data will be necessary. Any derived datasets, code and documentation produced by the project will be prepared for sharing in line with the terms of the original data sources and standard good practice in reproducible research.

Copyright and intellectual property ownership

State who will own the copyright and IPR of any new data that you will generate.

No new primary data will be generated. The project will produce derived data, code and documentation based on publicly available data. Copyright and intellectual property rights in these project-generated materials will be owned in accordance with City St George’s institutional policies, while the original source data will remain subject to the terms and conditions of the original data sources.

Responsibilities

Outline responsibilities for data management within research teams at all partner institutions

Data management will be the responsibility of the PI, who will oversee data organisation, documentation, storage and sharing throughout the project. The Research Assistant will support data preparation, file organisation, code documentation and version control under the PI’s supervision. As the project is based at City St George’s and does not involve formal project partners, no separate partner-institution data management responsibilities apply.

Preparation of data for sharing and archiving

Are the plans for preparing and documenting data for sharing and archiving with the UK Data Service appropriate?

Yes. The project will prepare and document the derived modelling dataset, code and associated materials in a way that supports future sharing and archiving where appropriate. Documentation will

include details of the original public data sources, preprocessing steps, variable definitions, and any transformations or aggregation applied during data preparation. Reproducible code will be used to ensure that the workflow from source data to analytical dataset is transparent and repeatable. In addition, derived data and code produced by the project will be shared through a public GitHub repository. As the research relies on publicly available data, any archiving or sharing through the UK Data Service will apply to the derived materials produced by the project rather than the original source data.

Is there evidence that data will be well documented during research to provide highquality contextual information and/or structured metadata for secondary users?

Yes. A combined modelling dataset will be created from publicly available data, and the project will document the original data sources, preprocessing steps, variable definitions, and any transformations or aggregation applied during data preparation. Clear file structures, naming conventions and reproducible code will be used throughout so that the workflow from source data to analytical dataset is transparent and reproducible. This will provide the contextual information and metadata needed to support secondary use.