Proper Dataset Documentation

Md Abu Sufian*

IVR Low-Carbon Research Institute, Chang'an University, Shaanxi, 710018, China

University of Leicester, Leicester, UK

March 28, 2024

Title

"AI-Enabled Study of Funding Cuts in the UK: Exploring Regional Mental Health Disparities through Machine Learning"

Executive Summary

This document encapsulates the comprehensive dataset underpinning the "AI-Enabled Study of Funding Cuts in the UK: Exploring Regional Mental Health Disparities" project. Sourced from the UK's Data.gov.uk and curated by the Department of Health and Social Care, this dataset embarks on a pivotal examination of how budgetary reductions impact mental health services across diverse UK regions. Utilizing advanced machine learning methodologies, the analysis presented herein delves into the intricate dynamics of service provision, patient care, and accessibility, elucidating the pronounced regional disparities engendered by fiscal austerity. The findings illuminate the stark realities and consequences of funding cuts, revealing critical insights into the disparities in healthcare delivery and outcomes. This exploration not only sheds light on the pressing challenges faced by mental health services but also underscores the urgent necessity for informed policy-making to safeguard and enhance patient care amidst financial constraints. This dataset documentation is designed to serve as an invaluable asset for a broad spectrum of stakeholders, including researchers, policy strategists, and healthcare practitioners, aiming to forge evidence-based solutions to fortify the UK's mental health service framework against the backdrop of economic austerity.

*Email: abusufian.tex.cu@gmail.com

Dataset Information

The dataset has been provided by the supervisor and is sourced from the open access platform Data.gov.uk, which is managed by the Department of Health and Social Care.

- **Dataset source:** Data.gov.uk
- **Department of Health and Social Care**

Supervisor Information

- Corresponding Authors (Supervisor): Mingbo Niu
- IVR Low-Carbon Research Institute, Chang'an University, Shaanxi, 710018, China
- Dept. of Signal Theory and Communications, University Carlos III of Madrid (UC3M), Leganes, 28911, Madrid, Spain

Versioning Information

- **Published by:** Department of Health and Social Care
- **Topic:** Government
- **Licence:** Open Government Licence

1 Data Collection Methodology:

The dataset for the "AI-Enabled Study of Funding Cuts in the UK: Exploring Regional Mental Health Disparities through Machine Learning" was sourced from data.gov.uk, a national repository for open data provided by the UK government. As this is a secondary data source, the collection methodology primarily involved the extraction of existing data made available by the Department of Health and Social Care.

The original data was compiled by the Department through a combination of administrative records, national surveys, and reporting systems from healthcare providers. The specific collection instruments and methods (e.g., electronic health records, patient surveys, financial reporting tools) are as per the standard protocols of the Department and designed to accurately reflect the healthcare services, operational metrics, funding allocations, staffing levels, and patient outcomes across various regions. While the dataset has been made publicly accessible for research and analysis, it is important to note that the granularity and scope of the data are determined by the original collection efforts of the governmental body. Therefore, any limitations in the primary data collection process, such as non-response in surveys or reporting biases in administrative records, may indirectly affect the secondary analysis conducted in this study.

For further details on the data collection instruments, methodologies, or processing steps undertaken by the Department of Health and Social Care, researchers are encouraged to refer to the corresponding documentation and metadata provided alongside the dataset on the data.gov.uk platform or to contact the Department directly for more comprehensive information.

2 Data Access and Licensing:

Any reader or research can access the dataset as it is open access free by the UK government. `https://www.data.gov.uk/searchfilters%5Bpublisher%5D=Department+of+Health+and+Social+Care`

DataFrame Index Information

The DataFramed referred to as `frame0-1`, includes the following variables with their respective data types:1

Table 1: Index and Data Types of DataFrame `frame0-1`

Variable	**Data Type**
Year	float64
Funding levels	float64
Staffing levels	object
Patient outcome	object
Emergency referral	float64
Discharges	float64
Associated emergency readmissions	float64
Urgent crisis referrals	float64
Access to care waiting time	float64
Geographic	object
Geographic-1	object
Gender	object
Age	float64
Indicator value patient satisfaction	object
Indicator value Policy changes	int64
Region	object

Data Types in DataFrame

Data types of the variables in the dataset in Table 2

Table 2: Data types of the variables in the dataset

Variable	**Data Type**
Year	float64
Funding levels	float64
Staffing levels	object
Patient outcome	object
Emergency referral	float64
Discharges	float64
Associated emergency readmissions	float64
Urgent crisis referrals	float64
Access to care waiting time	float64
Geographic	object
Geographic-1	object
Gender	object
Age	float64
Indicator value patient satisfaction	object
Indicator value Policy changes	int64
Region	object

Dataset Missing Values Analysis

After data cleaning, the following table represents the variables in the dataset with their respective count of missing values: Dataset missing values analysis in Table 3

Note: The data type for all variables has been verified as 'int64', confirming no missing values post-handling.

Synthetic DataFrame Index

The following table lists the variables of the synthetic DataFrame, referred to as `synthetic_frame`: Index of the synthetic DataFrame in Table 4.

Combined DataFrame Index

The combined DataFrame, denoted as `combined_frame`, consists of the following indices: Index variables of the combined DataFrame in Table 5.

3 Descriptive Analysis of Combined Data frame

Descriptive Analysis of Combined Data frame in Table 6

Table 3: Count of missing values per variable in the dataset.

Variable	**Missing Values**
Year	0
Funding levels	0
Staffing levels	0
Patient outcome	0
Emergency referral	0
Discharges	0
Associated emergency readmissions	0
Urgent crisis referrals	0
Access to care waiting time	0
Geographic	0
Geographic-1	0
Gender	0
Age	0
Indicator value patient satisfaction	0
Indicator value Policy changes	0
Region	0

Table 4: Index of the synthetic DataFrame.

Variables
num_professionals_pre_cuts
num_professionals_post_cuts
Waiting_time_pre_cuts
waiting_time_post_cuts
service_accessibility_pre_cuts
service_accessibility_post_cuts
hospitalization_rate_pre_cuts
hospitalization_rate_post_cuts
readmission_rate_pre_cuts
readmission_rate_post_cuts
patient_satisfaction_pre_cuts
patient_satisfaction_post_cuts
service_utilization_pre_cuts

4 Data Limitations and Assumptions:

The dataset may contain inherent limitations due to the nature of secondary data aggregation. Assumptions include the reliability and completeness of data as submitted by healthcare providers to the Department of Health and Social Care. Discrepancies in reporting standards and missing data are acknowledged as potential constraints on the analysis. We assume consistent data collection methodologies across different time periods and regions,

Table 5: Index variables of the combined DataFrame.

Index Variables
Year
Funding levels
Staffing levels
Patient outcome
Emergency referral
Discharges
Associated emergency readmissions
Urgent crisis referrals
Access to care waiting time
Geographic
Geographic-1
Gender
Age
Indicator value patient satisfaction
Indicator value Policy changes
Region
Num professionals pre cuts
Num professionals post cuts
Waiting time pre cuts
Waiting time post cuts
Service accessibility pre cuts
Service accessibility post cuts
Hospitalization rate pre cuts
Hospitalization rate post cuts
Readmission rate pre cuts
Readmission rate post cuts
Patient satisfaction pre cuts
Patient satisfaction post cuts
Service utilization pre cuts

although variations may exist.

5 Privacy and Ethical Considerations

The dataset was sourced from data.gov.uk, where data is presumed to be anonymized and stripped of any personally identifiable information in compliance with GDPR and other privacy legislation. Ethical considerations in the analysis of healthcare service data were strictly observed, ensuring that no private patient information could be discerned and that the data use complies with all ethical guidelines for secondary data analysis.

Table 6: Descriptive Analysis of Combined Data frame

Category	Variable	Count	Mean	Std Dev
Demographics	Age	454	28.82	3.51
Resources	Year	454	1972.5	13.56
	Funding Levels	454	7.41	4.35
	Staffing Levels	454	494.23	315.13
Outcomes	Patient Outcome	454	2087.84	9876.2
	Emergency Referral	454	35.82	215.64
	Discharges	454	21947.58	5978.86
	Associated Emergency Readmissions	454	836.31	228.28
	Urgent Crisis Referrals	454	443.88	2494.94
Accessibility	Access to Care Waiting Time	454	32.94	19.49
	Waiting Time Post Cuts	454	30.16	5.25
	Service Accessibility Pre Cuts	454	0.8	0.06
	Service Accessibility Post Cuts	454	0.62	0.09
Rates	Hospitalization Rate Pre Cuts	454	0.29	0.06
	Hospitalization Rate Post Cuts	454	0.38	0.05
	Readmission Rate Pre Cuts	454	0.29	0.03
	Readmission Rate Post Cuts	454	0.4	0.04
Satisfaction	Patient Satisfaction Pre Cuts	454	7.12	0.58
	Patient Satisfaction Post Cuts	454	8.48	0.4

6 Code Availability

The analysis was conducted using scripts developed in Python, utilizing libraries such as pandas, numpy, and scikit-learn for data manipulation and machine learning. While the specific codebase is proprietary to the research team at this stage, an overview of the algorithms and processes is available in the given github repository at the manuscript.

7 Use Cases

The dataset has been employed to analyze the impact of funding cuts on mental health services across regions. This has potential applications in policy analysis, resource allocation, and healthcare services research. The insights can inform governmental and organizational strategies to mitigate adverse effects and promote mental health welfare.

8 Change Log

Documentation and dataset versioning are crucial for maintaining the integrity and traceability of the data analysis process. A change log has maintained, documenting all updates to the dataset, revisions to the methodology, and alterations in the analysis scripts, providing a clear record for users to track modifications over time.

For dataset details, see Md Abu Sufian (2024) on UK funding cuts and mental health: [1].

Data Quality Assessment Results for Dataset Validation

Completeness

All columns have 0 missing values, indicating no missing data after preprocessing.

Consistency

No duplicate rows found. However, the 'staffing levels' and 'Patient outcome' columns show a wide range of unique values, indicating a need for standardization.

Accuracy

Column	Number of Inaccurate Entries
Age	0
Funding levels	0

Validity

Data types are correctly assigned to all columns post-processing.

Uniformity

Column	Standard Deviation
Year	15.36
Funding levels	4.95
Emergency referall	244.29
Discharges	6971.50
Associated emergency readmissions	260.27
Urgent crisis referalls	2826.38
Access_to_care_waiting_time	22.43
Age	4.26
Indicator_value_Policy_changes	5.07

Timeliness

Timeliness: 'Latest Year': 2022.0, 'Oldest Year': 1923.0

[1] https://github.com/datascintist-abusufian/AI-Enabled-Study-of-Funding-Cuts-in-the-UK-Exploring-Regi
blob/main/Proper_Dataset_Documentation_NHS_funding_cut.pdf

Outlier Detection

Column	Number of Outliers
Year	8
Funding levels	8
Emergency referall	2
Discharges	10
Associated emergency readmissions	7
Urgent crisis referalls	1
Access_to_care_waiting_time	9
Age	8
Indicator_value_Policy_changes	9

- Completeness: Confirms there are no missing values across all columns after preprocessing.
- Consistency: Highlights the need for further investigation into ’staffing levels’ and ’Patient outcome’ due to a wide range of values.
- Accuracy: Confirms there are no inaccurate entries in ’Age’ and ’Funding levels’.
- Validity: Indicates data types are appropriately assigned.
- Uniformity: Shows the standard deviation for each numeric column, giving an idea of the spread of values.
- Timeliness: Notes that the latest data point is from the year 2022.
- Outlier Detection: Lists the number of outliers detected in various columns, suggesting areas where data may deviate significantly from the norm.

9 Conclusion

The research project has been meticulously documented and validated through rigorous statistical methods. This ensures its reliability for any reader in-depth analyses further and makes it a valuable asset for stakeholders aiming to address the nuanced impacts of funding adjustments on mental health services.