Scientific Reproducibility: Challenges and Solutions in NEMSIS Data Analysis

A study based on the NEMSIS Dataset (NHTSA v3.5.0, 2022 data) and the article by Peters et al. (2022) "Differences in Out-of-Hospital Cardiac Arrest Management and Outcomes across Urban, Suburban, and Rural Settings."

Darby, Duggan, Jordan
DS 5110
Final Project

Research Using NEMSIS Data – Methodology

1: Acquire Data

2: Unpack Data

3: Create a Database

4: Develop Domain Knowledge

5: Define Features

6: Analyze Features

7: Report on Findings

Welcome to the NEMSIS Help Desk

Search for information

Welcome! You can submit a request for NEMSIS Help Desk using the options provided.

What can we help you with?

Request a Research Dataset
Request the public release research dataset.

Software Vendors
Vendor user accounts and general request for support for software vendors.
For compliance testing questions, please go to the compliance page at NEMSIS.org.

State Offices of EMS
Submit state resources, questions about user accounts, and general support for state EMS offices.

EMS Agency or Clinician
Request a custom data query and general support questions for EMS agencies.

Researchers
Requests for datasets and research queries, problems with dashboards or cubes, and general support questions for researchers.
For dataset requests, please go to NEMSIS.org.

Casey Gregor to Ian Duggan 28 Oct, 11:02 AM

Reply above this line.

Casey Gregor commented:

Below is the link for the 2022 NEMSIS Public-Release Dataset in ASCII format as requested. The link will expire in 3 business days. Thank you for your request. Please feel free to reach out if you have any further questions.

https://nemsis-prod-publicuse-datasets.s3.us-west-2.amazonaws.com/ASCII 2022.zip?AWSAccessKeyId=AKIAXWSRKUO24OVVL5GC&Expires=1730473284&Signature=C2ogUmZX2%2Frds27FutEaOUcsWEc%3D

Respectfully,

Casey F. Gregor, NRP

Business Data Analyst | NEMSIS Technical Assistance Center

University of Utah School of Medicine

Department of Pediatrics | Division of Critical Care

System & Directory Requirements

Request a copy of the NEMSIS Public-Release Research Dataset for 2022 from NEMSIS in ASCII (pipe-delimited) format. You can currently find the request form here. When you receive the file, verify that it has the following characteristics:

- filename: "ASCII 2022.zip"
- size: 18,021,941,320 bytes
- sha256 hash: 2fc87b18edf2e762be2d723c44413cb98064bcbc3e6a46468e3b42f25f521898 **Note:** if analysis is extended to a different year, file and data properties may vary.

Use a computer with at least 300 GB of available storage to hold the dataset. This work was done on a computer with 32 GB of RAM; less RAM may work, but is not advised as some analysis files if written inefficiently (multiple pandas dfs), may exceed 16 GB of RAM.

Discovery Account Access Guide

This guide is intended to assist subsequent students working on the NEMSIS 911 project with requesting access to a Discovery Account on Northeastern University's Remote Computing Platform. Follow the steps below to obtain access and set up your account.

Instructions for Requesting Access to the Discovery Remote Computing Cluster

Navigate to the Remote Computing Platform Website

- Click the following link to go to the platform: Remote Computing Platform Access Page.
- You will see a page describing the different access options for Northeastern's computing resources.
- Click on "Request Access to the Cluster".
- Enter your Northeastern credentials when prompted.

Overview of Original Study

- Study Name: "Differences in Out-of-Hospital Cardiac Arrest Management and Outcomes across Urban, Suburban, and Rural Settings" by Peters et al. (2022).
- Data Used: 2018 NEMSIS Dataset
- Primary Outcomes:
  - ROSC (Return of Spontaneous Circulation)
- Secondary Outcome
  - Binary indicator of survival at the end of the event.
- Methods: Multivariable logistic regression analysis
- Key Findings: Differences between urban, suburban, and rural OHCA management and outcomes.

Challenges in Reproducibility

- Complexity of the Dataset: NEMSIS data is large, multi-faceted, and requires precise filtering to extract relevant incidents.
- Ambiguity in Methodology: Lack of precise information on data filtering and analysis steps used in the original study, limits our ability to match the analysis.
  - Difficulty in ensuring whether all criteria used by Peters et al. were captured correctly (e.g., filtering out pediatric cases, traumatic injuries, and using precise outcomes).
- "Reproducibility is about transparency — making sure that the way analyses are conducted is documented clearly enough so that others can replicate the results" (Laurinavichyute et al. 2022).

Reproducibility Problems – A Recurring Theme

- In 2019 the Journal of Memory and Language instituted a case study of the reproducibility of articles published in the Journal of Memory and Language under the open data policy
- https://www.sciencedirect.com/science/article/pii/S0749596X22000195

- Under a strict reproducibility criterion, 20 (34%) papers were reproducible.
- Under a lenient criterion, 33 (56%) papers were reproducible.
- Reproducibility rate rose by almost 40% when the analysis code was provided.

Our Approach to Data

- **With this much data, using pandas as the tool for manipulating data was inefficient.**

  "Even datasets that are a sizable fraction of memory become unwieldy, as some pandas operations need to make intermediate copies." (Pandas Docs.)

- An SQlite database was constructed for more efficient querying of the data

```python
import sqlite3
from constants import table_definitions as tbl_defs
from constants import paths


"""
This script creates the data/NEMSIS_PUB.db sqlite database to store
data for further analysis.

See constants/table_definitions.py for the SQL used to create each table
"""
def main():
    # create sqlite .db object to store database
    conn = sqlite3.connect(paths.db_path)
    cursor = conn.cursor()

    # iterate over each item in the table_definitions, create table
    for table_name, create_table_sql in tbl_defs.table_definitions.items():
        cursor.execute(create_table_sql)

    # commit changes then close connection
    conn.commit()
    conn.close()

if __name__ == "__main__":
    main()
```

A constant set of table definitions are looped over to construct the SQlite database.

[Pandas]. 2024. *Scaling to large datasets.* Pandas Docs. https://pandas.pydata.org/docs/user_guide/scale.html

load_data_NEMSIS_db.py

- **Dynamic Table Creation and Loading**:
  - The script reads column names directly from the dataset, dynamically creating SQL tables to match the data structure.
  - It then iterates through the data and loads it row by row, making the process adaptable to different files with varying columns.
- **Handling Errors and Mismatches**:
  - The script includes error handling to skip rows with mismatched columns or missing data. This ensures the integrity of the loaded data and avoids unexpected issues during analysis.
- **Primary Key Management**:
  - A primary key, often PcrKey or another specified key, is assigned for each table. This is crucial for maintaining the uniqueness of entries, which is essential for reliable and consistent data retrieval.
- **Why This Matters for Reproducibility**:
  - A major goal of our project is reproducibility, which means future analysts need to be able to repeat our process accurately.
  - This script not only documents how the data is processed but also automates the setup of the database, ensuring that others can achieve the same data setup without manual intervention.
  - By incorporating flexibility, such as dynamic column assignment and error handling, this script provides a robust foundation for reproducibility.

Developing Domain Knowledge: Helper Scripts

- **query.py**

query the database, and save top-n rows of output to txt file

- **nemsis_text_format.py & nemsis_find_cols.py**

save text files with the head of each larger table, and find the source table given a list of column names

```
Query:
'''
    select distinct m.eMedications_03, m.eMedications_03Descr from FACTPCRMEDICATION m
    where m.eMedications_03Descr like '%epinephrine%'
    order by m.eMedications_03

'''

|  eMedications_03  |  eMedications_03Descr  |
|  1010751  |  Epinephrine 0.01 MG/ML / Lidocaine Hydrochloride 10 MG/ML Injectable Solution  |
|  1010759  |  EPINEPHrine 0.01 MG/ML / Lidocaine Hydrochloride 20 MG/ML Injectable Solution  |
|  1010767  |  Epinephrine / Lidocaine Injectable Solution [Xylocaine with Epinephrine]  |
|  1011648  |  Articadent 4 % with Epinephrine 1:100,000 Injectable Solution  |
|  1011809  |  Lignospan 2/1:100000 (lidocaine hydrochloride / epinephrine (as epinephrine bitar
|  1012792  |  Epinephrine 0.01 MG/ML / Mepivacaine Hydrochloride 20 MG/ML Injectable Solution
|  107602  |  Epinephrine / Lidocaine  |
|  107606  |  Bupivacaine / Epinephrine  |
|  1100194  |  10 ML EPINEPHrine 0.016 MG/ML Prefilled Syringe  |
|  1100200  |  50 ML Norepinephrine 0.016 MG/ML Prefilled Syringe  |
|  1150120  |  1 ML Epinephrine 0.01 MG/ML / Lidocaine Hydrochloride 10 MG/ML Prefilled Syringe
|  1150987  |  epinephrine 32 MCG/ML Injectable Solution  |
|  1163887  |  Epinephrine Injectable Product  |
|  1233582  |  Bupivacaine / Epinephrine / Fentanyl  |
|  1233778  |  10 ML EPINEPHrine 0.01 MG/ML Prefilled Syringe  |
```

A query for - "what kinds of medications show up in the data with descriptions like '%epinephrine%' ?"

Developing Domain Knowledge: Cont.

reports > query_results > Medications_Epinephrine_Dosage_Units.txt

```
Query:
'''
    select distinct m.eMedications_06 from FACTPCRMEDICATION m
    where m.eMedications_03 in (3992, 310116, 310132,
    317361, 328314, 328316, 330545, 372030, 377281,
    727316, 727373, 727374, 727386, 1100194, 1233778, 1305268)
    order by m.eMedications_06

'''

|  eMedications_06  |
|  3706001  | -- Grams (gms)
|  3706003  | -- Inches (in)
|  3706005  | -- International Units (IU)
|  3706007  | -- Keep Vein Open (kvo)
|  3706009  | -- Liters (l)
|  3706011  | -- may be deprecated, liters / minute (l/min fluid)
|  3706013  | -- Metered Dose (MDI)
|  3706015  | -- Micrograms (mcg)
|  3706017  | -- Micrograms per Kilogram per Minute (mcg/kg/min)
|  3706019  | -- Milliequivalents (mEq)
|  3706021  | -- Milligrams (mg)
|  3706023  | -- Milligrams per Kilogram Per Minute (mg/kg/min)
|  3706025  | -- Milliliters (ml)
```

A query for - "of the medications that may be related to epinephrine, what kinds of dosage units are reported?"

Future Research Question: *To analyze epinephrine treatment across different cases. Can/should a measure be created to standardize these units for easier comparison?*

Data Approach to Reproducibility: Constants Module

```
81        "FACTPCRARRESTROSC": """
82            CREATE TABLE IF NOT EXISTS FACTPCRARRESTROSC (
83                PcrKey TEXT PRIMARY KEY,
84                eArrest_12 TEXT,
85                FOREIGN KEY (PcrKey) REFERENCES Pub_PCRevents(PcrKey)
86            )
87        """,
88        "FACTPCRARRESTRESUSCITATION": """
89            CREATE TABLE IF NOT EXISTS FACTPCRARRESTRESUSCITATION (
90                PcrKey TEXT PRIMARY KEY,
91                eArrest_03 TEXT,
92                FOREIGN KEY (PcrKey) REFERENCES Pub_PCRevents(PcrKey)
93            )
94        """,
```

Table definitions for SQL, Keys, etc…

```
file_column_mapping = {
    'ComputedElements.txt': "PcrKey'~|~'USCensusRegion'~|~'USCensusDivision'~|~'
    'FACTPCRARRESTROSC.txt': "PcrKey'~|~'eArrest_12'",
    'FACTPCRARRESTRESUSCITATION.txt': "PcrKey'~|~'eArrest_03'",
    'FACTPCRARRESTWITNESS.txt': "PcrKey'~|~'eArrest_04'",
    'FACTPCRARRESTCPRPROVIDED.txt': "PcrKey'~|~'eArrest_09'",
    'FACTPCRMEDICATION.txt': "eMedications_01'~|~'PcrMedicationKey'~|~'PcrKey'~|
    'Pub_PCRevents.txt': "PcrKey'~|~'eDispatch_01'~|~'eDispatch_02'~|~'eArrest_1
}
```

Mapping of columns for tables of interest…

```
"""
eArrest_03 - Indication of an attempt to resuscitate the patient who is in
             cardiac arrest (attempted, not attempted due to DNR, etc.).

Rationale: other codes are excluded as they are cases where resuscitation
             was not attempted.

Included:
3003001 - Attempted Defibrillation
3003003 - Attempted Ventilation
3003005 - Initiated Chest Compressions
"""
eArrest_03_codes = [3003001, 3003003, 3003005]


"""
eResponse_05 - The type of service or category of service requested of
             the EMS Agency responding for this specific EMS event.

Rationale: response codes included, other codes not for scene response.

Included:
2205001 - Emergency Response (Primary Response Area)
2205003 - Emergency Response (Intercept)
2205009 - Emergency Response (Mutual Aid)
"""
eResponse_05_codes = [2205001, 2205003, 2205009]
```

Filter criteria for selecting relevant cardiac arrest events…

- How are features defined by Tables? Fields? Codes?

- Availability of Public Vs Privacy Restricted Elements of the NEMSIS Dataset

- "It was not clear from the available columns that a useful model could be built for any audience using the provided data." - Aaron Finn README.md

- How thing are filtered, should we - Include "Yes" or Exclude "No" ?

- Availability of State Vs National Elements of the NEMSIS Dataset

To REPRODUCE or not to REPRODUCE

- How are the txt files containing the tables parsed?

- we've found that some of the key OHCA variables seem to be very different in 2023 than prior years." - Author, Dr Cash

- How are unexpected values handled?

- Are missing values handled consistently? - Null vs " " (An Empty String)

So Many Uncertainties!

How Do We Move Forward?

Download and Read These Papers from our Repo

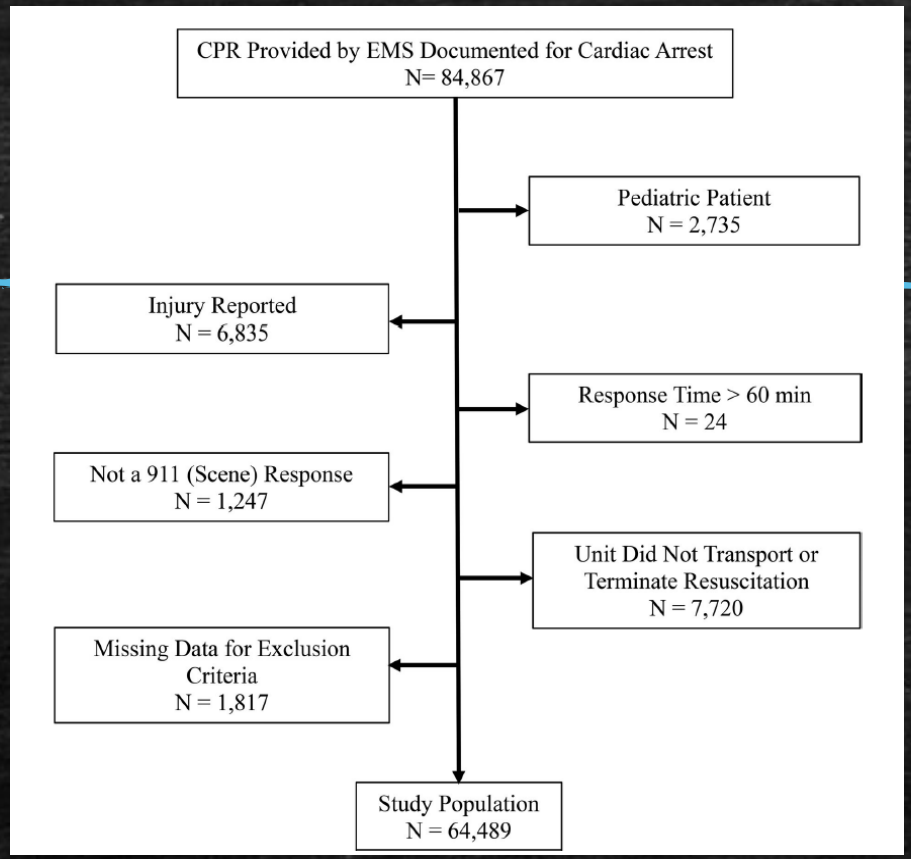
The Reproducibility Roadblock: Unmasking the Challenges of Reproduction

The Feature Definition Spreadsheet for NEMSIS-Data Based Research and Beyond

THE FEATURE DEFINITION SPREADSHEET

- What is it?
  - A Group of Spreadsheets Used Together to Organize and Define Features and Filters for a Project that are being Programatically Defined by Code from Rows in a Database
  - Prioritizes Reproducibility

- How do I use it?
  - Pick a Project to Reproduce or Create a New Project
    - Identify all of the features and filters required (Based on Prior Work or Domain Knowledge)
    - Load all of the reusable entries into the "validation_sheet"
    - Commit a group of lines in the features_sheet to represent a feature
      - Each line will tie to specific text or codes in the target field (A column in a related table)
      - Populate the line using the validated entried from the Dropdown List
      - Tie the code to the filter or feature group for clarity and reproducibility



PICK A PROJECT TO REPRODUCE

DIFFERENCES IN OUT-OF-HOSPITAL CARDIAC ARREST MANAGEMENT AND OUTCOMES ACROSS URBAN, SUBURBAN, AND RURAL SETTINGS

Feature Names

And Filters

To Reproduce

Prehospital Emergency Care Paper Reproduction				
Attribute Description	Confirmed by	Table in NEMSIS Data	Column Name	Code List Definiton
Cardiac Arrest	911 Team Member - Ben Darby	Pub_PCRevents.txt	eArrest_01	Cardiac Arrest Code List
CPR by EMS	911 Team Member - Ian Duggan	FACTPCRARRESTRESUSCITATION	eArrest_03	CPR by EMS Code List
Page age < 18 year Old?	911 Team Member - Tristan Jordan	ComputedElements.txt	ageinyear	N/A, No code, value listed is age to be imported as INT datatype
Had the Patient Sustained A Traumatic Injury?	Stakeholder - Qingchu Jin, PhD	Pub_PCRevents.txt	eArrest_02	Cardiac Arrest Etiology Code List
Transport not for 911	Emergency Care Paper Author - Rebecca Cash	Pub_PCRevents.txt	eResponse_05	Not A 911 Response Code List
Multiple unit check	Emergency Care Paper Author - Rebecca Cash	Pub_PCRevents.txt	eDisposition_12	No Transport or Terminate Codes List

THE FEATURE DEFINITION SPREADSHEET

Organizes Feature Name

Defines Researcher

Defines Data Source + Definition

Prehospital Emergency Care Paper Reproduction		
Data Definition Sources	**Data Definition Description**	**URL**
Nemsis Standard Data Dictionary	Detailed code defitions for every NEMSIS standard database field in all tables	https://nemsis.org/me
Nemsis Public Research Dataset	Details the fields available with the NEMSIS dataset in the PUBLIC dataset tables	https://nemsis.org/wp
Nemsis Code Lists	Defines NEMSIS code to name relationships for fields with many codes	https://nemsis.org/tec
Prehospital Emergency Care Paper	Contains findings for Cardiac Arrest Patient Filters and Characteristics	https://doi.org/10.108(

THE DATA_SOURCES_SHEET

Tables

Fields

Codes

Prehospital Emergency Care Paper Reproduction				
Methodology Confirmed?	**Confirmed by**	**Title from Prehospital Emergency Care Paper**	**Table in NEMSIS Data**	**Column Name**
--	--	--	--	--
Yes	911 Team Member - Ben Darby	Advanced airway management - endotracheal intub	ComputedElements.txt	ageinyear
No	911 Team Member - Ian Duggan	Advanced airway management - failed attempt	FACTPCRARRESTPROVIDED	eArrest_01
Pending	911 Team Member - Tristan Jordan	Advanced airway management - none	FACTPCRARRESTRESUSCITATION	eArrest_02
N/A	Emergency Care Paper Author - Rebecca Cash	Advanced airway management - supraglottic airway	FACTPCRARRESTROSC.txt	eArrest_03
	Nemsis Staff - Casey Gregor	Age < 18	FACTPCRARRESTWITNESS.txt	eArrest_04
	Stakeholder - Philip Bogden, Professor	Age in years, mean	FACTPCRMEDICATION.txt	eArrest_05
	Stakeholder - Qingchu Jin, PhD	ALS Level of Care	Pub_PCRevents.txt	eArrest_07
	Stakeholder - Teresa May, DO	Arrest witnessed by EMS		eArrest_09
	Stakeholder- Christine Lary, PhD	CPR Performed by EMS for Cardiac Arrest		eArrest_11
	Undefined	CPR prior to EMS arrival		eArrest_12

THE VALIDATION_SHEET

Define Valid Entries

By Column

For Use in Features_Sheet

Prehospital Emergency Care Paper Reproduction		
Characteristic Identifier	Code	Label
c7	2215001	BLS-First Responder/EMR
c7	2215003	BLS-Basic /EMT
c7	2215005	BLS-AEMT
c7	2215007	BLS-Intermediate
c7	2215023	BLS-Community Paramedicine
c7	2215009	ALS-AEMT
c7	2215011	ALS-Intermediate
c7	2215013	ALS-Paramedic
c7	2215015	ALS-Community Paramedicine
c7	2215017	ALS-Nurse
c7	2215019	ALS-Physician
c7	2215021	Specialty Critical Care

THE CODE_SHEET

Named Ranges

Of Codes

With Links to Feature

Prehospital Emergency Care Paper Reproduction

Attribute Identifier	Methodology Confirmed?	Title from Prehospital Emergency Care Paper	Attribute Description	Confirmed by	Table in NEMSIS Data	Column Name	Code List Definiton	Include/Exclude	Primary Data Source ID	Secondary Data Source ID	Code Script Name	Code Cells Range	Notes
f1	Yes	Event Type: Cardiac Arrest	Cardiac Arrest	911 Team Member - Ben Darby	Pub_PCRevents.txt	eArrest_01	Cardiac Arrest Code List	Include if True	Nemsis Standard Data Dictionary	--	filter_primary_NEMSIS_cases.py	Rows: 107 - 126	None
f2	Yes	CPR Performed by EMS for Cardiac Arrest	CPR by EMS	911 Team Member - Ian Duggan	FACTPCRARRESTRESUSCITATION	eArrest_03	CPR by EMS Code List	Include if True	Nemsis Standard Data Dictionary	--			None
f3	Yes	Age < 18	Page age < 18 year Old?	911 Team Member - Tristan Jordan	ComputedElements.txt	ageinyear	N/A, No code, value listed is age to be imported as INT datatype	Exclude if True	Nemsis Public Research Dataset	--			None
f4	Pending	Injury Reported	Had the Patient Sustained A Traumatic Injury?	Stakeholder - Qingchu Jin, PhD	Pub_PCRevents.txt	eArrest_02	Cardiac Arrest Etiology Code List	Include if True	Nemsis Standard Data Dictionary	--			Exclude if code present: 3002015 Trauma / mssg stakeholder about other exclusion conditions / eDispatch_01 = 2301073. I was able to remove only about 400 records from email of Yue Huang
f5	Pending	Not a 911 (scene) respone	Transport not for 911	Emergency Care Paper Author - Rebecca Cash	Pub_PCRevents.txt	eResponse_05	Not A 911 Response Code List	Include if True	Nemsis Standard Data Dictionary	--			Email forwarded by stakeholder for Author with details, clarification required
f6	Yes	Unit Did Not Transport or Terminate Resuscitation	Multiple unit check	Emergency Care Paper Author - Rebecca Cash	Pub_PCRevents.txt	eDisposition_12	No Transport or Terminate Codes List	Exclude if True	Nemsis Standard Data Dictionary	--			Confirmed by email from Author
f7	Yes	EMS Response Time > 60 Minutes	Chute Time	911 Team Member - Tristan Jordan	ComputedElements.txt	EMSChuteTimeMin	N/A, No code, value listed is time differential	Mean by Value	Nemsis Public Research Dataset	--			Calculate as eTimes_06 (arrival on scene) - eTimes_01 (initial dispatch time) // or from COMPUTEDELEMENTS (EMSDispatchCenterTimeSec + EMSSystemResponseTimeMin) Public.P.25
			Time from the 911 call to the patient scene	911 Team Member - Tristan Jordan	ComputedElements.txt	EMSSystemResponseTimeMin	N/A, No code, value listed is time differential		Nemsis Public Research Dataset	--			
f8	Yes	Missing data for exclusion criteria	Filter Column Missing Data	911 Team Member - Ben Darby	--	--	N/A, search for N/A values and empty strings	Special Rule	None	--			Exclude if any of the Incident Condition fields (cell A2) are missing or contain the NA values of 7701001 and 7701003
p1	Yes	Urbanicity	Urbanicity	911 Team Member - Tristan Jordan	ComputedElements.txt	Urbanicity	Urbanicity Code List	Sum by Value	Nemsis Public Research Dataset	--			Rural and wilderness are grouped as rural

Find the Link in Our Github Repo – Project-Duggani

THE

FEATURES_SHEET

Organized

Documented

Reproducible

MOVING FORWARD

- The Feature Definition Spreadsheet Serves as a Central Repository
  - Ensures Reproducibility
  - Allows Users to Create New or Define Existing Features from Prior Findings
  - Sets the stage for the next 911 group to move forward with step 6 (Analyize the data) and step 7 (Report on Findings)
- Clear Documentation:
  - Maintaining a clear README, code comments, and table definitions (e.g., table_definitions.py) to ensure transparency.
- Share the Code:
  - By documenting our code thoroughly and storing everything in a version-controlled repository (Github), we are making our work available for further scrutiny and continuation.

Citations

- G. A. Peters, A. J. Ordoobadi, A. R. Panchal, and R. E. Cash, "Differences in out-of-hospital cardiac arrest management and outcomes across urban, suburban, and rural settings," *Prehospital Emergency Care*, pp. 1–11, Dec. 2021, doi: https://doi.org/10.1080/10903127.2021.2018076.
- A. Laurinavichyute, H. Yadav, and S. Vasishth, "Share the code, not just the data: A case study of the reproducibility of articles published in the Journal of Memory and Language under the open data policy," *Journal of Memory and Language*, vol. 125, p. 104332, Aug. 2022, doi: https://doi.org/10.1016/j.jml.2022.104332.
- F. Prinz, T. Schlange, and K. Asadullah, "Believe it or not: how much can we rely on published data on potential drug targets?," Nature Reviews Drug Discovery, vol. 10, no. 9, pp. 712–712, Aug. 2011, doi: https://doi.org/10.1038/nrd3439-c1.