NEMSIS & Cardiac Arrest Survivability

Aaron Fihn
December 6, 2023

Dataset Description

- **NEMSIS**: National Emergency Medical Services Information System
- NHTSA program collecting EMS data from US states
- Data released for research requests
- 51 MM de-identified patient care reports (PCRs) from EMS activations in 2022

"Information documented by the EMS clinician about the patient encounter. It identifies the type of injury, scene of injury, medications administered, procedures performed, and more."

https://nemsis.org/wp-content/uploads/2022/11/NEMSIS-Guide-to-Data-Sets-1.pdf

Original Project Goals

1. Create a **reproducible** data pipeline to filter records to only cardiac arrests
2. Analyze data to find effects of urban/rural environment on cardiac arrest survival
3. Model cardiac arrest survivability

Data Issues

- Too large to hold in-memory (> 100 GB)
- No patient identifiers
- Convenience sample
  - States have different criteria for data submission
- No geographic identifiers
- Missing data
  - *“In most cases, NEMSIS data are not missing at random and analyses, therefore, are subject to bias if missing data are ignored.”*
- Data Quality
  - “Checked for completeness, logical consistency, and proper formatting”
  - Problematic data is flagged for state review
  - If uncorrected, **files are included anyway**

https://nemsis.org/wp-content/uploads/2023/04/2022-NEMSIS-RDS-340-User-Manual_v1-FINAL.pdf

Goal 1: Reproducible Data Pipeline

Clone repo:

```
git clone https://github.com/ds5500/scope-aaronfihn
```

Create the Conda environment and activate it:

```
conda env create -f environment.yml
conda activate 911-NEMSIS
```

Drop the zipped NEMSIS dataset into /data/raw and run the makefile:

```
make
```

What does the Makefile do?

1. zip utility: corrupted ZIP -> repaired ZIP
2. zip utility: repaired ZIP -> 42 unfiltered text files (175 GB)
3. Python script: Filter key tables to CSVs
4. Python script: Filter entire dataset down to cardiac arrests and save as pandas pickle file
5. Perform analysis and save results

Cardiac arrests only: 51MM → 291K records

Urbanicity and outcome both present: 291K → 184K records

Goal 2: Urbanicity affects cardiac arrest survival?

Overall survival rate about 19%

- Literature gives 10% - 25%

Small differences in survival rate:

- Urban/Suburban: 19.0%
- Rural/Wilderness: 17.8%

Chi-square independence test:

- Difference is *not* caused by random chance
- $p < 0.001$

	Alive	Dead	**Total**
Urban/Suburban	31,797	135,368	**167,165**
Rural/Wilderness	3,049	14,090	**17,139**
Total	**34,846**	**149,458**	**184,304**

Goal 3: Modeling cardiac survivability

Who is this model for?

What are they doing with it?

A useful model:

- Only uses features known to the user
- Helps make actionable decisions

Goal 3: Rejected models

First responders: Survivability by hospital capability

- They can't choose to make a more capable hospital exist
- Selection effects: worse patients taken to more capable hospitals?

Hospital administrators: Identify frequent or at-risk patients?

- No patient identifiers included

Policymakers: Survivability by city/county?

- No geographic identifiers included