DRIVENDATA

HEALTH

Unsupervised Wisdom: Explore Medical Narratives on Older Adult Falls

Use unsupervised machine learning approaches to extract insights from emergency department narratives about how, when, and why older adults (age 65+) fall. Competition hosted by Centers for Disease Control and Prevention.

$70,000 in prizes

2 weeks left

612 joined

Compete! →

Navigation

Problem Description

Your goal in this challenge is to explore the application of unsupervised machine learning methods on emergency department (ED) visit narratives about older (age 65+) adult falls. Ultimately, insights gained through such analyses can help inform policies and interventions to reduce older adult falls.

> [NEW!] DrivenData and CDC/NCIPC are hosting a Q&A event on September 6th, 2023 from 1:30 - 2:30 PM ET via Zoom! The event will be recorded and shared with all solvers. To learn more about the event and register, check out the Q&A announcement.

Data

The data for this challenge come from a large, public dataset: the National Electronic Injury Surveillance System (NEISS). NEISS data are managed and maintained by the Consumer Product Safety Commission and come from a representative sample of emergency departments in the United States. NEISS only includes injury records when a product is involved. The information in this data has been manually extracted from ED health records.

There are three data files provided on the data download page.

- `primary_data.csv`: Limited set of ~115,000 ED visit records from verified unintentional falls
- `supplementary_data.csv`: Broader set of ~418,000 ED visit records with imperfect fall identification
- `variable_mapping.json`: Mapping between encoded integers and their string values

You are required to use the primary dataset in your submission. Use of the supplementary dataset is optional.

Both primary and supplementary datasets have the following columns:

- `cpsc_case_number` - Unique case ID
- `narrative` - Up to 400 character description of the incident (who, what, why, when, where, how)
- `treatment_date` - Date of ED treatment
- `age` - Age of the patient on the date of treatment
- `sex` - Patient's sex

- `race` - Patient's race
- `other_race` - Optional field used when there is no applicable race code
- `hispanic` - Whether the patient is Hispanic, Latino/Latina, or of Spanish origin
- `diagnosis` - Most severe clinical diagnosis
- `other_diagnosis` - Optional field when there is no applicable diagnosis code
- `diagnosis_2` - Secondary clinical diagnosis if there is more than one
- `other_diagnosis_2` - Optional field when there is no applicable diagnosis code and there is more than one diagnosis
- `body_part` - Body part with the most severe injury associated with the diagnosis
- `body_part_2` - Body part associated with diagnosis_2
- `disposition` - Patient outcome (i.e., treated, transferred, held for observation, etc.)
- `location` - Where the incident occurred
- `fire_involvement` - Whether the incident involved smoke inhalation or unexpected flames or smoke
- `alcohol` - Whether the patient consumed alcohol prior to or during the incident
- `drug` - Whether the patient used, or is regularly using, a drug(s) or medication(s) that contributed to the incident or severity of the injury
- `product_1` - Consumer product involved in the incident
- `product_2` - Optional second consumer product involved in the incident, where `0` means no second product
- `product_3` - Optional third consumer product involved in the incident, where `0` means no third product

Remember that the variables in the primary and supplementary datasets contain encoded numerical values. Check out the community code posts on how to use the `variable_mapping.json`. We've included both a Python version and an R version.

For more detail on the variables in this dataset, check out the NEISS coding manual.

The **primary data** has the following filters:

- Age: 65+
- Years: 2019-2022
- Limited to records from unintentional falls (externally verified)

It is required to use the primary data because it *only* contains unintentional falls, which is the focus of the challenge.

The **supplementary data** has the following filters:

- Age: 65+
- Years: 2013 - 2022
- Limited to records where the narrative contains "fall" or "fell"

The supplementary data are provided because a larger corpus of narratives may be helpful for distilling more generalizable language patterns and insights.

Most, but not all, of the records in the primary data are also contained in the supplementary data. There are ~8,000 fall records that are only in the primary data and not the supplementary data because they do not contain the keywords "fall" or "fell" but have been externally verified to be about falls.

Note that the fall-identification keyword filter in the supplementary data is imperfect. This means there will be false positives (e.g. records that include "fall" or "fell" but are not about a person falling) and false negatives (e.g. fall-related ED visits that were not captured by the keyword filter).

The length of the narrative field was also expanded to 400 characters in 2019, so in the supplementary dataset, you'll find shorter narratives for data before 2019.

External data are allowed in this competition, so you are also welcome to query the NEISS API directly with your desired filters.

More about the narrative field

You'll find that the `narrative` field shares a consistent structure. Below is an excerpt from the NEISS coding manual, which contains the instructions given to medical abstractors.

> - Age and sex must be at the start of the comment (i.e., 10YOM, 11MOF).
> - Be descriptive and include pertinent details about the incident (i.e., who, what, why, when, where, how). Describe the sequence of events, alcohol/drug involvement, and the affected body part(s) in the middle of the comment.

- The name and spelling of the product(s) must be correct. Include information about the brand, manufacturer, and model when known.
- Include the patient's alcohol concentration/level (BAC/BAL) whenever alcohol use is associated with the incident. If the patient's alcohol concentration/level was not taken or recorded, the comment must state this.
- The relevant clinician's diagnoses should be at the end of the comment exactly as written in the ED record and denoted with "DX:". This abbreviation helps distinguish the clinician's diagnoses from other details about symptoms and complaints. If there are multiple diagnoses, separate them with punctuation marks. If there are no clinical diagnoses in the ED record, put "NO DX" at the end of comment.
- Use sentence case (i.e., capitalize only the first word of each sentence and "DX:") and correct any spelling errors.

Labels

This challenge is the first of its kind: unsupervised! That means there are no labels for this competition.

External Data and Models

External data and pretrained models are permitted if you have the rights, licenses, and permissions to use them. **If you use any external data or pretrained models, you must clearly document them as part of your submission.**

Large language models (LLMs), including proprietary models, are permitted for this competition. However, you must access them via programmatic API access and not user inferfaces. This means, for example, if you want to use OpenAI's ChatGPT, you must use the OpenAI API and **not manually interact** with the ChatGPT web application.

To help get you started exploring LLMs, we've provided:

- A community code post for working with the Falcon 7B model to demonstrate how to work with an open-source LLM

- Embeddings for the primary dataset narratives using OpenAI's `text-embedding-ada-002 model`. For more about embeddings, check out the OpenAI guide.

If you use any external datasets or pretrained models, we encourage you to share them with other participants either in the Community Code section or on the competition forum.

Submission Format

A full submission for this challenge is a zip archive with the extension `.zip`, which includes the following two items:

1. Notebook of analysis
2. Executive summary

You are allowed to make only one submission. To make changes, you can delete and re-upload your submission as many times as you like. Only the last entry submitted will be considered.

See the following sections for further details on requirements for each.

Notebook

Your analysis should be presented in a **single** notebook file demonstrating "literate programming". Your notebook should clearly and logically integrate code, explanation, and visuals to communicate your analysis and results.

Your notebook must satisfy the following requirements:

- The programming language must be either Python or R
- The analysis uses the `narrative` field from `primary_data.csv` in a substantive way
- The following notebook file formats are accepted:
  - Jupyter Notebook
  - R Markdown
  - Quarto
- Submissions should be viewable with rendered outputs without requiring rerunning

    - For Jupyter notebooks, this means you must cleanly execute the entire notebook before submitting
    - For R Markdown and Quarto, you must include both the .rmd/.qmd source file and a rendered version (e.g., PDF or HTML) of the document
- Clearly document any external datasets or pretrained models that are used

Executive Summary

The executive summary should succinctly summarize the key ideas of your analysis. It should explain what you tried, why, what worked or didn't, and what you found. Executive summaries must follow these requirements:

- Up to 1 page of text; up to 3 pages total including references, figures, and tables
- 8.5 x 11 inch paper size with minimum margin of 1 inch
- Minimum font size of 11
- Minimum single-line spacing
- PDF file format

Suggested template

The following is a suggested template for your executive summary.

> ### Key findings
>
> - What are 2-3 takeaways from your exploration?
>
> ### Summary of your approach
>
> - Describe the data source(s) you used (e.g., if supplementary or external data was used, how the data was subsetted, if at all).
> - What did you do and why (e.g., preprocessing, key features, algorithms you used, unique or novel aspects of your solution, etc.)?
> - What worked and what didn't? How did you evaluate the performance of your approach?
> - What are some limitations of your analysis?

> ### Visualizations
>
> - Do you have any useful tables, charts, graphs, or visualizations from the process (e.g., exploring the data, testing different features, summarizing model performance, etc.)?

Evaluation

Winners will be selected by a judging panel of domain experts and researchers. Submissions will be judged according to the following weighted criteria:

- **Novelty (35%)**
  To what extent does this submission utilize creative, cutting-edge, or innovative techniques? This can be demonstrated in any or all parts of the submission (e.g., preprocessing, embeddings, models, visualizations).
- **Communication (25%)**
  To what extent are findings clearly and effectively communicated? This includes both text and visuals.
- **Rigor (20%)**
  To what extent is this submission based on appropriate and correctly implemented methods and approaches (e.g., preprocessing, embeddings, models) with adequate sample sizes?
- **Insight (20%)**
  To what extent does this submission contain useful insights about the effectiveness of unsupervised machine learning methods at uncovering patterns in this data and/or informative findings that can advance the research on circumstances related to older adult falls?

Midpoint Feedback and Prizes

You will have an opportunity partway through the competition to submit an executive summary and receive feedback from the judges. The feedback will be focused on improving the elements outlined in the judging criteria in the previous section (novelty, communication, rigor, insight). See the previous section on executive summary format requirements. Depending on the

number of submissions, feedback will either be provided individually or in a single forum post capturing overall trends.

Judges will also select three midpoint submissions for "most promising" bonus prize awards. These bonus prize winners will be selected using the same evaluation criteria as the final evaluation and receive $2,500 each.

Executive summary submissions for midpoint feedback are due by **August 23, 2023**.

Guiding Questions

We know this is an open-ended challenge. To help, we are providing some guiding lines of inquiry to help you get started. Feel free to use these or try something completely different!

- What information about falls is contained in the narrative but not captured by the other variables?
- What precipitating events — actions that happen right before the fall — can be identified?
- How do falls (e.g., severity, type of fall injury) and fall circumstances (e.g., precipitating event, activity involved) differ among various demographic groups (e.g., by race, sex, age)?
- Are there trends over time in the types of falls that occur?
- What kinds of people, places, or things are associated with more or more severe falls? Are there alternative explanations for these associations?
- What risk factors are associated with falls?

The goal of this challenge is to explore the application of unsupervised machine learning techniques to this dataset. Below are some examples of analysis that would be a good fit for this challenge, as well as examples of analysis that are not a good fit.

Good fit

- Remove words from the narratives that are found in other variables (e.g., male, female, body parts) and cluster the remaining text to identify what information is in the narratives that is not captured elsewhere. Then use word embeddings to remove similar words and repeat the analysis.
- Design a way to identify the "precipitating event" before a fall (e.g., tripping on carpet) and compare differences in precipitating events among various demographic groups.

- Explore various clustering algorithms using the provided embeddings. Use the ChatGPT API to produce summaries of the clusters. Compare the results of using the embeddings versus using more traditional bag-of-words preprocessing techniques (e.g., stop word and punctuation removal, stemming and lemmatization, and token weighting such as TF-IDF).

Not a good fit

- Trying to identify coders or hospitals based on regional dialect or semantic structure of narratives.
- Focusing on gaps in data quality (e.g., "5% of the time this demographic field does not align with narrative").
- Focusing only on information contained in the non-narrative variables.
- Producing visualizations that don't tell a clear story (e.g., word clouds).
- Many different techniques without a compelling hypothesis or clear rationale that ties them together into a coherent analysis.
- Overly simplistic approaches like a count vectorizer with logistic regression.

Community Code

This competition is the first time we have a Community Code section! This is a place where you can share and view code and analysis from other participants in the competition. To encourage contributions, judges will award a "most helpful shared code" bonus prize of $2,500 to the individual or team that contributes the most high-quality and useful content during the competition. You can access the Community Code section via the navigation sidebar.

Good luck!

Good luck! Check out the Community Code section for some ways to get started, and consider contributing to it yourself. If you have any questions, you can visit the user forum!

DRIVENDATA

We run data science competitions that create AI solutions for social good.

Work with us

As a partner
As a competitor
Join a competition
Careers

About us

What we do
Who we are
Blog

Get in touch

Contact us
Twitter
LinkedIn

Terms
Privacy
Copyright policy