Summary

Using text classification methods on the samples labeled by ChatGPT, we obtained a full dataset of falls with four new variables: prior activity, intrinsic cause, outdoor location, and assisted living.

We showed how these new indicators vary by age and sex, highlighted the temporal dynamics of intrinsic falls, and estimated how fall circumstances affect the risk of hospitalization.

Methodology

First, we labeled a sample of falls/not falls narratives using ChatGPT API and merged it with the primary dataset. Then we trained an SVM model for predicting falls in the entire NEISS database (18+ year-olds, 2013-2022) and obtained the whole dataset of falls for this period. Next, we separately labeled the four variables in the sample of fall-related narratives with ChatGPT API and used them to predict those variables in the full dataset of falls.

Based on the previous literature, the following fall-related variables were labeled by ChatGPT:

1. Prior activity, multiple-choice options: *Lying in bed, Walking up/downstairs, Walking in general, Bathing, Riding, Standing up/down, Sitting, Doing sports/exercises, Reaching/leaning/bending, Climbing, Carrying, Other activity, Not found*

2. Intrinsic cause of the fall (e.g., syncope, faint, weakness), a binary variable

3. Outdoor location of the fall, a binary variable

4. Assisted living (e.g., fall in nursing home), a binary variable

```
'''
You are an intelligent medical worker trying to code information from fall-related emergency department narratives.
Classify the cause of the fall as 'intrinsic' (patient passed out and fall due to the internal reasons: syncope, faint, weakness etc.) or
'not intrinsic' (all other cases).
Examples:
    input: 'blacked out and fell to the floor landing on her buttocks.'
    output: 'intrinsic'
    input: 'was mopping floor, lost balance, slipped on wet floo
    output: 'not intrinsic'
    input: 'became dizzy and fell in the shower'
    output: 'intrinsic
Here is the narrative:
'''
```

Figure 1. Example of ChatGPT prompt for the labeling of intrinsic causes

The final database contained 1,147,085 fall-related incidents for all adult patients during the 2013-2022 period, including 495,821 records of 65+ year-olds with the newly generated variables described above.

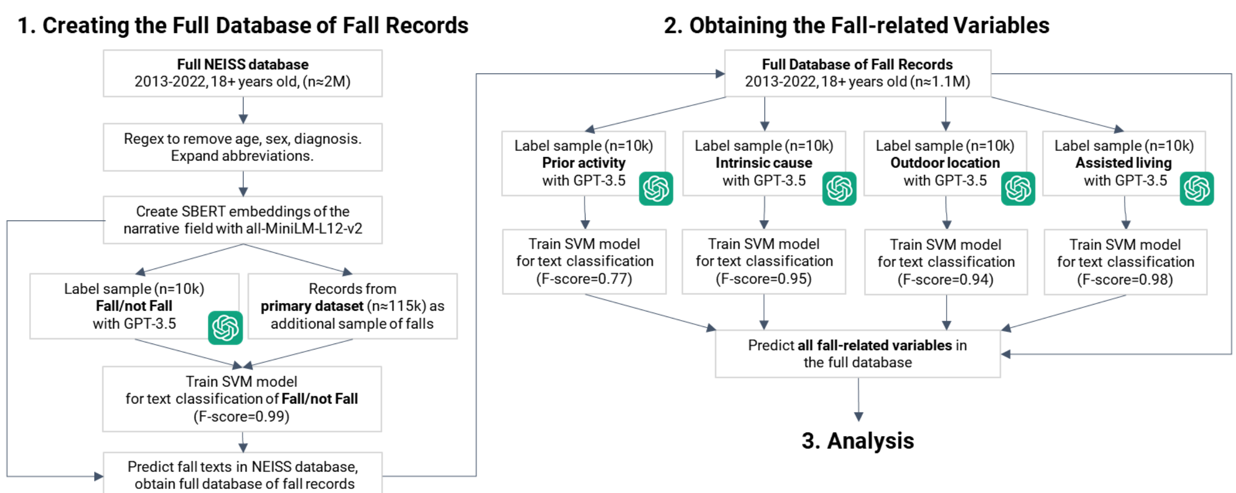

Figure 2. Data processing pipeline

Analysis

Table 1. Fall types by sex and age groups for older adults, % by column

	Female n=314,668	Male n=181,150	65-74 years n=167,876	75-84 years n=163,446	85+ years n=164,499
Intrinsic cause	4.9	7.4	6.5	6.2	4.6
Outdoor location	5.5	11.5	12.4	7.4	3.3
Assisted living	9.3	7.7	3.3	7.0	15.9
Prior activity					
Lying in bed	14.6	14.4	11.5	14.4	17.8
Walking up/downstairs	12.3	12.0	16.5	12.6	7.4
Walking in general	11.8	7.1	9.8	10.6	9.8
Bathing	9.1	8.9	8.3	9.0	9.8
Riding	1.3	4.2	4.4	1.9	0.7
Standing up/down	2.4	2.2	1.9	2.4	2.7
Sitting	2.2	2.2	1.8	2.2	2.7
Doing sports/exercises	1.5	2.8	3.2	1.9	0.8
Reaching/leaning/bending	2.1	1.6	1.5	1.9	2.5
Climbing	0.7	4.0	3.6	1.6	0.4
Carrying	1.5	1.5	1.8	1.6	1.1
Other activity	2.0	2.1	2.1	2.0	2.0
Not found	38.3	37.0	33.5	37.8	42.3

Table 1 shows that most falls of older adults happened while lying in bed, using the stairs, walking, and bathing. Substantially more accidents occurred during physical activities (riding, sports, climbing) among men than women. We can also notice a greater proportion of falls from the lying position and in assisted living facilities among the oldest group (85+) compared to other ages. Interestingly, around a third of narratives did not contain prior activities (e.g., "fall to the floor"). Moreover, the amount of unspecified cases increases with age, suggesting some variability in how different groups report their incidents and the potential of bias associated with this.

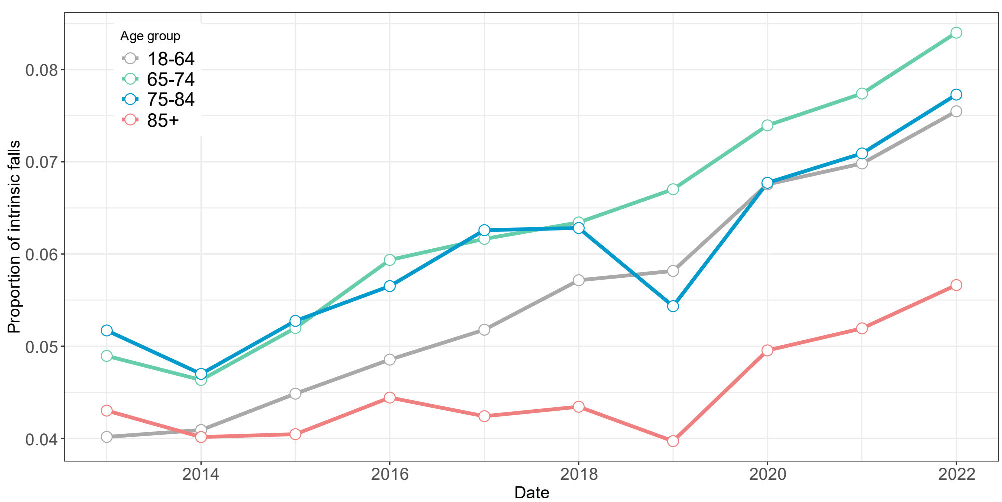

Figure 3. Proportion of intrinsic falls over time

In *Figure 3*, we can observe a distinct rise in intrinsic falls (syncope, fainting, weakness, etc.) over time for all age groups. It is remarkable that the upward trend is not unique to older adults: the same pattern appears for the 18-64-year-olds.

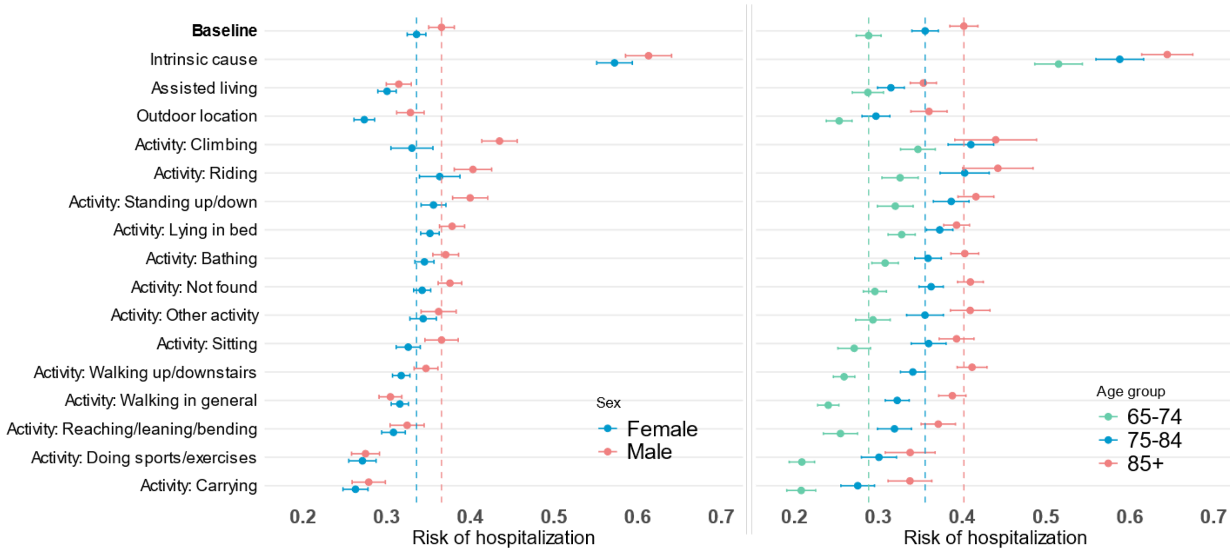

Figure 4. Risk of hospitalization for older adults by age and sex

We ran a set of separate GLM regressions for each sex and age group category to estimate how the newly extracted variables affect the risk of hospitalization, controlling for race, drug or alcohol involvement in the incident, stratum, year, and month. *Figure 4* demonstrates that compared to the baseline risk, the intrinsic causes of falls significantly increase the probability of hospitalization. Contrarily, being in assisted living and outdoors reduce such a risk. Riding and climbing are among the most unsafe activities for older adults, especially for males. The 85+ years old patients experience higher hospitalization rates for most activities relative to other ages: even falls during walking can frequently lead to hospitalization for the them.

Conclusion

This work demonstrated how labeling large text samples using ChatGPT coupled with the ML classification methods can be applied to extract various fall-related variables from the narrative field. This method is adaptable to introducing other potential variables depending on the research objectives and is well-suited for large sample sizes.

The constraints of this methodology include reliance on proprietary access to ChatGPT API for labeling the data and the necessity to manually devise one’s set of variables and categories. Moreover, when calibrating this approach, a researcher should simultaneously deal with two types of errors: the first one stemming from ChatGPT inaccurately labeling the narrative sample and the second one – from the mispredictions of the classification model on the full dataset.

In our subsequent steps, we aim to focus on increasing the classification accuracy for the prior activity variable, systematically assess the quality and bias of the ChatGPT labeling using a manually coded subsample, and attempt to fill the unspecified information on prior to a fall activity using methods for the missing values imputation. Additionally, we plan to explore some intriguing substantive relationships in the data, residing on our initial findings.