Predicting survival times from narratives using survival models? by *Research4Good*

Rationale: Timely admission to the emergency department (ED) is an important determinant of outcomes in elderly patients. Yet, unnecessary hospital visits increase the burden on health systems and risk bringing anxiety and stress to patients and their family members and caregivers. Accordingly, we **examined how the time till hospital visit (nicknamed as "delay time") as currently captured in the narratives may be predictive of patient outcomes.** Once "delay times" were extracted, we also **developed prognostic survival models that predict time to an adverse outcome using data collected at baseline** (narrative, baseline, and delay times). As secondary analyses, we examined: (IIa) two acuity-based patient outcomes: time of fall till hospital treatment and time of fall till hospitalization; (IIb) correlations between narratives' word embeddings with severity rankings derived by the "severity regressor" that was developed by Kang et al.

Methods. We employed data from both *primary* and *supplementary* as shown in Table2. Our evaluation set consists of cases from a wide period (2013-2020 as opposed to 2019-2022), which allows us to examine temporal generalization, and that these subsets yield sizes comparable to open survival datasets reported in recent benchmarks, e.g. SUPPORT (n=9,105), FLCHAIN (n=7,894).

As problem statement instructs, we removed parts of each narrative that were already captured in the other columns, i.e. starting characters on sex, age, and end phrases that follow the marker "DX" (or equivalent markers such as "***", ">>"). Then, we performed preprocessing steps on the narratives (details on p2). Next, we computed five different sets of word embeddings: all-mpet-base-v2 (WB1), all-MiniLM-L6-v2 (WB2), paraphrase-mpnet-base-v2 (WB3), universal sentence encoder (WB4), and LEALLA (WB6). We also included the challenge's OpenAI embeddings (WB11) in our analyses.

Derivations of survival times. We inferred the number of hours from the time of fall incident to the time of hospital visit by searching for keywords, such as "1DAY AGO", "YESTERDAY", "LAST NITE". Cases whose survival times cannot be determined were excluded in the survival analyses. Patients whose dispositions do not match the outcome definitions were treated as censored (these include patients who left the hospital before being seen).

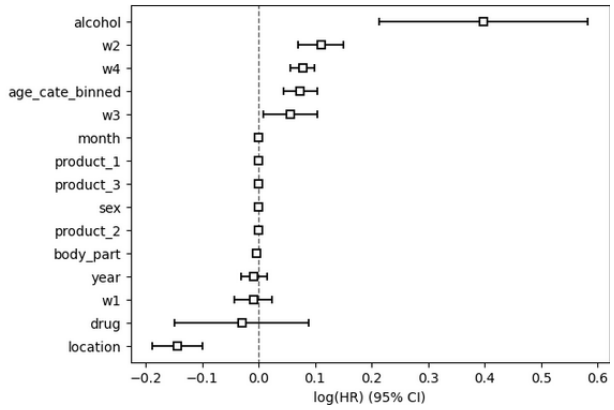
Outcome def#3 employed more involved processing of the "DX" narratives: 1) "DX:", "DX|", ">>" were unified to " DX. "; 2) We searched for keywords defining the 46 regions employed by Chung's method, and additionally searched for combined phrases "HEAD INJURED" or "HEAD INJU"; 3) each "DX" narrative generated a 46-feature vector and then got fed into Kang's severity regressor.

Model development. We performed Bayesian optimization (Optuna package) to find the optimal parameters of eXtreme Gradient Boosters (XGB) survival models that predict *P* probabilities of experiencing an adverse outcome at times 1 to P. For baseline comparison only**,** we fitted regularized Cox's regression models. For each outcome, we developed two survival models (XGB/Cox) under >5 combinations of 3 input types: patient's baseline data, raw word embeddings and their dimensionality reduced versions.

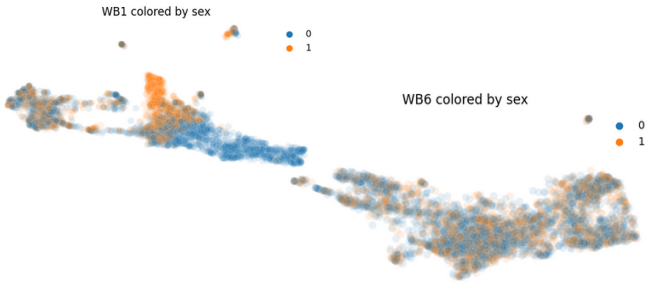
Key messages. Based on our key findings as detailed on the last page, we recommend that "delay time" be tracked, as our analyses showed that this additional information can be used to predict adversity in patient outcomes, with predictive accuracy reaching 0.72 in C-index. Furthermore, long lie, which refers to an inability to get up after falling, is a marker of illness. According to literature, 50% of fallers experienced long lies. Future work shall examine incidence rates of long lies over years, as increased rates may call for the development of new protocols/remedies (e.g. wearable alarms/ smartwatches/ GPS widgets to alert their caregivers and/or given to recurring fallers). We note that "delay time" is different from the hour of the fall (morning/evening), which is also not tracked in the current NEISS coding manual but has long been suggested in literature (e.g. SPLATT mnemonic).

Contributions. All our analysis scripts can be switched to public-readable mode on Kaggle by Oct 6 1159 UTC; the research community can regenerate our results and work with the intermediate files archived on Kaggle once the challenge host permits (only kept private now per challenge rules).

Results of Cox's regression analysis. We dropped fire involvement due to its sparsity in order to bypass degenerate mathematical formulation. We omitted diagnosis as they correlate with disposition (likewise DX information in the narratives were stripped out before calculating word embeddings). Log of the hazard ratios and their confidence intervals are shown in the figure below.



We also interpreted the word embeddings via UMAP. Word embeddings generated by appeared to capture sex differences:



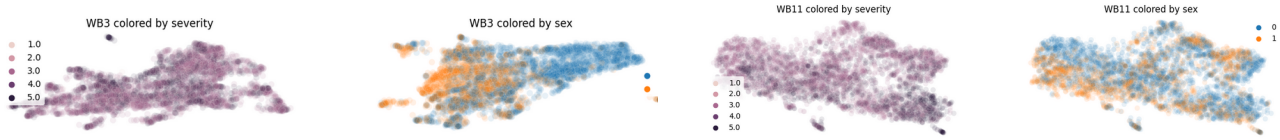


Table1. Splits. After merging two CSV files into a dataframe, we created non-overlapping subsets:

	Development set		Evaluation set 1	Evaluation set 2
Abbreviation/ code:	"trn1"	"trn2"	"val"	"tst"
Periods	2019-2022	same	same	2013-2022
# of cases with race "white"	1700	1735	3495	13,277

Table2. Median time to ED visit

Race	White	Black/ AA	Mixed	Asian	Amer. Indian/ Native	Native Hawaiian	White hispanic	Black hispanic	Cambodian
Time to ED	12 6-18	12 6-18	12 7.5-18	12 12-24	15 6-18	6 9-15	12 12-24	12 5.5-12	12 6-18

Table 3. Performance of prognostic models for predicting survival time till hospitalization. Aggressive training.

C-index	WB1	WB2	WB3	WB4	LEALLA	OpenAI	WB19	WB20	All WBr		BL only
XGB (Eval1)	60.7	56.7	63.8	61.8	61.9	65.87	60.84	59.4	70.9	Cox (Eval1)	56.5
XGB (Eval2)	61.3	58.3	64.7	63.0	63.1	n/a	61.12	59.4	71.9	Cox (Eval2)	55.9

Table 4. Performance (measured by C-index) of prognostic models for predicting survival time till treatment. BL denotes use of baseline data only (d=11, i.e. age category, sex, product_1, product_2, product_3, year, month, race, drug, alcohol, location). "WBr" denotes all dimensionally reduced word embeddings.

Model	WB1	WB2	WB3	WB4		BL only	BL+WRr
XGB (Eval1)	71.4	69.6	70.1	71.9	Cox (Eval1)	55.9	61.5
XGB (Eval2)	73.2	71.2	70.9	72.1	Cox (Eval2)	56.4	60.7

Findings. We found that the use of "trn1" for model development was too aggressive; as shown in Table 3, none of the developed models could achieve accuracy in C-index higher than 0.70.
When using the entire development set, the use of embeddings from "universal sentence encoder" achieved predictive performance in C-index of 71.9 on Evaluation#1 when predicting survival time till hospital treatment. The performance generated was not impacted by differences in time; models achieved C-index 72.1 on Evaluation#2. Word embeddings from LELLA gave comparable performance but required more compute hours. Hence, we recommend the use of the former. We also found that unsupervised dimensionality reduction generated by all-mpet-base-v2 and all-MiniLM-L6-v2 tended to yield "sex-based" clusters while the more computationally more intensive models (OpenAI and LEALLA) did not generate such "bias".

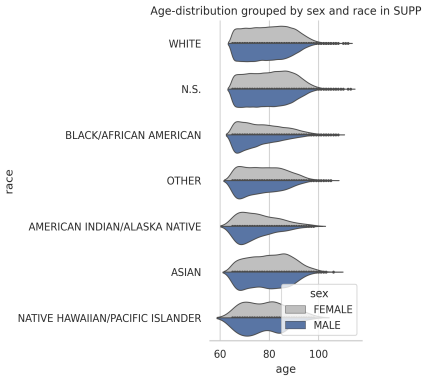
Summary listing of our contributions:

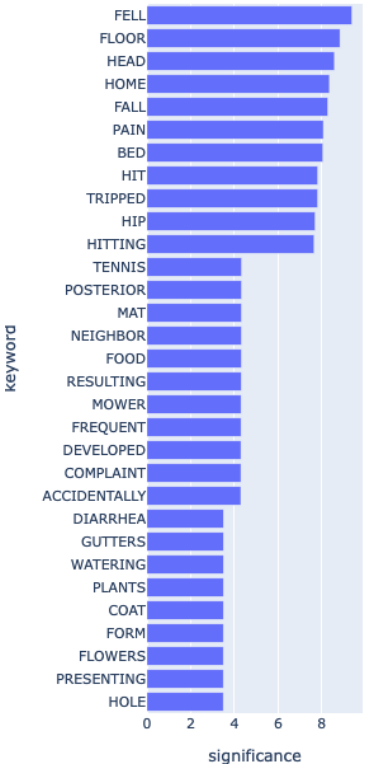
- We examine all available samples by merging the data in *primary.csv* and *supplementary.csv* to define *development* and internal *evaluation* sets to allow for evaluation of generalization.
- We benchmarked different types of word embedding models, and developed over dozens of survival models, each of which involved >3 input combinations, i.e. using raw word embeddings and their dimensionality reduced versions, and their combined effect on patient's baseline data (e.g. sex, age category, location of fall, etc).

What did not work?

- We explored the applicability of the severity score regressor developed by Kang et al., which calculates a score based on the severity of injuries of up to three body parts. Since our cohort captured only up to two body parts and their severities are not available, we used this regressor only as a ranker. The Spearman's correlation between the ranks and "delay times" were low and not statistically significant in all data splits.

Exploratory visual analyses from midpoint submission:
Below is a plot on Age, Sex, Race; the figure on the right shows the most common words.





Appendix

Preprocessing steps: 1) typos that were manually observed during preliminary analysis were corrected; 2) triplets of symbols (e.g. *, >) were reduced to one instance or spelled out (e.g. "&" changed to "and", @ changed to "at"; 3) common medical abbreviations were replaced with English words; 4) lemmatization was performed using SPACY package (ensuring special words remain unmodified, e.g. "left foot" would not be converted to "leave foot"); 5) Words that were not recognizable by PySpellChecker (e.g. when character space is missing between two English words, e.g. "ANDFELL") were split into two substrings if the first substring was deemed valid per PySpellChecker. Note that no training data was used to optimize the processing steps.