Predicting 30-Day Mortality in Sepsis Patients Using Text-Extracted Frailty Indicators from MIMIC-III

Michelangelo Pagán

May 27th, 2025

1 Introduction

Clinical frailty is an important measure of physiological vulnerability to various health stressors. In critically ill patients with sepsis—a leading cause of ICU mortality—integrating frailty into risk assessment may improve outcome prediction.

Current risk stratification models rely on structured clinical data such as vitals, lab results, and comorbidity indices. However, frailty remains underutilized, largely because it is rarely captured in structured EHR fields. While administrative frailty indices exist, they require structured inputs that overlook latent signals of frailty related to diet, activity, and functional deficits.

This project addresses that gap by extracting binary frailty indicators from clinical notes, supplementing traditionally collected lab values. We investigate whether these text-derived frailty signals can improve prediction of 30-day mortality in ICU patients with sepsis. This approach of integrating frailty indicators from free text into predictive models could provide a low-cost, non-invasive enhancement to existing models, especially in settings lacking structured frailty measures.

2 Background

Since the emergence of the ”frailty phenotype” as a clinical concept, it has been linked to adverse outcomes such as falls, re-hospitalization, mortality, and increased ICU length of stay. In septic ICU patients, significant differences in physiological and laboratory-based frailty index (FI-Lab) scores have been found between survivors and non-survivors using the MIMIC-IV database (Ding et al., 2024). Similarly, the Modified Frailty Index (MFI), derived from ICD-10 codes, has been associated with elevated mortality rates and 90 and 180 days post-discharge (Li et al., 2024). While these findings underscore clinical frailty's prognostic value, they rely on administrative codes or structured data that are prone to missingness and inconsistent EHR documentation.

To address this limitation, researchers have begun exploring latent indicators of frailty that are not in structured chart values. Frailty is strongly determined by functional activity, activity patterns, changes to diet, fatigue, and muscle weakness that appear in clinical notes but are absent in structured EHR data. Prior work has used Natural Language Processing (NLP) to derive frailty scores from free text, such as using word embeddings to answer a binary frailty questionnaire called the Tillburg Frailty Indicator (TFI) from clinical notes (Wijesinghe et al., 2024).

The purpose of this project is to determine whether frailty indicators can be extracted meaningfully from clinical notes without complicated frailty indexes or NLP methods which may be computationally expensive. Instead of using black-box models, this approach uses interpretable binary indicators for frailty by detecting keywords in notes. These features can be easily integrated into conventional modeling pipelines, allowing for transparency and reproducibility of modeling across multiple contexts.

3 Methods

3.1 Data

The data for this analysis were drawn from the MIMIC-III v1.4 dataset, containing de-identified records from over 40,000 hospital visits between 2001 and 2011 (Johnson et al., 2016). A sepsis cohort was extracted according to the Sepsis-3 criteria using a script provided in the mimic_sepsis repository published by Microsoft. This cohort includes all patients who developed sepsis in the ICU, with observations 24 hours before and 48 hours after presumed sepsis onset. The sepsis cohort includes only the first ICU stay, patients ≥ 18 years old, and ICU stays ≥ 1 day. All vitals extracted through this script were z-normalized and saved as a CSV file for further processing.

3.2 Text Preprocessing and Frailty Feature Engineering

Data from the ADMISSIONS, PATIENTS, ICUSTAYS, and NOTEEVENTS tables were joined onto this cohort using admission and ICU stay IDs. Frailty-related keywords were compiled from online clinical sources (see Supplementary Table S1). Discharge summaries were selected as the clinical text source since they provide a comprehensive snapshot of the patient's stay at the hospital and ICU, including pre-existing conditions, physical deficits, and other frailty indicators that do not appear in structured lab data. Vital signs and lab metrics were aggregated by summary statistics, depending on the variable (see Supplementary Table S2). Finally, to align with the 30-day mortality outcome, only patients who survived their first ICU stay were included.

Once discharge summaries were linked to ICU stays, the History of Present Illness (HPI) section was extracted a cleaned using a Python script. Keyword matches with each of the clinical terms were counted and added as binary columns in the dataframe, along with a total 'frailty score' derived from the sum of these binary columns. A binary column for 30-day mortality was derived from the PATIENTS table date-of-death (DOD) column.

3.3 Modeling Approach

With the final dataset of demographics, normalized vitals, and binary frailty flags with 30-day mortality, the data was split into training and testing sets with mortality as the true label. Data was divided using an 80-20 training-testing split. Then, three model sets were evaluated:

1. **Baseline**: Only demographics and vitals for training.
2. **Frailty Only**: Demographics and frailty for training.
3. **Full**: Demographics, frailty, and vitals for training.

Class imbalance was accounted for through the use of the RandomOverSample function from imblearn's over_sampling module, which randomly over-sampled the minority class (patients who died within 30 days of discharge). Logistic regression from the sklearn module was applied, as well as Extreme Gradient Boosting (XGBoost) from the xgboost module and a Support Vector Classifier (SVC) from sklearn. Models were evaluated through comparison of AUCROC curves, precision, sensitivity, specificity, and AUC values.

4 Results

The performance metrics for the logistic regression models are presented in Table 1. The baseline model achieved an AUC of 0.734. The frailty-only model showed reduced performance with an AUC of 0.715, but the full model combining structured features from frailty indicators with existing lab values improved to an AUC of 0.753. This improvement demonstrates the additive value of extracting frailty indicators from text data using a keyword matching strategy from discharge summaries. Across all models, sensitivity was moderate but precision was low, indicating a tendency towards false positives and over-prediction of mortality. While the frailty-only model underperformed in overall discrimination, several frailty indicators - such as cachexia, malnutrition, and difficulties with Activities of Daily Living (ADLs) - emerged as meaningful predictors of 30-day mortality. In the full model, the top ten features by importance were frailty-extracted binary features, especially cachexia, deconditioning, and dependence in ADLs. Vital signs and demographic data that were important in these models included Blood Urea Nitrogen (BUN) and Age, indicating that frailty can be most effective when used alongside traditional clinical variables.

Metric	**Baseline (ROS)**	**Frailty Only (ROS)**	**Full Model (ROS)**
AUC	0.734	0.715	0.753
Accuracy (%)	65.5	69.7	71.2
Sensitivity (%)	66.3	57.8	62.0
Specificity (%)	65.5	70.3	71.7
Precision (%)	8.9	9.0	10.0

Table 1: Performance metrics for logistic regression models with Random Over-Sampling (ROS).

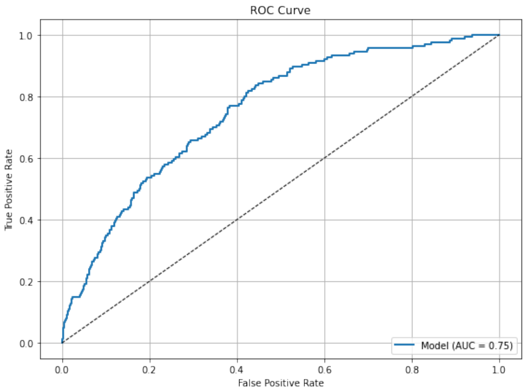

Figure 1: AUROC Curve for Randomly Oversampled Logistic Regression with all Input Features.

For all model variations, XGBoost consistently performed with a sensitivity less than 0.4, as it biased its predictions towards the majority class despite random over-sampling and other class balancing techniques.

5 Discussion and Conclusions

This analysis demonstrates that incorporating frailty indicators from free-text clinical notes can improve mortality prediction in sepsis patients, independent of structured lab and administrative data in EHRs. These features resulted in improved predictive performance while adding minimal complexity to the models (keyword match counts can be found in Supplementary Table S3). The enhanced performance of models including frailty indicators suggests that these derived features capture latent signals of patient condition that are not present in vitals alone, aligning with previous literature asserting frailty as a key vulnerability indicator in sepsis. These indicators can be derived without necessitating deep learning or heavy pre-processing, allowing for their adaptability and up-scaling across EHR systems.

However, several limitations must be acknowledged. First, discharge summaries in this analysis may not capture temporal information present in other notes (nursing, progress, social work, etc.) useful for health forecasting. Second, discharge disposition - such as transfer to hospice or rehabilitation - was not considered and may confound mortality outcomes in older adults. Third, although binary encoding of frailty keyword presence is easily interpretable, it may miss subtle linguistic variations, overlook negations in phrasing (such as ”not losing weight”), and ignore important contextual clues in discharge summaries. Additionally, demographic information included in this study was limited to age and gender, not accounting for known impacts of socioeconomic factors and network strength in outcomes. Finally, the inherent class imbalance in 30-day post-discharge mortality outcomes likely affected model calibration, as reflected in precision-recall tradeoffs of the logistic regression model.

In the future, this analysis can be expanded upon to account for some of these limitations and extract clinical frailty signals from text data more effectively. Expanding frailty detection through token-based matching patterns or more sophisticated NLP methods like ClinicalBurt could extract frailty signals beyond keywords and include negation, temporality, and sentiment analysis. Expanding text sources beyond discharge summaries would also build richer frailty profiles. Including in-hospital mortality cases and validating against established frailty measures like the Hospital Frailty Risk Score (HFRS) and Clinical Frailty Scale (CFS) will be essential for strengthening the generalizability and clinical relevance of these findings.

References

Ding, H., Li, X., Zhang, X., Li, J., and Li, Q. (2024). The association of a frailty index derived from laboratory tests and vital signs with clinical outcomes in critical care patients with septic shock: a retrospective study based on the mimic-iv database. *BMC Infectious Diseases*, 24(1):573.

Johnson, A. E. W., Pollard, T. J., and Mark, R. G. (2016). MIMIC-III Clinical Database (version 1.4).

Li, X., Tang, Y., Deng, X., Zhou, F., Huang, X., Bai, Z., Liang, X., Wang, Y., and Lyu, J. (2024). Modified frailty index effectively predicts adverse outcomes in sepsis patients in the intensive care unit. *Intensive and Critical Care Nursing*, 84:103749.

Wijesinghe, Y. V., Xu, Y., Li, Y., and Zhang, Q. (2024). A phrase-based questionnaire–answering approach for automatic initial frailty assessment based on clinical notes. *Computers in Biology and Medicine*, 170:108043.

Supplementary Materials

Supplementary Table S1: Frailty Keywords Used for Text Extraction

Frailty Keywords
frail, frailty, deconditioned, deconditioning, cachectic, cachexia, sarcopenia, decline in health, decline in function, failure to thrive, poor functional status, chronically ill, failure to progress, wheelchair bound, walker, cane, mobility aid, limited mobility, difficulty walking, unsteady gait, shuffling gait, requires assistance to ambulate, uses wheelchair, fall, falls, fall risk, frequent falls, recent fall, immobile, bed bound, bedridden, nonambulatory, easily fatigued, fatigued, tired all the time, fatigue, lethargy, exhaustion, low energy, poor stamina, unable to complete tasks, limited activity tolerance, muscle wasting, muscle weakness, generalized weakness, grip weakness, decreased strength, loss of strength, unable to lift self, reduced muscle tone, unintentional weight loss, losing weight, weight loss of, lost weight, malnourished, malnutrition, poor appetite, not eating, eating less, anorexia, hypoalbuminemia, decreased oral intake, slowed cognition, memory issues, poor concentration, social withdrawal, confusion, delirium, disoriented, withdrawn, apathetic, depressed, isolated, adl assistance, requires help with adls, dependent for activities of daily living, needs help dressing, incontinent, toileting assistance, dependent on caregiver, requires 24 hour care, home health aide, assisted living, rehab facility, progressive decline, overall decline, functionally limited, no longer independent, physically weak, debilitated, functional decline, senile, aging related decline, discharge to skilled nursing facility, poor rehab potential

Table 2: List of frailty-related keywords used to generate binary indicators from unstructured clinical notes.

Supplementary Table S2: Aggregation Strategy for Structured Features

Feature Name	Aggregation Function
o:age, o:gender, o:re_admission	first
o:mechvent, o:SOFA, o:SIRS, o:Shock_Index	max
a:action, r:reward	sum
o:Weight_kg, o:HR, o:SysBP, o:MeanBP, o:DiaBP, o:RR, o:Temp_C, o:FiO2_1	mean
o:Potassium, o:Sodium, o:Chloride, o:Glucose, o:Magnesium, o:Calcium	mean
o:HB, o:WBC_count, o:Platelets_count, o:PTT, o:PT, o:INR	mean
o:Arterial_pH, o:paO2, o:paCO2, o:Arterial_BE, o:HCO3, o:Arterial_lactate	mean
o:PaO2_FiO2, o:cumulated_balance, o:SpO2, o:BUN, o:Creatinine	mean
o:SGOT, o:SGPT, o:Total_bili	mean
o:input_total, o:input_4hourly, o:output_total, o:output_4hourly	sum
o:max_dose_vaso	max
o:GCS	min

Table 3: Feature-level aggregation strategy used in preprocessing vital signs for model input.

Supplementary Table S3: Histogram of Frailty Keywords Matched from Discharge Summaries

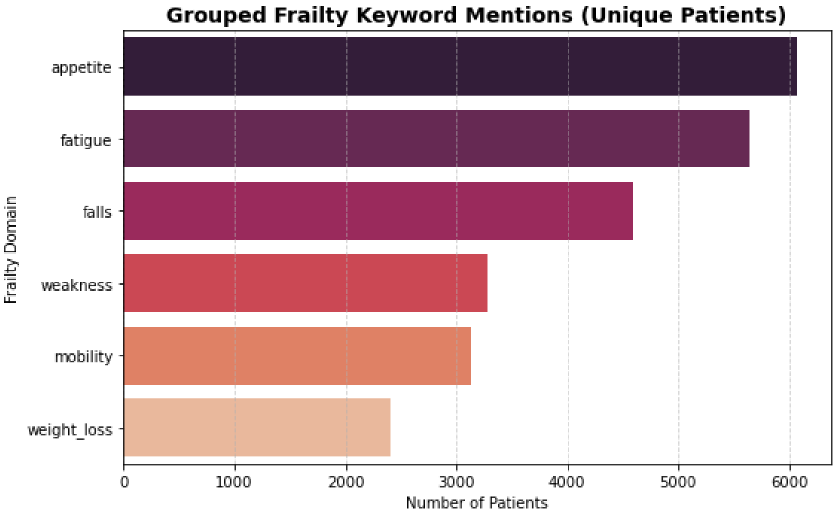

Figure 2: Histogram of frailty keyword grouped by category.