Natural Language Processing Methods to Identify Oncology Patients at High Risk for Acute Care with Clinical Notes

Claudio Fanconi, B.Sc.[1,2], Marieke van Buchem, M.Sc.[1,3], Tina Hernandez-Boussard, Ph.D., M.P.H., M.Sc.[1]
[1]Stanford University, Stanford, California, United States
[2]ETH Zürich, Zürich, Switzerland
[3]Leiden University Medical Center, Leiden, The Netherlands

Abstract

Clinical notes are an essential component of a health record. This paper evaluates how natural language processing (NLP) can be used to identify the risk of acute care use (ACU) in oncology patients, once chemotherapy starts. Risk prediction using structured health data (SHD) is now standard, but predictions using free-text formats are complex. This paper explores the use of free-text notes for the prediction of ACU in leu of SHD. Deep Learning models were compared to manually engineered language features. Results show that SHD models minimally outperform NLP models; an ℓ_1-penalised logistic regression with SHD achieved a C-statistic of 0.748 (95%-CI: 0.735, 0.762), while the same model with language features achieved 0.730 (95%-CI: 0.717, 0.745) and a transformer-based model achieved 0.702 (95%-CI: 0.688, 0.717). This paper shows how language models can be used in clinical applications and underlines how risk bias is different for diverse patient groups, even using only free-text data.

Introduction

Oncology patients undergoing chemotherapy often need acute care utilisation (ACU) and hospitalisation after starting chemotherapy. These interventions account for nearly half of the costs associated with oncology care in the United States[1,2]. Evidence suggests roughly 50% of these healthcare encounters are potentially preventable through early outpatient interventions[3,4]. A previous paper by Peterson et al[5] introduced a machine learning (ML) model, using structured health data (SHD) from electronic health records (EHR), to identify patients at high risk for ACU after chemotherapy initiation. In total, they trained their model using 760 inputs and retained 125 to predict the risk of ACU. This work, and others, highlight the potential of data-driven models to predict ACU risk using SHD[6–8].

However, most EHRs are not mapped to a common data model and they are not necessarily standardised between different facilities. To replicate other hospital's predictive models they could require intensive data preparation. On the other hand, 96% of hospitals in the US collect digital clinical notes from physicians and nurses in 2019[9]. Natural language processing (NLP) methods can extract useful information from these unstructured clinical texts.

NLP methods have already proven useful in clinical applications, e.g. for predicting critical care outcomes in intensive care units[10], classifying procedures and diagnoses[11], or predicting outcomes after an ischemic stroke[12]. In particular, deep learning-based language models have become popular in recent years[13], being used to identify 30-day hospital readmissions[14,15] or Statin non-use[16].

This study aims to assess the added value for identifying patients at risk of needing ACU by replacing tabular inputs with features from unstructured clinical notes or by combining both modalities. Second, we aim to investigate whether novel deep learning language models outperform traditional language feature extraction and linear models. We investigated these aspects by developing five predictive models of ACU risk trained with different inputs and compared their predictive performance and utility when used at the point of care.

Methods

Data Collection

In 2019, the Centers for Medicare & Medicaid Services (CMS) introduced the Chemotherapy Measure (also referred to as OP-35), a quality measure that captures hospital admissions or emergency department visits of adult patients related to potentially preventable diagnoses within 30 days of starting outpatient chemotherapy[17]. Based on this measure, a study population at a comprehensive cancer center, including a large tertiary outpatient clinic, was assembled for risk prediction at 30, 180 and 365 days after chemotherapy initiation[5]. The OP-35 metric itself is used as a label for supervised learning and defines a positive event.
For the SHD inputs, we use the original 760 features from Peterson et al.[5] extracted from the same EHR database, such as demographic, social, vital sign, procedural, diagnostic, medication, laboratory, health care utilisation, and cancer-specific data generated 180 days before the first date of chemotherapy. For a detailed description of how the patient cohort was extracted, the inclusion and exclusion criteria for the OP-35 metric, and a full list of features, please see the original paper[5].
Based on the above study population, we matched patients to their respective progress notes and the history and physical (H&P) notes from the EHR database (Epic Systems Corp). We removed notes of less than 100 words, as these were mainly erroneous entries, and notes of more than 5,000 words, as these often contained long copies of previous notes and laboratory analyses. We also removed history notes with mentions of clinical trial consents, as based on our review these were copy-paste texts. Finally, we extract and aggregate the most recent clinical notes (at most three) created 180 days before the patient started chemotherapy, same as in the SHD collection from Peterson et al.[5]. If a patient had no clinical records in the EHR database, they were removed from the study population.
The cohort was previously randomly divided into a training set (80%) and a test set (20%) for modelling, and we, therefore, kept exactly these patient sets (except the ones without any clinical notes) to obtain comparable results.

Model Development

Five different risk prediction models were compared in this study: *Tabular LASSO*, *Language LASSO*, *Fusion LASSO*, *Language BERT* and *Fusion BERT*.
Tabular LASSO. This model is a logistic regression with an ℓ_1-penalty (also known as Least Absolute Shrinkage and Selection Operator - LASSO). The inputs were the 760 structured health data points from Peterson et al.[5].
Language LASSO. The Language LASSO model is an ℓ_1-penalised logistic LASSO regression with manually generated inputs from the clinical notes. The notes were preprocessed as follows: First, we removed special characters and personal, organisational, date and time entities using SpaCy's[18] part of speech tagging. Then we tagged negated terms with a "not_" using SpaCy's negator library. We removed auxiliary words, adpositions, determiners, interjections and pronouns. Subsequently, we lemmatised[19] the remaining words. Finally, we followed the method of Marafino et al.[10] by filtering out the 2,000 most frequent terms[1] of all the notes and weighting these words using the Term Frequency-Inverse Document Frequency (TF-IDF) algorithm. The Language LASSO has 2,000 input features corresponding to the 2,000 most frequently occurring words.
Fusion LASSO. The Fusion LASSO is also a logistic regression LASSO model. This time it uses both, the tabular data and TF-IDF values, as input features. We combined them to inspect if data extracted from the clinical notes has added value to SHD.
Language BERT. This model is a deep learning-based Bidirectional Encoder Representation of Transformers[20,21] (BERT). This model does not require manual feature engineering and can consume clinical notes with little preprocessing. We used a pre-trained distilBERT[22] model as the encoding structure, as it

---

[1]We also tested the algorithm with 500, 1,000, and 3,000 most occurring terms. While increasing the number of terms increased the predictive performance, the increase became insignificant after 2,000. We omitted these results from the main paper because of space constraints.
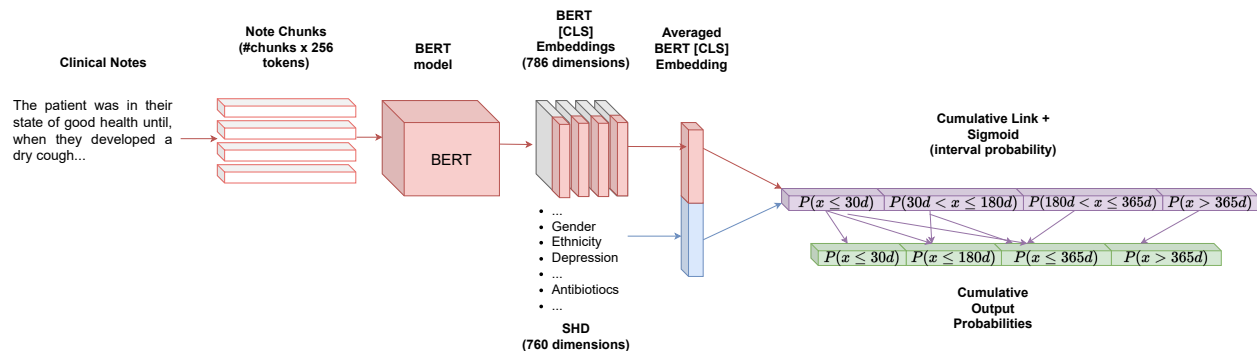
Figure 1. Overview of the Fusion BERT model. The Language BERT only contains the upper (red) encoder architecture before leading into the output probabilities (purple and green).

requires less computation than the more familiar BERT or ClinicalBERT[14] models[2]. DistilBERT and other transformer models are often computationally limited by the input size of the text (often referred to as token length in the deep learning literature). To avoid these GPU memory overflows, we decomposed the clinical notes into chunks of at most 25 sequences, each 256 tokens (1 token ≈ 1 word), and ran them through the neural network. We aggregated the outputs of the transformers (also called embeddings) by averaging over the embeddings of the corresponding clinical note. We connected the averaged embedding (representing a complete clinical note) linearly to a single output neuron, whose value is divided into four sections. A sigmoid function is applied to assign a probability to each of these values. These four slices represent the probability distribution of an ACU event within the time intervals emanating from the different ground truth labels ($P(x \leq 30d)$, $P(30d < x \leq 180d)$, $P(180d < x \leq 365d)$ and $P(x > 365d)$). Since a patient who experienced an ACU event within the first 30 days is also eligible for an event within 180 days and 365 days, we add the corresponding probabilities to get the original ground truth interpretation of an ACU within 30 days ($P(x \leq 30d)$), 180 days ($P(x \leq 180d)$), 365 days ($P(x \leq 365d)$) and not within 365 days ($P(x > 365d)$).

Fusion BERT. The fusion BERT model is the same as the language BERT model, except that the corresponding SHD are concatenated with the output embedding. The newly-concatenated embedding was then linearly connected to the output neuron. Figure 1 shows an overview of the fusion BERT.

The regularisation hyperparameters of the LASSO models were determined by tenfold cross-validation grid search, while the hyperparameters of the two BERT models were determined by using 20% of the training data as validation data. While the LASSO models are trained on each time interval of the label (30d, 180d, and 365d) individually, the BERT models are trained on all the labels simultaneously. We applied a cumulative link loss[23] to train the neural network with backpropagation, to ensure the ordinal regression structure[24] (e.g avoid cases where $P(x \leq 30d) > P(x \leq 180d)$, as in this use case any patient that had an ACU event within 30 days, subsequently is marked to have had an ACU within 180).

Model Evaluation

We first evaluated the models by their discriminative performance with the area under the receiver operating characteristic curve (AUROC), the area under the precision-recall curve (AUPRC) and the negative log-likelihood (cross-entropy) with a 1'000-fold bootstrap to obtain 95% confidence intervals.

We reported the number of SHD used during risk prediction. We did this for the LASSO models by summing the number of non-zero model coefficients originating from SHD. For the Fusion BERT, we counted the number of connections of tabular features to the output neuron that have values less than 0.001, as the backpropagation algorithm is not optimized for feature selection, unlike the LASSO.

To assess the model calibration, calibration curves[25] were developed. Finally, we assessed the initial clinical utility of these four models through a Decision Curve Analysis (DCA)[26]. A DCA plots the net benefit across a range of decision thresholds and quantifies the number of true positives versus false positives. The curves

---

[2]Using a pre-trained ClinicalBERT over a pre-trained distilBERT did not yield any significant improvements in prediction. We, therefore, omitted its results from the paper due to space constraints.

for the prediction models were compared to two alternative clinical strategies: treat all (everyone is treated as if they will have an ACU event) and treat none (nobody is treated as if they will have an ACU event).

To test the discriminatory power of the model in a setting similar to that in which it might be used at the point of care, the test cohort was stratified into high, medium and low-risk groups based on the tertile of predicted risk. Kaplan-Meier[27] survival curves for OP-35 events are used to examine the separation between risk groups for language LASSO and language BERT on 180-day ACU risk prediction. In addition, the ten highest and lowest coefficients of the language LASSO model are presented. This helps us to determine the importance of certain keywords in the clinical notes.

Since Peterson et al.[5] have reported unfair algorithmic results for ACU prediction from structured data, we investigated whether language features might also be affected. We present and compare the empirical cumulative distributions of predicted risk score percentiles for subgroups to assess how the models predicted each subgroup's risk for OP-35 events. Specifically, we examine demographic values (i.e. race and insurance type) and tumour stage on the Language LASSO model for 180-day ACU risk prediction.

Results

Table 1. Information about the complete patient cohort (train and test set) eligible for the OP-35 metric for 30, 180, and 365 day prediction. "std" stands for standard deviation.

Patient Characteristic	Total Cohort (N=6,938)	OP-35 Events within 30 days (n=936, 13.5%)	OP-35 Events within 180 days (n=2,202, 31.7%)	OP-35 Events within 365 days (n=2,704, 39.0%)	No OP-35 Events within 365 days (n=4,234, 61.0%)
Age, mean± std					
At diagnosis	58.7±14.3	57.2±15.3	57.7±15.1	57.9±15.0	59.2±13.9
At first chemotherapy	60.5±14.4	58.9±15.2	59.4±15.1	59.6±15.0	61.0±14.0
Sex, No. (%)					
Female	3,659 (52.7)	474 (50.6)	1132 (51.4)	1417 (52.4)	2242 (53.0)
Race, No (%)					
White	3,804 (54.8)	461 (49.3)	1,113 (50.5)	1,379 (51.0)	2,425 (57.3)
Asian	1,619 (23.3)	226 (24.1)	536 (24.3)	649 (24.0)	970 (22.9)
Black	188 (2.7)	42 (4.5)	88 (4.0)	100 (3.7)	88 (2.1)
Other or unknown	1,327 (19.1)	207 (22.1)	465 (21.1)	576 (21.3)	751 (17.7)
Ethnicity, No. (%)					
Non Hispanic/Latino	5,989 (86.3)	788 (84.2)	1,867 (84.8)	2,280 (84.3)	3,709 (87.6)
Hispanic or Latino	855 (12.3)	142 (15.2)	327 (14.9)	414 (15.3)	441 (10.4)
Cancer type, No. (%)					
Breast	1,321 (19.0)	113 (12.1)	275 (12.5)	346 (12.8)	975 (23.0)
Gastrointestinal	819 (11.8)	93 (9.9)	291 (13.2)	366 (13.5)	453 (10.7)
Thoracic	774 (11.2)	107 (11.4)	258 (11.7)	326 (12.1)	448 (10.6)
Lymphoma	700 (10.1)	170 (18.2)	345 (15.7)	382 (14.1)	318 (7.5)
Head and neck	658 (9.5)	90 (9.6)	208 (9.4)	238 (8.8)	420 (9.9)
Pancreas	585 (8.4)	99 (10.6)	214 (9.7)	280 (10.4)	305 (7.2)
Prostate	520 (7.5)	11 (1.2)	46 (2.1)	70 (2.6)	450 (10.6)
Gynecologic	513 (7.4)	70 (7.5)	176 (8.0)	218 (8.1)	295 (7.0)
Genitourinary	461 (6.6)	76 (8.1)	184 (8.4)	219 (8.1)	242 (5.7)
Other	587 (8.5)	107 (11.4)	205 (9.3)	259 (9.6)	328 (7.7)
Cancer stage, No. (%)					
Stage I	1,099 (15.8)	123 (13.1)	281 (12.8)	338 (12.5)	761 (18.0)
Stage II	1,415 (20.4)	141 (15.1)	336 (15.3)	410 (15.2)	1005 (23.7)
Stage III	964 (13.9)	131 (14.0)	351 (15.9)	429 (15.9)	535 (12.6)
Stage IV	1,898 (27.4)	327 (34.9)	759 (34.5)	937 (34.7)	961 (22.7)
Unknown	1,562 (22.5)	214 (22.9)	475 (21.6)	590 (21.8)	972 (23.0)
Insurance, No. (%)					
Medicare	2,683 (38.7)	323 (34.5)	788 (35.8)	970 (35.9)	1,713 (40.5)
Private	2,450 (35.3)	328 (35.0)	747 (33.9)	898 (33.2)	1,552 (36.7)
Medicaid	599 (8.6)	130 (13.9)	258 (11.7)	307 (11.4)	292 (6.9)
Other or unknown	1,206 (17.4)	155 (16.6)	409 (18.6)	529 (19.6)	677 (16.0)

A total of 6,938 patients were included in the study cohort compared to the original cohort, which included 8,439 patients. The mean age at chemotherapy initiation was 60.5 years (± 14.4 years), and 52.7% were female. A total of 936 patients (13.5%) met the primary criteria of having at least one OP-35 event within the first 30 days of starting chemotherapy, 2,202 (31.7%) within the first 180 days and 2,704 (39.0%) within the first year. The majority of patients in the cohort were white (n=3,804, 54.8%), followed by Asian patients (n=1,619, 23.3%), then other and unknown races (1,327, 19.1%) and least represented were black patients

(n=188, 2.7%). The most common cancer type was breast cancer (n=1,321, 19.0%), gastrointestinal tumours (n=819, 11.8%), thoracic cancer (n=774, 11.2%) and lymphoma (n=700, 10.1%), which accounted for more than half of all data. ACU events occurred most frequently in lymphoma (30d: n=170, 18.2%; 180d: n=345, 15.7%; 365d: n=382, 14.1%) and least frequently in prostate cancer (30d: n=11, 1.2%; 180d: n=46, 2.1%; 365d: n=70, 2.6%) across all time periods. Most chemotherapy patients had a stage IV tumour (n=1,898, 27.4%), which was also most common in ACU events (30d: n=327, 34.9%; 180d: n=759, 34.5%; 365d n=937, 34.7%). The most common type of insurance in the cohort was Medicare (n=2,863, 38.7%) and private health insurance (n=2,450, 35.3%). The cohort characteristics are summarised in Table 1.

Table 2. Resulting metrics on the test set of the tabular, language and fusion LASSO models, as well as the language and fusion BERT, trained on 30, 180 and 365 days ACU prediction with the 95%-CI in the brackets. The best-performing metrics for every label type are marked in bold. We also display the number of SHD used for prediction in the third column, where "N/A" means that SHD was used for prediction.

Label	Model	No. SHD	AUROC	AUPRC	Cross-Entropy
30	Tabular LASSO C=0.02	83	0.775 (0.757,0.792)	**0.411** (0.373,0.447)	0.344 (0.329,0.358)
	Language LASSO C=0.03	N/A	0.726 (0.707,0.744)	0.294 (0.264,0.323)	0.363 (0.346,0.379)
	Fusion LASSO C=0.02	73	**0.778** (0.760,0.795)	0.410 (0.372,0.447)	**0.341** (0.326,0.356)
	Language BERT	N/A	0.710 (0.692,0.729)	0.259 (0.235,0.282)	0.435 (0.415,0.455)
	Fusion BERT	419	0.766 (0.749,0.784)	0.315 (0.286,0.343)	0.393 (0.377,0.406)
180	Tabular LASSO C=0.03	221	0.748 (0.735,0.762)	0.623 (0.600,0.647)	0.540 (0.527,0.552)
	Language LASSO C=0.02	N/A	0.730 (0.717,0.745)	0.577 (0.555,0.601)	0.558 (0.546,0.570)
	Fusion LASSO C=0.02	101	**0.765** (0.752,0.779)	**0.632** (0.610,0.655)	**0.530** (0.517,0.543)
	Language BERT	N/A	0.702 (0.688,0.717)	0.543 (0.517,0.567)	0.625 (0.603,0.644)
	Fusion BERT	419	0.753 (0.741,0.767)	0.620 (0.597,0.644)	0.548 (0.536,0.558)
365	Tabular LASSO C=0.02	150	0.763 (0.752,0.775)	**0.704** (0.685,0.724)	0.559 (0.549,0.569)
	Language LASSO C=0.02	N/A	0.732 (0.719,0.745)	0.639 (0.618,0.661)	0.585 (0.575,0.595)
	Fusion LASSO C=0.02	115	**0.770** (0.759,0.782)	0.702 (0.683,0.722)	**0.553** (0.541,0.563)
	Language BERT	N/A	0.709 (0.695,0.723)	0.617 (0.594,0.640)	0.666 (0.647,0.683)
	Fusion BERT	419	0.760 (0.748,0.774)	0.695 (0.675,0.714)	0.565 (0.554,0.575)
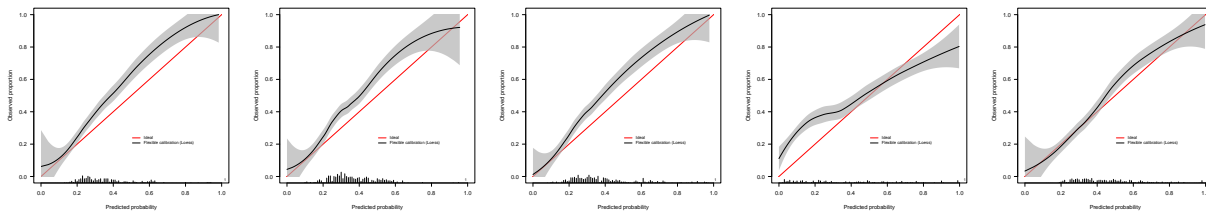
(a) Tabular LASSO **(b)** Language LASSO **(c)** Fusion LASSO **(d)** Language BERT **(e)** Fusion BERT

Figure 2. Calibration curves of the 180-day ACU risk prediction models. The red line indicates ideal calibration, while the black line is the flexible calibration with the 95% confidence interval, generated with the prob.cal.ci.2 function[25].

Model Performance

Table 2 lists the AUROC, AUPRC, and cross-entropy scores including the 95% confidence intervals of the five risk models for 30-day, 180-day and 365-day ACU prediction. For the 30 day acute care risk prediction, the Fusion LASSO model performs best on AUROC (0.778, 95%-CI: 0.760, 0.795) and cross-entropy (0.341, 95%-CI: 0.326, 0.356), using 73 SHD features. The highest AUPRC score has the Tabular LASSO (0.411, 95%-CI: 0.373, 0.447) compared to the event rate of 13.5%, using 83 tabular variables.

For 180-day ACU prediction, the Fusion LASSO model performs best in all metrics with 101 SHD features. The Language LASSO has a 0.730 (95%-CI: 0.717, 0.745) AUROC score and the Language BERT achieves 0.702 (95%-CI: 0.688, 0.717), both of them without using any structured data.

In the full-year ACU prediction, we observe that the Fusion LASSO scores again the highest C-stastic (0.770, 95%-CI:0.759, 0.782) and the lowest cross-entropy loss (0.553, 95%-CI:0.541, 0.563), using 115 tabular features, while the Tabular LASSO has the highest AUPRC score (0.704, 95%-CI:0.685, 0.724), using 150 SHD points. We show the flexible calibration curves for the 180-day models in Figure 2, where we observe a risk underestimation of the three LASSO models and underestimation of low risk patients and overestimation of high risk patients with the Language BERT model.

The Fusion BERT uses the most SHD points (419 tabular inputs) for all three label types to make predictions.

Exploration of Clinical Usage of Language Models

The Decision curve analysis for the 180-day ACU prediction showed that the net benefit of the Language-BERT model yields a negative benefit when the decision threshold for treatment is chosen above 0.6 (Figure 3) and less or equal net benefit than treating every patient with a threshold below 0.19. The other models, including the Language-LASSO model, have positive benefit values for decision thresholds below 0.7.

The Kaplan-Meier survival curves for OP-35 events showed good separation between risk groups (Figure 4, p < 0.001 for each group by log-rank test) for the two language-only models. By 180 days after the start of chemotherapy, 64 (13.9%) of the 462 low-risk patients in the language LASSO prediction had an OP-35 event and 76 (16.5%) in the language BERT prediction. On the other hand, 246 (53.2%) of the 462 high-risk patients had an event for the speech LASSO prediction and 238 (51.5%) for the language BERT prediction. Figure 5 shows the relative importance of the ten highest and lowest coefficients of the language LASSO model for the 180-day prediction. The words "Admission", "Failure", "Pain" and "Palliative" are among the ten highest coefficients, while "Breast", "PSA (Prostate-specific antigen)", "Nourished" and "Prostate" are among the ten lowest coefficients.
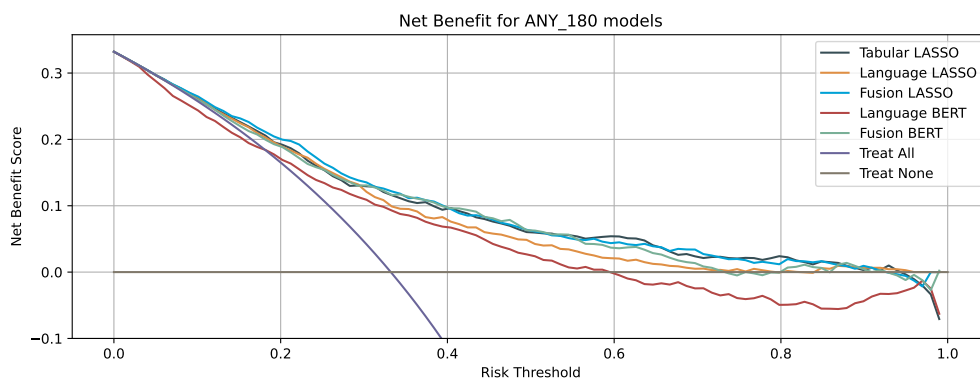

Figure 3. Net benefit curves of the tabular, language and fusion models. The purple curve indicates the benefit of all the patients treated, whereas the grey curve indicates the benefit is no patient is treated
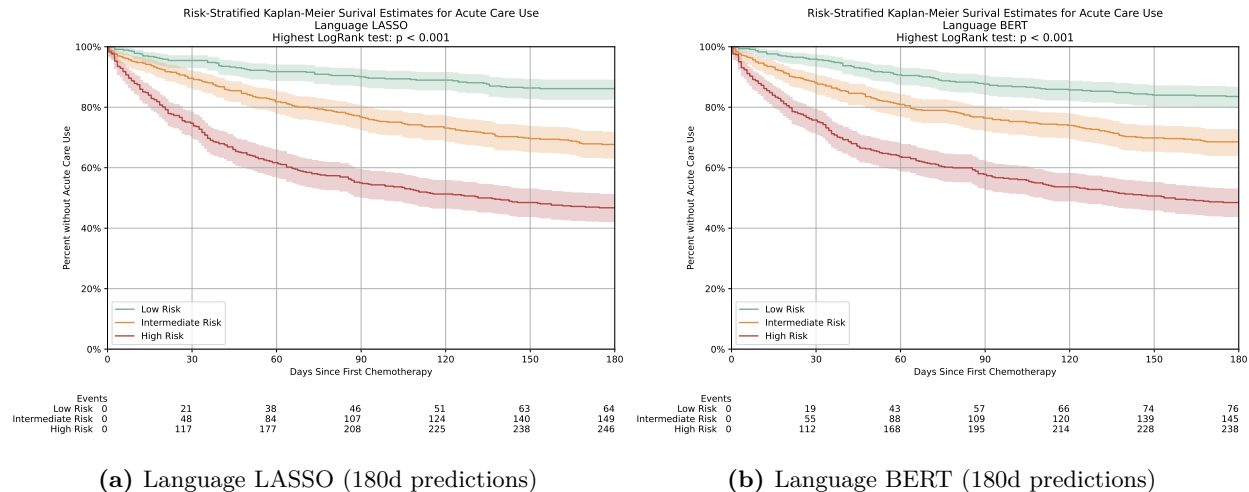
(a) Language LASSO (180d predictions)　　　　　**(b)** Language BERT (180d predictions)

Figure 4. Kaplan-Meier curves for ACU events for patients in the test cohort stratified by predicted risk. The shaded area represents the 95%CIs.
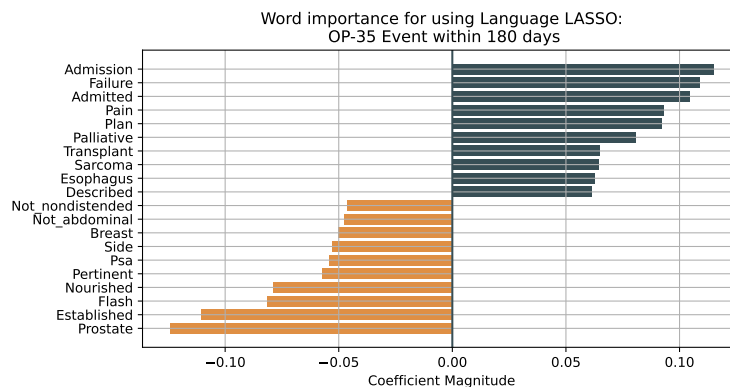

Figure 5. Coefficient magnitudes for the Language LASSO for 180-day ACU prediction, displaying the ten highest and ten lowest. The coefficients in this model are single words found in the clinical notes before the ACU event.

Sensitivity Analysis

Figure 6a shows that black patients are predicted to have a disproportionately higher risk than white, Asian or other-race patients. We note that the number of black patients is at least seven times smaller than that of non-black races. The cumulative risk by different insurance types is displayed in Figure 6b, where we note a risk overestimation of Medicaid patients.
In Figure 6c we see that the risk predictions for patients with stage III and IV tumours are overestimated, while the predictions for patients with stage I, II and unknown stage cancer are underestimated.

Discussion

Identifying methods to improve external generalisability is a priority for the medical informatics community. This paper presented an analysis of NLP to identify patients at risk of seeking acute care for different time intervals, a method which is scalable across sites and required minimal data preprocessing. It presents a comparison of logistic regression LASSO using structured health data with manually generated language features and the combination of both modalities. In addition, these models are compared to deep learning-based
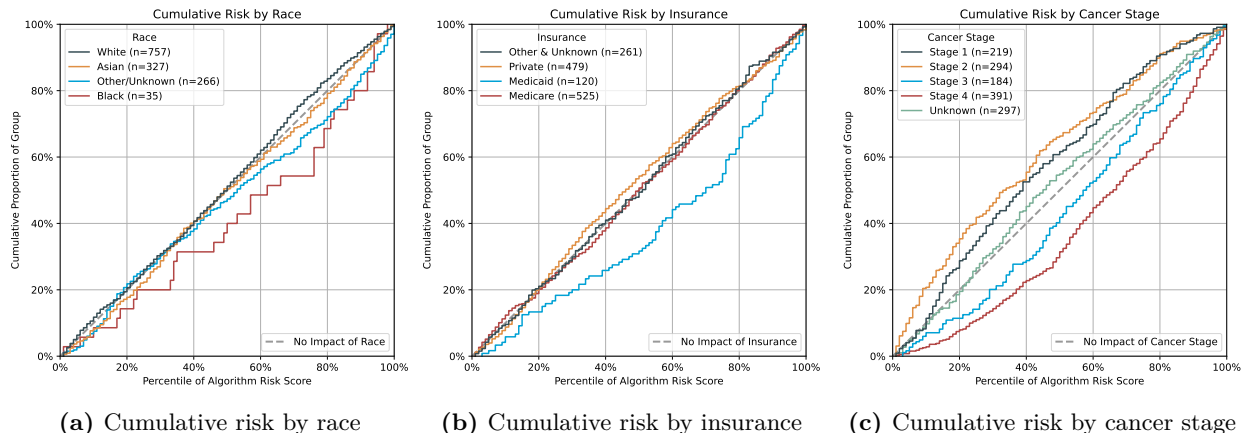
(a) Cumulative risk by race   **(b)** Cumulative risk by insurance   **(c)** Cumulative risk by cancer stage

Figure 6. Cumulative risk of the Language LASSO on 180 ACU prediction, stratified by race (a), insurance type (b) and over cancer stage (c).

transformers using either full-text clinical records only or in combination with structured health data.

Results demonstrate tabular LASSO outperforming language LASSO and language BERT at all time intervals. Nevertheless, we point out that both language-based methods achieve good discriminative performance (AUROC ≥ 0.7) even without the use of tabular health data, especially for 180-day risk prediction. On the other hand, combining the two input modalities (clinical notes and SHD) did not yield significant improvements over using tabular data alone. Regarding the selection model, results show that the popular BERT-based classifier does not outperform ℓ_1-penalised logistic regression with TF-IDF features and the fusion of both input modalities. This is likely due to the aggregation of chunks of the lengthy clinical documents into a single output probability, which makes deep learning training difficult.

Our study has several strengths; first, we show that NLP methods can be used instead of high-dimensional SHD to identify chemotherapy patients at risk for acute treatment. Secondly, our method optimises considerably training as we developed a transformer-based model that is trained simultaneously on multiple risk intervals in the form of nested ordinal regression, so that the computationally intensive training of the model for the different labels is not required.

Three main clinical implications can be drawn from this study. First, ACU risk prediction models for chemotherapy patients perform well using free-text data from the last (at most three) H&P and progress notes from physicians before chemotherapy. This implies that NLP methods could be easily implemented across sites and/or facilities as they only require access to written medical notes without the need for re-structuring or mapping of structured data, potentially saving costs in feature collection. Second, we show that LASSO coefficients can be used by clinicians to understand relative word meaning when making a prediction. BERT models lack this type of interpretive mechanism that allows clinicians to build confidence in the model. Third, in the sensitivity analysis, we find that certain groups may be subject to risk bias, despite not using demographic values explicitly as inputs. Although this needs further investigation, the results suggest that clinicians should be aware of these subgroup differences when interpreting the results of ML models.

Our study has limitations. First, more sophisticated transformer language models can handle longer notes[28] or also cross-modality[29]. However, these techniques require significant computation, which may not be feasible at most institutes. Second, our models have been validated only on one dataset for risk prediction in acute care. It would be beneficial to test the language models in a variety of medical problems. Finally, the work relies on data from a single academic cancer center, which may be not generalisable to other populations. In summary, this study demonstrates the utility of using free-text data to identify patients at risk of needing acute care once they have started chemotherapy. It is an alternative to structured health data, which may require significant preprocessing and may not be generalisable across settings. We show that the Language LASSO is a suitable model, especially for 180-day prediction, and that it is still interpretable. This work advances our knowledge on risk prediction models and provides an alternative for cross site generalisability. Hospital systems may consider these techniques to validate risk models.

Data and Code Availability

Under the terms of the data sharing agreement for this study, we cannot share source data directly. Requests for anonymous patient-level data can be made directly to the authors. All experiments were implemented in Python using the library SciKit-learn[30] for the metrics and the classical logistic regression model, and PyTorch[31] for the transformer models. We used R to create the calibration plots. The code for the logistic LASSOs and our analysis is available on www.github.com/su-boussard-lab/nlp-for-acu, while the code for the BERT model and training is available on www.github.com/su-boussard-lab/bert-for-acu.

References

1. Brooks GA, Li L, Uno H, Hassett MJ, Landon BE, Schrag D. Acute hospital care is the chief driver of regional spending variation in Medicare patients with advanced cancer. Health Aff (Millwood). 2014 Oct;33(10):1793-800.

2. Yabroff KR, Lamont EB, Mariotto A, Warren JL, Topor M, Meekins A, et al. Cost of care for elderly cancer patients in the United States. J Natl Cancer Inst. 2008 May;100(9):630-41.

3. Adelson KB, Dest V, Velji S, Lisitano R, Lilenbaum R. Emergency department (ED) utilization and hospital admission rates among oncology patients at a large academic center and the need for improved urgent care access. Journal of Clinical Oncology. 2014;32(30_suppl):19-9. PMID: 28141471. Available from: https://doi.org/10.1200/jco.2014.32.30_suppl.19.

4. Uno H, Jacobson JO, Schrag D. Clinician assessment of potentially avoidable hospitalization in patients with cancer. Journal of clinical oncology : official journal of the American Society of Clinical Oncology. 2014;32 30_suppl:4.

5. Peterson DJ, Ostberg NP, Blayney DW, Brooks JD, Hernandez-Boussard T. Machine Learning Applied to Electronic Health Records: Identification of Chemotherapy Patients at High Risk for Preventable Emergency Department Visits and Hospital Admissions. JCO Clinical Cancer Informatics. 2021;(5):1106-26. PMID: 34752139. Available from: https://doi.org/10.1200/CCI.21.00116.

6. Brooks GA, Kansagra AJ, Rao SR, Weitzman JI, Linden EA, Jacobson JO. A Clinical Prediction Model to Assess Risk for Chemotherapy-Related Hospitalization in Patients Initiating Palliative Chemotherapy. JAMA Oncology. 2015 07;1(4):441-7. Available from: https://doi.org/10.1001/jamaoncol.2015.0828.

7. Grant RC, Moineddin R, Yao Z, Powis M, Kukreti V, Krzyzanowska MK. Development and validation of a score to predict acute care use after initiation of systemic therapy for cancer. JAMA Network Open. 2019;2.

8. Daly RM, Gorenshteyn D, Gazit L, Sokolowski S, Nicholas K, Perry C, et al. A framework for building a clinically relevant risk model. Journal of Clinical Oncology. 2019.

9. Office of the National Coordinator for Health Information Technology: National trends in hospital and physician adoption of electronic health records; 2022. https://www.healthit.gov/data/quickstats/national-trends-hospital-and-physician-adoption-electronic-health-records.

10. Marafino BJ, Park M, Davies JM, Thombley R, Luft HS, Sing DC, et al. Validation of prediction models for critical care outcomes using natural language processing of electronic health record data. JAMA Netw Open. 2018 Dec;1(8):e185097.

11. Marafino BJ, Davies JM, Bardach NS, Dean ML, Dudley RA. N-gram support vector machines for scalable procedure and diagnosis classification, with applications to clinical free text data from the intensive care unit. J Am Med Inform Assoc. 2014 Sep;21(5):871-5.

12. Heo TS, Kim YS, Choi JM, Jeong YS, Seo SY, Lee JH, et al. Prediction of stroke outcome using natural language processing-based machine learning of radiology report of brain MRI. J Pers Med. 2020 Dec;10(4):286.

13. Wu S, Roberts K, Datta S, Du J, Ji Z, Si Y, et al. Deep learning in clinical natural language processing: a methodical review. J Am Med Inform Assoc. 2020 Mar;27(3):457-70.

14. Alsentzer E, Murphy J, Boag W, Weng WH, Jin D, Naumann T, et al. Publicly available clinical bert embeddings. In: Proceedings of the 2nd Clinical Natural Language Processing Workshop. Minneapolis, Minnesota, USA: Association for Computational Linguistics; 2019. p. 72-8. Available from: https://www.aclweb.org/anthology/W19-1909.

15. Huang K, Altosaar J, Ranganath R. ClinicalBERT: modeling clinical notes and predicting hospital readmission. arXiv:190405342. 2019.

16. Sarraju A, Coquet J, Zammit A, Chan A, Ngo S, Hernandez-Boussard T, et al. Using deep learning-based natural language processing to identify reasons for statin nonuse in patients with atherosclerotic cardiovascular disease. Communications Medicine. 2022;2.

17. 2019 chemotherapy measure facts admissions and emergency department (ED) visits for patients receiving outpatient chemotherapy hospital outpatient quality reporting (OQR) program (OP-35);. https://qualitynet.cms.gov/files/5dcc6762a3e7610023518e23?filename=CY21_OQRChemoMsr_FactSheet.pdf.

18. Eyre H, Chapman AB, Peterson KS, Shi J, Alba PR, Jones MM, et al. Launching into clinical space with medspaCy: a new clinical text processing toolkit in Python. In: AMIA Annual Symposium Proceedings 2021; (in press, n.d.). Available from: http://arxiv.org/abs/2106.07799.

19. Miller GA. WordNet: A Lexical Database for English. Commun ACM. 1992;38:39-41.

20. Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, et al. Attention is all You need; 2017. Available from: https://arxiv.org/pdf/1706.03762.pdf.

21. Devlin J, Chang MW, Lee K, Toutanova K. BERT: pre-training of deep bidirectional transformers for language understanding. ArXiv. 2019;abs/1810.04805.

22. Sanh V, Debut L, Chaumond J, Wolf T. DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. arXiv; 2019.

23. Pedregosa F, Bach FR, Gramfort A. On the consistency of ordinal regression methods. J Mach Learn Res. 2017;18:55:1-55:35.

24. Rosenthal E. Spacecutter: ordinal regression models in pytorch; 2018. Available from: https://www.ethanrosenthal.com/2018/12/06/spacecutter-ordinal-regression.

25. Van Calster B, Nieboer D, Vergouwe Y, De Cock B, Pencina MJ, Steyerberg EW. A calibration hierarchy for risk models was defined: from utopia to empirical data. J Clin Epidemiol. 2016 Jun;74:167-76.

26. Vickers AJ, van Calster B, Steyerberg EW. A simple, step-by-step guide to interpreting decision curve analysis. Diagnostic and Prognostic Research. 2019;3.

27. Kaplan EL, Meier P. Nonparametric estimation from incomplete observations. J Am Stat Assoc. 1958 Jun;53(282):457.

28. Beltagy I, Peters ME, Cohan A. Longformer: The long-document transformer. arXiv:200405150. 2020.

29. Khadanga S, Aggarwal K, Joty SR, Srivastava J. Using clinical notes with time series data for ICU management. ArXiv. 2019;abs/1909.09702.

30. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, et al. Scikit-learn: machine learning in python. Journal of machine learning research. 2011;12(Oct):2825-30.

31. Paszke A, Gross S, Massa F, Lerer A, Bradbury J, Chanan G, et al. PyTorch: An Imperative Style, High-Performance Deep Learning Library. In: Advances in Neural Information Processing Systems 32. Curran Associates, Inc.; 2019. p. 8024-35. Available from: http://papers.neurips.cc/paper/9015-pytorch-an-imperative-style-high-performance-deep-learning-library.pdf.