Available online at ScienceDirect

Resuscitation

journal homepage: www.elsevier.com/locate/resuscitation

Original paper

Large-scale automated phenotyping of cardiac arrest and withdrawal of life-sustaining therapy using electronic health record data

Catherine Clive[a,b,*], *Arjun Singh*[d], *Bram Overmeer*[a,c], *Spencer Boris*[a,b], *Lydia Peterson*[a,b], *Jaden Searle*[a,b], *Greg Hooke*[a,b], *Niels Turley*[a], *Marta Fernandes*[d], *Aditya Gupta*[a], *Manohar Ghanta*[a], *Valdery Moura Junior*[a,f], *S. Mukeriji*[d], *Sahar Zafar*[d], *Edilberto Amorim*[e], *M. Brandon Westover*[a,#], *Haoqi Sun*[a,#]

Abstract

Aims of the study: Anoxic brain injury following cardiac arrest is a leading cause of death in the United States. Withdrawal of life-sustaining therapy (WLST) is a common end-of-life decision in these patients, but its contributing factors and outcomes remain poorly understood. We developed machine learning models to enable large-scale, automated phenotyping to identify patients who died following WLST.

Methods: We used structured and unstructured EHR (Electronic Health Record) data from two major hospitals to train models that identify (1) patients with cardiac arrest and coma, and (2) patients who died after WLST. Performance was evaluated using the area under the receiver operating characteristic (AUROC) and precision-recall (AUPRC) curves, as well as other precision metrics.

Results: On holdout (internal) testing the models achieved AUROC/AUPRC values of 0.984/0.968 (cardiac arrest) and 0.992/0.991 (WLST). Cross-hospital evaluation showed strong performance for the cardiac arrest phenotype but variable generalizability for the WLST phenotype, with sensitivity depending on the training site. Population-level error rates were low (<0.5 %) for the cardiac arrest phenotype; estimates for WLST varied by hospital.

Conclusion: These models establish a reproducible framework for automated cohort identification. Nearly half of comatose post-arrest patients died following WLST, with 42 % of these deaths occurring within 72 h, highlighting the impact of early prognostication decisions. The models enable rapid cohort identification for research on neuroprognostication, including how WLST decisions may perpetuate self-fulfilling prophecies. Broader validation across health systems and larger cohorts will improve generalizability and inform evidence-based end-of-life decision-making.

Institutional review board approval: Mass General Brigham IRB BIDMC: 2022P000481; MGB: 2013P001024.

All procedures complied with institutional and national ethical standards; informed consent was waived for use of de-identified data.

Keywords: Machine learning, Electronic health records, Cardiac arrest, Brain injury, Prognostication, Natural language processing

Introduction

About 80 % of cardiac arrest patients remain comatose, with many failing to regain consciousness for days.[1–3] Accurate identification of postarrest coma and withdrawal of life-sustaining therapy (WLST) from electronic health records (EHRs) is essential for understanding practice patterns and outcomes after cardiac arrest. Withdrawal of life-sustaining therapy (WLST) is frequently chosen in cases where recovery may still be possible, confounding outcome prediction and contributing to "self-fulfilling prophecies", where poor outcomes occur because they were anticipated.[4]

Existing registries of data such as International Classification of Diseases (ICD) codes incompletely capture neurologic outcomes,

* Corresponding author at: Department of Neurology, Beth Israel Deaconess Medical Center, Boston, MA 02215, USA.
E-mail addresses: kittyclive@gmail.com, clive.catherine@yahoo.com (C. Clive).
Co-senior author.

https://doi.org/10.1016/j.resuscitation.2025.110919
Received 16 August 2025; Received in Revised form 25 November 2025; Accepted 26 November 2025

0300-9572/© 2025 Elsevier B.V. All rights are reserved, including those for text and data mining, AI training, and similar technologies.

and manual chart review is infeasible at scale.[5–8] Automated phenotyping using unstructured note data can fill this gap by simulating manual review in exponentially shorter time if methods are transparent and generalizable.[9,10]

We present a reproducible phenotyping pipeline that achieves high accuracy on internal holdout data and demonstrates feasibility for cross-hospital use, while highlighting limitations in cross-site generalizability (particularly for WLST) that warrant further multicenter validation. This model lays the groundwork for studying how WLST decisions are made and how they may be influenced by prior cases, potentially contributing to self-fulfilling prognostic cycles.[11]

Methods

Cohort construction

We used records from the Mass General Brigham (MGB; including Massachusetts General and Brigham and Women's Hospitals) deidentified institutional database (2013–2023), and Beth Israel Deaconess Medical Center's (BIDMC) publicly available MIMIC-III database (2001–2012).[12–14]

Data sources included physician discharge summaries, International Classification of Diseases (ICD) codes, and Current Procedural Terminology (CPT) codes. Only discharge summaries exceeding 500 words were included to ensure sufficient clinical detail. To balance positive and negative cases for training, we created an enriched cohort of 3000 unique patients using ICD and CPT codes associated with cardiac arrest and neurological injury, selected per clinical guidelines and expert consensus.[14] Detailed selection criteria and code lists are provided in Supplemental Section S1 and Table S1.

Determining ground truth

A standardized case-classification protocol was developed in collaboration with four board-certified neurologists (M.B.W., E.D.A., S.F.Z., S.S.M.). Cases were labeled positive if documentation indicated both cardiac arrest and coma. To avoid bias, annotators were blinded to algorithmic predictions and hospital of origin during labeling. The full manual annotation procedure is detailed in Supplemental Table S2.

For Model 2, the positive cases from Model 1 served as input for WLST classification. WLST determinations followed a protocol reviewed by two neurologists (M.B.W., E.A.). A case was classified as WLST only if the patient died during hospitalization and death was explicitly attributed to withdrawal of medical care. Deaths from other causes (e.g., brain death) or WLST without death were excluded. Among 938 cases meeting criteria for cardiac arrest and coma (740 [78.9 %] from MGB; 198 [21.1 %] from BIDMC), 444 (47.3 %) were identified as WLST. Time to WLST was approximated as the interval between admission and discharge.

Matrix creation

Features for Model 1 were derived from ICD codes and keywords extracted from clinical notes. During manual annotation, influential terms were recorded as candidate features. Text preprocessing included stemming and negation handling using standard NLP tools. Keywords occurring fewer than 30 times across the entire dataset were excluded.[15] Features were used to construct a binary feature matrix, with each feature coded as "1" if present and "0" if absent. Negated terms were treated as separate features; if both a term and its negation appeared, both were assigned a "1". For additional information see Supplemental Section S2 and Supplemental Table S3.

Model development and validation

Binary feature matrices served as inputs for predictive modeling. Model 1 used 3000 patients split 1500/1500; Model 2 used 938 patients split 469/469. Splits were fixed prior to tuning and testing data were never used for model selection. For external validation, a single model trained on one hospital's training set was applied to the other hospital's holdout set, ensuring no patient overlaps.

We compared random forest and logistic regression models using ICD-only and ICD + keyword feature sets to assess the value of unstructured text. Binary cut-points were chosen by maximizing Youden's J on the training set and applied unchanged to testing and external data. The final model was trained on the entire training set after hyperparameter tuning by cross-validation on training folds.[16] Model optimization procedures are detailed in Supplemental Section S3.

Cross-hospital generalizability analysis

To test generalizability of model 1, a model was trained on BIDMC notes from the Master Training Set (n = 511) and tested on MGB notes from the Master Holdout Set (n = 1011). This was repeated with MGB (n = 989) training and BIDMC (n = 489) testing. This was repeated for model 2 with MGB training (n = 366) and BIDMC testing (n = 104), and with BIDMC training (n = 94) and MGB testing (n = 374). We acknowledge that the available sample sizes for this nested Model 2 analysis (e.g., n = 94 for the BIDMC training set) are below standard recommendations for machine learning derivation. Consequently, this cross-hospital assessment was undertaken as an exploratory analysis to investigate the impact of training size and site-specificity on model performance.

Error analysis

After final model evaluation all misclassified cases were manually reviewed to determine the sources of error. A cohort reconstruction was conducted using 40,000 patients (20,000 patients from each hospital), to estimate the error rate in a randomly selected population.

The probability of classification error (Pe) was calculated using the following formula:

$$Pe = (Pe_{++} \cdot p_{++}) + (Pe_{+-} \cdot p_{+-}) + (Pe_{-+} \cdot p_{-+}) + (Pe_{-} \cdot p_{-})$$

where Pe represents the overall error probability, and p the proportion of cases in each classification group.

Results

Cohort construction

The final study cohort included 3000 unique patients, with 2000 from MGB and 1000 from BIDMC. It contained a relatively higher number of males (57.4 %). The most common race in the cohort was White, followed by Black/African American, Asian, and American Indian or Alaska Native, with 6.8 % of patients identifying as Hispanic. Race was unavailable for 13.3 % of the cohort. See Table 1 for cohort demographic specifics. Supplemental Tables S5 and S6 contain further details about cohort creation.

Table 1 – Patient demographics for MGB and BIDMC. Included are the number of participants, the average age including one standard deviation from the mean race, and ethnicity.

Hospital	ICD+, CPT+	ICD+, CPT−	ICD−, CPT+	ICD−, CPT−
Participants				
MGB	500	500	500	500
BIDMC	250	250	250	250
Average Age (St. Dev.)				
MGB	57.9 (18.9)	59.3 (18.1)	56.5 (21.9)	55.4 (23.6)
BIDMC	62.2 (15.3)	68.7 (14.8)	62.1 (18.3)	60.69 (23.8)
Female (%)				
MGB	166 (33.2 %)	184 (36.8 %)	228 (45.6 %)	256 (51.2 %)
BIDMC	89 (35.6 %)	123 (49.2 %)	110 (44.0 %)	121 (48.4 %)
White (%)				
MGB	329 (65.8 %)	346 (69.2 %)	369 (73.8 %)	378 (75.6 %)
BIDMC	175 (70.0 %)	179 (71.6 %)	195 (78.0 %)	183 (73.2 %)
African American (%)				
MGB	81 (16.2 %)	67 (13.4 %)	54 (10.8 %)	38 (7.6 %)
BIDMC	35 (14.0 %)	36 (14.4 %)	25 (10.0 %)	13 (5.2 %)
Asian (%)				
MGB	10 (2.0 %)	17 (3.4 %)	11 (2.2 %)	22 (4.4 %)
BIDMC	8 (3.2 %)	4 (1.6 %)	8 (3.2 %)	8 (3.2 %)
American Indian or Alaska Native (%)				
MGB	3 (0.6 %)	2 (0.4 %)	1 (0.2 %)	1 (0.2 %)
BIDMC	1 (0.4 %)	0 (0.0 %)	0 (0.0 %)	1 (0.4 %)
Unavailable/Other (%)				
MGB	77 (15.4 %)	67 (13.4 %)	64 (12.8 %)	61 (12.2 %)

(*continued on next page*)

Table 1 (continued)				
Hospital	**ICD+, CPT+**	**ICD+, CPT−**	**ICD−, CPT+**	**ICD−, CPT−**
BIDMC	33 (13.2 %)	31 (12.4 %)	22 (8.8 %)	45 (18.0 %)
Hispanic (%)				
MGB	49 (9.8 %)	50 (10.0 %)	33 (6.6 %)	37 (7.4 %)
BIDMC	7 (2.8 %)	12 (4.8 %)	11 (4.4 %)	6 (2.4 %)

Model performance

Random forest models consistently outperformed logistic regression and were selected for final evaluation. Detailed performance metrics are provided in Supplemental Table S4, with training ROC and PR curves by fold in Supplemental Figs. S1–S3. This final model is a single model trained on all training data after hyperparameter selection. 95 % CIs for binary metrics used Clopper–Pearson exact method; AUROC/AUPRC CIs were derived from 1000 bootstrap resamples. See Table 2 for performance metrics and Fig. 1 for AUROC and AUPRC curves. The model demonstrated high specificity (e.g., 96.1 %) and positive predictive value (PPV) (e.g., 91.4 %), indicating a low false positive rate and a high degree of reliability for positive classifications on the holdout set.

Feature importances

Feature importances were calculated to identify which keywords contribute most to prediction of the target phenotype. Afterwards, model training features were reviewed by experts in neurology (M.B.W, E. D.A., S.F.Z, S.S.M.) to confirm that they were reasonable indicators of the phenotypes (Fig. 2).

External validation

External validation was performed by training a single model on the training cases from one hospital and evaluating it on the holdout cases from the other hospital. We did not recalibrate probability thresholds between sites.

For Model 1 performance remained robust across training/testing hospital pairs. For Model 2, performance was variable: when trained on BIDMC and tested on MGB sensitivity was markedly reduced (sensitivity = 0.171; F1 = 0.290), while the reverse train/test pairing yielded substantially better sensitivity (sensitivity = 0.605). These differences appear driven by (1) limited WLST-positive cases in the BIDMC training set, and (2) differences in documentation language between hospitals. We therefore interpret external validation results as demonstrating feasibility of cross-hospital phenotyping for cardiac arrest but limited generalizability for WLST without larger multi-site training data or site-specific recalibration. Full performance metrics are in Table 3.

Error analysis

An error analysis was performed for each model, and incorrect predictions were categorized based on the most common causes. Supplemental Section S4 for the primary sources of error, listed in a decreasing order of prevalence.

Population error rates

We estimated the error rate of model 1 in a reconstructed cohort, in which the prevalence of cardiac arrest matches that in the general

Table 2 – AUROC, AUPRC, and F1 scores for evaluating performance, and sensitivity, specificity, positive predictive value (PPV), negative predictive value (NPV), and confusion matrix for evaluating practical utility of final models.

	Model 1	**Model 2**
AUROC (95 % CI)	0.984 (0.979–0.989)	0.992 (0.9857–0.996)
AUPRC (95 % CI)	0.968 (0.958–0.977)	0.991 (0.984–0.996)
F1 (95 % CI)	0.900 (0.880–0.919)	0.955 (0.933–0.973)
Sensitivity: TP/(TP + FN) (95 % CI)	0.879 (0.846–0.907)	0.968 (0.935–0.987)
Specificity: TN/(TN + FP)(95 % CI)	0.966 (0.953–0.976)	0.950 (0.916–0.973)
PPV: TP/(TP + FP) (95 % CI)	0.923 (0.895–0.946)	0.942 (0.903–0.969)
NPV: TN/(TN + FN) (95 % CI)	0.945 (0.929–0.958)	0.972 (0.944–0.989)
Confusion matrix	TP: 420, FP: 35 FN: 58, TN: 987	TP: 212, FP: 13 FN: 7, TN: 246

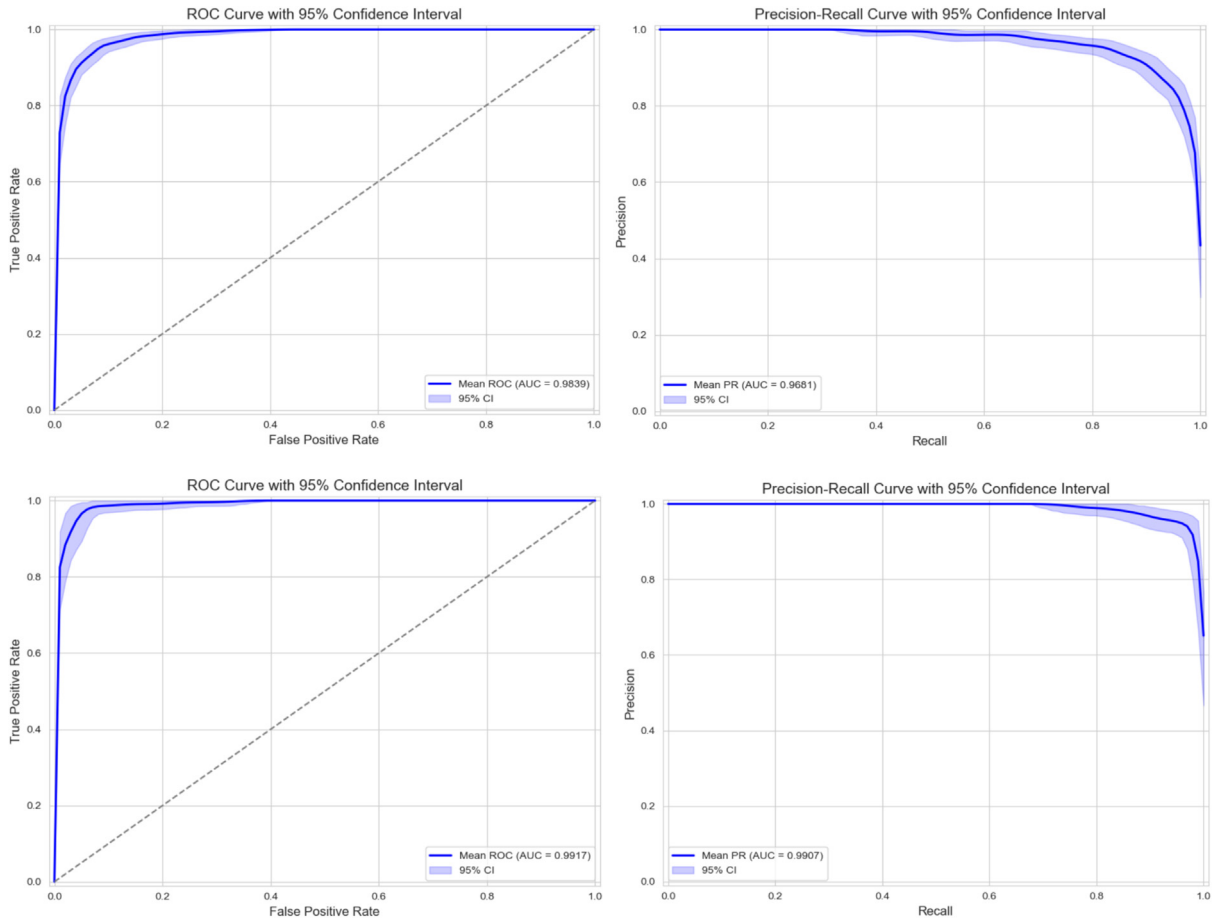

Fig. 1 – AUROC and AUPRC curves from the final models. The solid line is the empirical curve, and the shaded area is the 95 % CI.

hospital population. The estimated error rate of model 1 for MGB patients was 0.51 %, and 0.18 % for BIDMC patients, yielding an overall error rate of 0.34 %. Within model 2, the estimated error rate for BIDMC was 0.39 % and for MGB was 0.03 %, yielding an overall error rate of 0.42 %. See Supplemental Table S6 for the observed error rates for each cohort group (*Pe*), and Supplemental Table S7 for a reconstructed cohort occurrence (*p*).

WLST clinical results

Among the 444 patients who underwent WLST, 189 (42.6 %) died within 3.0 days of admission. The time from documented admission to discharge-death was used as a proxy for time-to-WLST and was 6.65 days (SD 8.63 days), with a range of 0.02–72.76 days. The 25th, 50th, and 75th percentiles of WLST duration were 1.26, 3.94, and 8.52 days, respectively (see Fig. 3).

Discussion

Principal findings

Our study shows that combining structured ICD codes with NLP-extracted keywords yields high accuracy on internal holdout data and provides a practical pipeline for scalable phenotyping. However, cross-site variability, particularly for WLST detection, limits generalizability. By automating what has traditionally required labor intensive chart review, these models correctly identified patients with cardiac arrest, and those who then died of WLST. Beyond cohort assembly, this model opens the door to granular analyses of neuroprognostication practices, facilitating investigations into whether WLST timing and language contribute to self fulfilling prophecies. Moving forward, embedding our pipeline into multicenter EHR platforms, and validating it in larger, more diverse cohorts will be critical to ensure robustness and to inform evidence based end of life quality improve efforts.

Our study demonstrates that combining keyword and ICD code-based features improves EHR phenotyping performance, enhancing our ability to perform retroactive analysis studies. We also found that models based solely on keywords were extremely effective. Adding ICD codes improved classification by only a single additional correct note out of 1500, suggesting that free-text features alone may be sufficient for phenotyping tasks of this nature.

When examining important features, the model prioritized terms describing the direct actions and status changes associated with WLST (e.g., 'CMO', 'expired') over phrases describing the rationale (e.g., 'poor prognosis'), suggesting that documentation of care transitions provides a stronger predictive signal.

Both logistic regression and random forest models performed well, with random forest models performing slightly better. Random forest models often outperform logistic regression in EHR phenotyp-

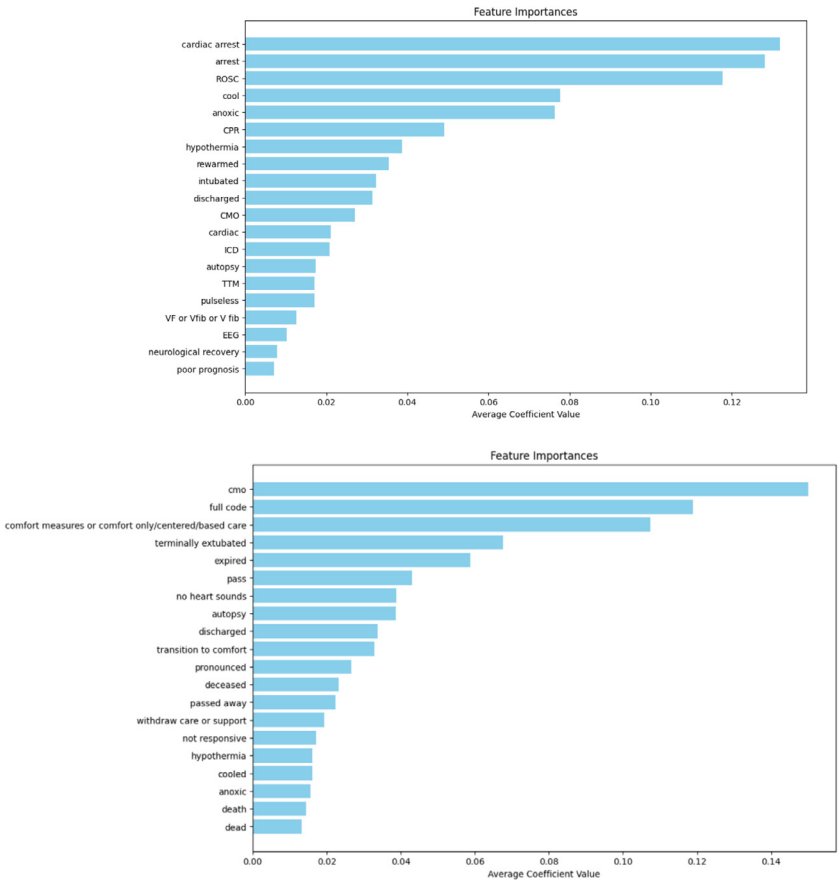

Fig. 2 – Feature importances averaged across cross-validation folds used during training; final importance values were inspected on the single final model. The *x*-axis represents the average coefficient value of each feature, while the *y*-axis lists the most influential features. Higher values indicate greater impact on model predictions. Values were determined using Gini impurity.

ing tasks due to their ability to capture nonlinear relationships.[17] Given the intended use (cohort assembly for retrospective research), we prioritized high specificity to avoid false positive inclusion; users seeking different trade-offs should recalibrate thresholds to their application.

Clinical implications

Our by-hand annotation of the entire cohort showed that 47.3 % of patients who were diagnosed with cardiac arrest and coma died of WLST, with nearly 42 % of those patients receiving WLST within less than 72 h. Current guidelines recommend against WLST before 72 h have passed, and the high percentage of WLST within 72 h suggests that crucial prognostic judgments are being made within this early timeframe. This is a critical window where more objective and accurate prognostic tools could help ensure decisions align with guidelines and prevent premature WLST in patients with recovery potential.[18] Withdrawing life support directly results in the death of the patient, and may represent a self-fulfilling prophecy in many cases. Importantly, these models are not ready for clinical decision support; they are intended to identify retrospective cohorts for research into practice patterns and ethics. However, using this pipeline to study accurate early prognostication could potentially save the lives of patients who have potential for delayed neurological recovery.

Comparison to prior research

HR-based phenotyping has been used to classify acute neurological injury patterns. Unsupervised clustering of post–cardiac-arrest data delineated early brain injury subtypes linked to outcomes in over 2000 patients, and ICD-based algorithms have achieved >90 % precision for identifying out-of-hospital arrests in administrative data.[18,19] Our model advances these efforts by automating phenotyping of cardiac arrest, coma, and WLST, enabling large-scale cohort generation without manual review.

Combining natural language processing (NLP) of clinical notes with structured codes consistently outperforms code-only models. Across 172 ICU phenotypes, text features exceeded structured data in 84 % of cases, and disease-specific studies have shown 15–30 % gains in precision when both are integrated.[9,10] Our hybrid approach, merging ICD codes with curated NLP keywords, achieved AUROCs >0.97, demonstrating high-fidelity phenotyping at scale.

Manual chart review remains the gold standard for identifying WLST but lacks scalability. In one multicenter cohort, ~25 % of in-hospital deaths followed WLST for neurologic reasons.[20] Recent

Table 3 – AUROC, AUPRC, and F1 scores for evaluating performance, and sensitivity, specificity, positive predictive value (PPV), negative predictive value (NPV), and confusion matrix for evaluating practical utility of for external validation step.

	Model 1	
Training set/Testing set	**BIDMC/MGB**	**MGB/BIDMC**
AUROC (95 % CI)	0.968 (0.957–0.976)	0.941 (0.917–0.961)
AUPRC (95 % CI)	0.951 (0.935–0.964)	0.846 (0.784–0.897)
F1 Score (95 % CI)	0.834 (0.804–0.864)	0.703 (0.630–0.766)
Sensitivity: TP/(TP + FN) (95 % CI)	0.757 (0.710–0.799)	0.808 (0.719–0.878)
Specificity: TN/(TN + FP) (95 % CI)	0.966 (0.948–0.978)	0.868 (0.830–0.900)
PPV: TP/(TP + FP) (95 % CI)	0.928 (0.844–0.895)	0.622 (0.535–0.704)
NPV: TN/(TN + FN) (95 % CI)	0.871 (0.844–0.895)	0.944 (0.914–0.965)
Confusion matrix	TP: 283, FP: 22 FN: 91, TN: 615	TP: 84, FP: 51 FN: 20, TN: 334
	Model 2	
AUROC (95 % CI)	0.903 (0.871–0.935)	0.888 (0.818–0.948)
AUPRC (95 % CI)	0.900 (0.857–0.939)	0.838 (0.706–0.927)
F1 Score (95 % CI)	0.290 (0.210–0.234)	0.708 (0.561–0.820)
Sensitivity: TP/(TP + FN) (95 % CI)	0.171 (0.119–0.234)	0.605 (0.434–0.760)
Specificity: TN/(TN + FP)(95 % CI)	0.990 (0.963–0.999)	0.939 (0.852–0.983)
PPV: TP/(TP + FP) (95 % CI)	0.939 (0.798–0.993)	0.852 (0.663–0.958)
NPV: TN/(TN + FN) (95 % CI)	0.560 (0.506–0.614)	0.805 (0.699–0.887)
Confusion matrix	TP: 31, FP: 2 FN: 150, TN: 191	TP: 23, FP: 4 FN: 15, TN: 62

NLP systems have detected code-status discussions (e.g., DNR/DNI, comfort care) with F1 scores >0.90, matching expert review.[21] Our model builds on these advances by automatically identifying WLST events with similar accuracy and minimal manual input.

Limitations & future direction

Limiting data to two Boston hospitals restricted generalizability, necessitating broader validation to capture regional linguistic variation. Our 500-word threshold excluded ~80 % of notes, limiting applicability to detailed summaries. Future work should adapt models to briefer documentation. Critically, the rigorous nested validation produced small training sets (e.g., $n = 94$); the resulting sensitivity drop highlights that this methodology requires substantial training data to capture linguistic variety for robust cross-hospital generalizability. This likely explains the reduced sensitivity observed when Model 2 was externally applied across institutions. Additional limitations include the use of admission-to-death intervals as proxies for WLST timing, potential error propagation from Model 1 to Model 2 due to training on positive cases from Model 1, and the need for periodic retraining as terminology evolves. Finally, future research must distinguish direct causes of WLST to better understand the influence of prognostication on withdrawal decisions.

Conclusion

In summary, our NLP augmented phenotyping framework transforms EHR data into a rapid, reliable means of identifying patients with cardiac arrest, coma, and WLST. By reducing manual abstraction while retaining high accuracy, this approach empowers large scale retrospective studies of neuroprognostication and end of life decision making. Future work will leverage these phenotypes to quantify practice variation, examine the impact of documentation lan-

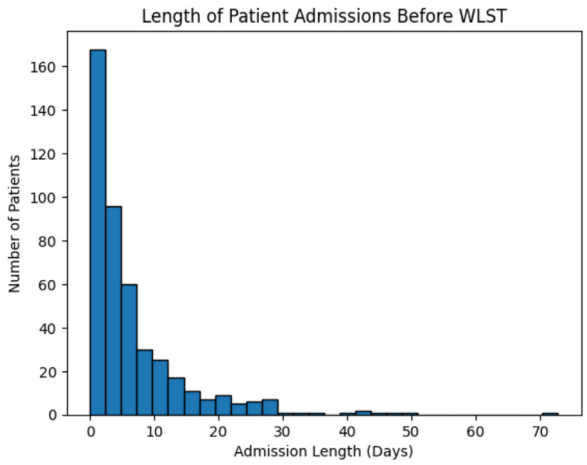

Fig. 3 – Patient admission lengths before WLST. The *x*-axis represents admission length in days, while the *y*-axis shows the number of patients that were admitted before passing away from a withdrawal of care.

guage on WLST timing, and ultimately guide the development of standardized, ethically informed prognostic tools.

Source of support

This work was supported by a grant from the NIH (R01HL161253).

CRediT authorship contribution statement

Catherine Clive: Writing – original draft, Visualization, Validation, Software, Methodology, Investigation, Formal analysis. **Arjun Singh:** Software, Data curation. **Bram Overmeer:** Software, Methodology. **Spencer Boris:** Software, Methodology. **Lydia Peterson:** Software, Methodology. **Jaden Searle:** Software, Methodology. **Greg Hooke:** Writing – review & editing, Software, Methodology. **Niels Turley:** Software, Resources, Data curation. **Marta Fernandes:** Writing – review & editing, Formal analysis, Conceptualization. **Aditya Gupta:** Resources, Data curation. **Manohar Ghanta:** Resources, Data curation. **Valdery Moura Junior:** Resources, Data curation. **S. Mukeriji:** Writing – review & editing, Supervision, Project administration, Methodology, Conceptualization. **Sahar Zafar:** Supervision, Project administration, Methodology, Conceptualization. **Edilberto Amorim:** Writing – review & editing, Validation, Methodology, Conceptualization. **M. Brandon Westover:** Writing – review & editing, Supervision, Methodology, Funding acquisition, Conceptualization. **Haoqi Sun:** Writing – review & editing, Supervision, Software, Project administration, Methodology.

Data availability

The code and data used in this study will be made publicly available at the time of publication through the Brain Data Science Platform (https://bdsp.io/). The code for this project is available at https://github.com/bdsp-core/NAX-CardiacArrest.

Declaration of competing interest

Dr. Westover is a co-founder, scientific advisor, and consultant to Beacon Biosignals and holds a personal equity interest in the company. All other authors declare no competing interests.

Appendix A. Supplementary material

Supplementary material to this article can be found online at https://doi.org/10.1016/j.resuscitation.2025.110919.

Author details

[a]*Department of Neurology, Beth Israel Deaconess Medical Center, Boston, MA 02215, USA* [b]*Brigham Young University, Provo, UT 84602, USA* [c]*Utrecht University, 3584 CS Utrecht, the Netherlands* [d]*Department of Neurology, Massachusetts General Hospital, Boston, MA 02114, USA*[e]*Department of Neurology, Weill Institute for Neurosciences, University of California, San Francisco, San Francisco, CA 94110, USA* [f]*Department of Medicine, Massachusetts General Hospital, Boston, MA, USA*

REFERENCES

1. Tsao CW, Aday AW, Almarzooq ZI, Alonso A, Beaton AZ, Bittencourt MS, Boehme AK, Buxton AE, Carson AP, Commodore-Mensah Y, Elkind MSV, Evenson KR, Eze-Nliam C, Ferguson JF, Generoso G, Ho JE, Kalani R, Khan SS, Kissela BM, Knutson KL, Levine DA, Lewis TT, Liu J, Loop MS, Ma J, Mussolino ME, Navaneethan SD, Perak AM, Poudel R, Rezk-Hanna M, Roth GA, Schroeder EB, Shah SH, Thacker EL, VanWagner LB, Virani SS, Voecks JH, Wang N-Y, Yaffe K, Martin SS, on behalf of the American Heart Association Council on Epidemiology and Prevention Statistics Committee and Stroke Statistics Subcommittee. Heart disease and stroke statistics—2022 update: a report from the American Heart Association. Circulation 2022;145:e153–639. https://doi.org/10.1161/CIR.0000000000001052.
2. Henson T, Rawanduzy C, Salazar M, Sebastian A, Weber H, Al-Mufti F, Mayer SA. Outcome and prognostication after cardiac arrest. Ann NY Acad Sci 2022;1508:23–34. https://doi.org/10.1111/nyas.14699.
3. Marion DW. Coma due to cardiac arrest: prognosis and contemporary treatment. F1000 Med Rep 2009;1:89. https://doi.org/10.3410/M1-89. PMID: 20948689; PMCID: PMC2948325.
4. Kromm J, Davenport A, Elizabeth Wilcox M. Neuroprognostication after cardiac arrest. Chest Crit Care 2024;2(3)100074. https://doi.org/10.1016/j.chstcc.2024.100074. ISSN 2949-7884.
5. Masica A, Collinsworth A. Leveraging electronic health records in comparative effectiveness research. Prescript Excell Health Care Newslett Suppl 2012;1(14) [Google Scholar][Ref list]Available from: http://jdc.jefferson.edu/cgi/viewcontent.cgi?article=1105&context=pehc.
6. Smoller JW. The use of electronic health records for psychiatric phenotyping and genomics. Am J Med Genet Part B 2018;177B:601–12. https://doi.org/10.1002/ajmg.b.32548.
7. Botsis T, Hartvigsen G, Chen F, Weng C. Secondary Use of EHR: data quality issues and informatics opportunities. Summit Transl Bioinform 2010;2010:1–5. PMID: 21347133; PMCID: PMC3041534.
8. Hsu J, Pacheco JA, Stevens WW, Smith ME, Avila PC. Accuracy of phenotyping chronic rhinosinusitis in the electronic health record. Am J Rhinol Allergy 2014;28(2):140–4. https://doi.org/10.2500/ajra.2014.28.4012.

9. Wei W-Q, Teixeira PL, Mo H, Cronin RM, Warner JL, Denny JC. Combining billing codes, clinical notes, and medications from electronic health records provides superior phenotyping performance. J Am Med Inform Assoc 2016;23(1):e20–7.
10. Shivade C, Raghavan P, Fosler-Lussier E, et al. A review of approaches to identifying patient phenotype cohorts using EHR. J Am Med Inform Assoc 2014;21(2):221–30.
11. Mertens M, King OC, van Putten MJAM, et al. Can we learn from hidden mistakes? Self-fulfilling prophecy and responsible neuroprognostic innovation. J Med Ethics 2022;48:922–8.
12. Johnson A, Pollard T, Mark R. MIMIC-III clinical database (version 1.4). PhysioNet; n.d. https://doi.org/10.13026/C2XW26.
13. Johnson AEW, Pollard TJ, Shen L, Lehman LH, Feng M, Ghassemi M, Moody B, Szolovits P, Celi LA, Mark RG. MIMIC-III, a freely accessible critical care database. Sci Data 2016;3:160035.
14. Ehrenstein V, Kharrazi H, Lehmann H, Taylor CO. Obtaining data from electronic health records. Tools and technologies for registry interoperability, registries for evaluating patient outcomes: a user's guide. Addendum 2 – NCBI Bookshelf; 2019. Available from: https://www.ncbi.nlm.nih.gov/books/NBK551878/.
15. Bird S, Klein E, Loper E. Natural language processing with Python. O'Reilly Media Inc.; 2009.
16. Saarela M, Jauhiainen S. Comparison of feature importance measures as explanations for classification models. SN Appl Sci 2021;3:272. https://doi.org/10.1007/s42452-021-04148-9.
17. Junying Wei HC. An interpretable machine learning model for predicting in-hospital mortality in ICU patients with ventilator-associated pneumonia. PLoS One 2025. https://doi.org/10.1371/journal.pone.0316526.
18. Elmer J, Torres C, Aufderheide TP, Austin MA, Callaway CW, Golan E, Herren H, Jasti J, Kudenchuk PJ, Scales DC, Stub D, Richardson DK, Zive DM. Association of early withdrawal of life-sustaining therapy for perceived neurological prognosis with mortality after cardiac arrest. Resuscitation 2016;102:127–35. ISSN 0300-9572.
19. Elmer J, Coppler PJ, May TL, et al. Unsupervised learning of early post arrest brain injury phenotypes. Resuscitation 2020;153:154–60.
20. Hwang U, Richardson LD, Wang NE, et al. Emergency department processes of care: a natural language processing approach to identify discussions of end of life care. Acad Emerg Med 2014;21(1):33–40.
21. Lee SJ, Lopez L, Rhee Y, Burns JP. Automated detection of code status orders from clinical notes: a natural language processing approach. J Biomed Inform 2018;85:65 72.