Assessment of a zero-shot large language model in measuring documented goals-of-care discussions

Robert Y. Lee, MD, MS [1,2]
Kevin S. Li, MS [1,3]
James Sibley, BS [1,4]
Trevor Cohen, MBChB, PhD [1,3]
William B. Lober, MD, MS [1,3,4,5]
Danae G. Dotolo, PhD, MSW, MA [1,2]
Erin K. Kross, MD [1,2]

[1] Division of Pulmonary, Critical Care, and Sleep Medicine, University of Washington, Seattle, USA
[2] Cambia Palliative Care Center of Excellence at UW Medicine, University of Washington, Seattle, USA
[3] Division of Biomedical and Health Informatics, Department of Biomedical Informatics and Medical Education, University of Washington, Seattle, WA
[4] Department of Biobehavioral Nursing and Health Informatics, University of Washington, Seattle, WA
[5] Department of Global Health, University of Washington, Seattle, WA

Address correspondence to:

Robert Y. Lee, MD, MS, Assistant Professor of Medicine
Division of Pulmonary, Critical Care, and Sleep Medicine
Cambia Palliative Care Center of Excellence at UW Medicine
University of Washington School of Medicine, Seattle, WA

Address: Harborview Medical Center, 325 9th Ave., Box 359762, Seattle, WA 98104
Phone: 206-744-6359; fax: 206-744-8584; e-mail: rlee06@uw.edu

Word count: 3632
References: 49
Tables: 3
Figures: 1
Appendices: 1

ABSTRACT

Context: Goals-of-care (GOC) discussions and their documentation are important process measures in palliative care. However, existing natural language processing (NLP) models for identifying such documentation require costly task-specific training data. Large language models (LLMs) hold promise for measuring such constructs with fewer or no task-specific training data.

Objective: To evaluate the performance of a publicly available LLM with no task-specific training data (zero-shot prompting) for identifying documented GOC discussions.

Methods: We compared performance of two NLP models in identifying documented GOC discussions: Llama 3.3 using zero-shot prompting; and, a task-specific BERT (Bidirectional Encoder Representations from Transformers)-based model trained on 4,642 manually annotated notes. We tested both models on records from a series of clinical trials enrolling adult patients with chronic life-limiting illness hospitalized over 2018-2023. We evaluated the area under the receiver operating characteristic curve (AUC), area under the precision-recall curve (AUPRC), and maximal F_1 score, for both note-level and patient-level classification over a 30-day period.

Results: In our text corpora, GOC documentation represented <1% of text and was found in 7.3-9.9% of notes for 23-37% of patients. In a 617-patient held-out test set, Llama 3.3 (zero-shot) and BERT (task-specific, trained) exhibited comparable performance in identifying GOC documentation (Llama 3.3: AUC 0.979, AUPRC 0.873, and F_1 0.83; BERT: AUC 0.981, AUPRC 0.874, and F_1 0.83).

Conclusion: A zero-shot large language model with no task-specific training performed similarly to a task-specific trained BERT model in identifying documented goals-of-care discussions. This demonstrates the promise of LLMs in measuring novel clinical research outcomes.

Abstract word count: 250 / 250

KEY MESSAGE

This article reports the performance of a publicly available large language model with no task-specific training data in measuring the occurrence of documented goals-of-care discussions from electronic health records. The study demonstrates that newer large language AI models may allow investigators to measure novel outcomes without requiring costly training data.

Key Message word count: 50 / 50

Keywords: Natural language processing, large language models, artificial intelligence, electronic health records, goals of care

Running title: Large language models to measure GOC discussions

Introduction

Advances in natural language processing (NLP) hold great promise for allowing researchers to measure previously inaccessible constructs from unstructured electronic health records (EHR). Because many important clinical constructs in EHRs are linguistically complex and difficult to capture with rule-based or keyword searches,[1] researchers have used machine learning (ML) models trained on manually curated and labeled datasets—a process called *supervised learning*—to extract complex constructs from EHR text.[2] However, acquiring and labeling training datasets for supervised learning can be prohibitively expensive, especially for rare constructs. This problem, which has been termed the "annotation bottleneck," is an important barrier for the use of supervised learning in clinical applications of NLP.[2]

Many recent advances in NLP have been driven by large language models (LLMs), a newer class of ML models pre-trained on vast and diverse corpora of unlabeled text. Through their pre-training, LLMs acquire a broad representation of linguistic structures and relationships that allow them to perform a wide range of NLP tasks with little or no task-specific training data.[3] Recent generations of LLMs have gained widespread attention through their deployment as artificial intelligence (AI) chatbots, which at least imitate human-like knowledge in many areas.[4] As such, there has been growing interest in using LLMs to measure linguistic constructs without the need for large amounts of training data. In these approaches, LLMs are prompted to classify text with minimal or no task-specific training data (referred to as *few-shot* and *zero-shot learning*, respectively), instead relying on careful prompting in conjunction with the model's "out-of-the-box" pre-training to apply previously learned patterns to the task at hand (**Table 1**).[5]

Goals-of-care discussions are an important process measure in palliative care research, and likely mediate the delivery of patient-centered, goal-concordant care.[6–8] However, timely and high-quality assessment of patients' goals of care remains a significant shortcoming in health systems.[9–11] Additionally, because goals-of-care discussions are frequently documented in unstructured text such as progress notes or other inconspicuous locations, documentation of historical discussions can be difficult for clinicians to unearth during clinical care.[12–14] This has led to interest in using NLP to measure goals-of-care discussions.[15–20] To date, the use of NLP to identify this linguistically complex outcome has facilitated at least two clinical trials;[19–22] and, while current clinical EHRs rely on either search-based approaches or the use of semi-structured documentation,[12,23] real-time ability for clinicians to efficiently locate goals-of-care documentation seems likely to improve delivery of goal-concordant care.[24]

Our research group has studied the use of NLP to identify EHR-documented goals-of-care discussions for patients with serious illness. In our previous work, a supervised learning model based on BERT (Bidirectional Encoder Representations from Transformers)[25,26] demonstrated promising performance in identifying goals-of-care documentation for hospitalized patients, with an area under the receiver operating characteristic curve [AUC] of 0.962 and area under the precision-recall curve [AUPRC] of 0.824 for adjudicating individual notes.[19] However, acquiring the task-specific training and validation data for this model required hundreds of human abstractor-hours,[19] limiting our ability to adapt this model toward measuring other constructs or conducting larger trials.

In a recent report, a zero-shot LLM was used to identify advance care planning and goals of care documentation for outpatient oncology patients.[18,27] To our knowledge, no study has yet evaluated the use of zero-shot LLMs to identify goals-of-care discussions in a general population of seriously ill hospitalized patients against manually adjudicated standards. To explore the potential of modern LLMs for current and future research tasks, we evaluated the

performance of a publicly accessible LLM using zero-shot prompting to identify documented goals-of-care discussions for hospitalized patients with serious illness.

Methods

This diagnostic study evaluated the zero-shot performance of Llama 3.3 (Meta AI, ai.meta.com), a publicly available pre-trained general-purpose LLM, in predicting the presence of documented goals-of-care discussions in EHR notes of hospitalized patients with serious illness. We compared Llama predictions against gold-standard results from manual abstraction, and also compared Llama performance against the benchmark of a previous BERT-based model that was fitted using supervised ML to manually-abstracted training data.[19] All study procedures were approved by the University of Washington Human Subjects Division Institutional Review Board (STUDY00011002).

Data sources

There are three datasets used in this trial (**Table 2**): (1) a pilot set, comprising 4,642 EHR notes from 150 patients enrolled in a pilot trial of a patient- and clinician-facing communication-priming intervention;[19,28] (2) the Trial 1 test set, comprising 2,974 EHR notes from a sample of 160 of 2,512 patients enrolled in a pragmatic trial of a clinician-facing-only communication-priming intervention;[21,29] and, (3) the Trial 2 held-out test set, comprising 11,574 EHR notes from 617 patients enrolled in a comparative-effectiveness trial of two communication priming interventions.[29] The detailed inclusion criteria for patients in each dataset have been previously reported.[19,21,28,29] Briefly, all patients were hospitalized at any of 2-3 study hospitals in the UW Medicine health system, Seattle, WA; were of advanced age (≥ 80 years); or, were of age 55 years or older with one or more chronic life-limiting illnesses as defined by diagnosis codes. Each dataset consisted of all notes for the sampled patients collected over similar time frames that were authored by attending and trainee physicians, advance practice providers, and subinterns. Patients in the Trial 1 test set were randomly selected from all enrolled participants on the date of an interim analysis in February 2021, with intentional enrichment for patients with Alzheimer disease and related dementias (ADRD; 80/160 [50%]) to support an interim subgroup analysis in the parent trial. (The Trial 1 test set is slightly larger here than in previous reports due to the inclusion of notes preceding parent trial randomization and notes from one patient whose records were inadvertently sampled from the wrong index admission.) The interventions tested in all three trials (variants of the Jumpstart Guide) did not alter existing clinician EHR documentation processes or templates for documentation of goals-of-care discussions.[21,28,29]

By definition, the zero-shot Llama 3.3 LLM had no task-specific training dataset, and all LLM prompt engineering and hyperparameter selection was conducted using notes drawn from the pilot set. The Llama model was tested on both the Trial 1 test set and the Trial 2 held-out test set. The reference BERT-based model was trained and validated on the pilot set, and subsequently evaluated on the Trial 1 test set as previously reported.[19] For this study, the same model was also tested on the Trial 2 held-out test set.

Natural language processing

Llama 3.3 is a 70-billion-parameter generative LLM with a 128,000-token context length, developed and pre-trained by Meta AI on an undisclosed collection of public online data. (*Tokens* are an NLP term for words and subwords that form the smallest unit of NLP analysis; the *context length* of a model defines the largest text passage it can analyze at once.) Most EHR notes may be analyzed in their entirety within Llama 3.3's context length; notes exceeding this context length were partitioned evenly and assigned the highest predicted probability of

their constituent partitions. Notes were preprocessed using regular expressions to remove common headers, footers, allergy and medication sections, and a palliative care quality improvement checklist consisting only of discrete data. To predict the probability of documented goals-of-care discussions in each candidate note or partition, we prompted a newly initialized Llama 3.3 model with our operational definition and the candidate text (**Appendix 1**). We then examined the raw scores (logits) underlying the model's generated response, which were transformed into probabilities that reflect the model's confidence in predicting an affirmative ("Yes") vs. negative ("No") response.

The reference BERT-based model for identifying documented goals-of-care discussions[19] is an instance of BioClinicalBERT[26] that was trained by our team on labeled notes from the pilot set (**Table 2**). BioClinicalBERT is a publicly-available customization of $BERT_{BASE}$ (Bidirectional Encoder Representations from Transformers—base model),[25] a 110-million-parameter model with a 512-token context length developed and pre-trained by Google Research (google.com) on the full text of English Wikipedia and BookCorpus (a dataset of 11,000 unpublished books).[30] BioClinicalBERT was further pre-trained by Alsentzer *et al* on PubMed abstracts, PubMed Central manuscripts, and EHR notes from the MIMIC III database.[26,31] Notes were partitioned into *BERT passages* of ≤ 512 BERT tokens over common whitespace patterns; the BERT-predicted probability for each note was defined by the highest predicted probability of its constituent passages.

Llama 3.3 and other recent Llama models have been released by Meta AI under a licensing agreement and use policy that allows for free non-commercial use by the public. To ensure privacy and prevent model developers from training on protected health information, we ran Llama locally on a secure in-house server equipped with a 16-core AMD Epyc 9124 processor (Advanced Micro Devices, amd.com), a single Nvidia RTX A6000 graphics processing unit (GPU) (Nvidia Corporation, nvidia.com), and 48 gigabytes of video random-access memory (VRAM). Both BioClinicalBERT and the original BERT family of models have been made open-source and publicly available by their respective developers.[32,33]

Adjudication of ground truth

For this study, goals-of-care discussions were defined as discussion of the overarching aims of medical care for a patient[34] that extended beyond routine code status discussions or standalone citations of past advance care planning documents. This pragmatic definition was designed to capture typical documentation of discussions by both specialty palliative care and other clinicians, while excluding routine admission code status discussions or automated listings of advance care planning documents on file.[21,35] The definition was operationalized using a medical record abstraction manual that has been previously published.[19]

Ground-truth for BERT passage- and note-level presence or absence of EHR-documented goals-of-care discussions was determined for all notes in the pilot set and Trial 1 test set using manual human chart abstraction, as previously described.[19] During these abstraction efforts, abstractors met regularly as a group to discuss their findings, and instances of disagreement were resolved by consensus. Quality assurance practices, such as co-review of passages by senior research coordinators and investigators, were applied at the passage level.

The Trial 2 held-out test set, which included 617 patients—substantially more than the pilot (150 patients) and Trial 1 (160 patients) test sets—was adjudicated for presence and timing of first human-confirmed documented goals-of-care discussion within 30 days of randomization using BERT NLP-screened human abstraction. In this approach, the BERT model, which was trained on the pilot set and previously tested on the Trial 1 test set, was applied to the Trial 2 held-out

test set using a highly-permissive (i.e., sensitive) screening threshold to select passages for human adjudication for goals-of-care content. This approach enabled measurement of a much larger number of patient-level outcomes than would be feasible in a manually-abstracted validation sample. For adjudicating the held-out test set, the screening threshold was set to the 98.5th percentile of BERT-predicted passage-level probabilities for goals-of-care content, resulting in 8,952 of 559,596 BERT passages (1.6%) screening positive; 3,718 of 11,574 notes (32%) contained at least one screen-positive passage, and 533 of 617 patients (86%) had at least one screen-positive passage. Evaluation of the selected BERT screening threshold in the Trial 1 test set yielded 99.3% note-level sensitivity (95%CI 97.6%, 99.9%) and 100% patient-level sensitivity (one-sided 97.5%CI 97.7%, 100%).[19] Because not all BERT-positive notes in the Trial 2 held-out test set were adjudicated by humans, note-level performance in the Trial 2 held-out test set was evaluated using a 2,136-note case-control sample of all notes containing human-confirmed goals-of-care content (304 notes) and up to 3 randomly selected notes per patient without goals-of-care content (i.e., with BERT score less than 98.5th percentile; 1,832 notes). In addition to the same quality assurance practices applied during abstraction of the pilot and Trial 1 test sets, Trial 2 records coded as positive or possibly-positive by abstractors were further co-reviewed by a single clinician (E.K.) in the context of the patient's chart to ensure consistency of the outcome measure. Patient-level performance in the Trial 2 held-out test set was evaluated by comparing NLP predictions for all 11,574 notes against patient-level results of NLP-screened human abstraction for all 617 patients.

Statistical analysis

We evaluated the performance of each language model using receiver operating characteristic (ROC) curve analysis, which compares sensitivity and specificity across all discrimination thresholds; precision-recall (PR) analysis, which compares positive predictive value (PPV; or, precision) and sensitivity (recall) across all discrimination thresholds;[36] and, evaluation of the F_1 score, a metric commonly used to describe the predictive performance of information retrieval systems that is defined as the harmonic mean of PPV and sensitivity (range, 0 to 1). ROC and PR analyses were evaluated both graphically and also by their respective areas-under-the-curve, with AUC indicating area under the ROC curve, and AUPRC indicating area under the precision-recall curve. As F_1 is a threshold-dependent measure, we reported the highest observed F_1 score across all discrimination thresholds. In all analyses, patient-level probability was defined as the maximal NLP-predicted probability of all constituent notes or passages; and, patient-level ground truth was defined as the union of ground truth for all constituent notes. Statistical analyses were performed using Stata/MP, version 19.0 (StataCorp LLC, stata.com). AUC and AUPRC were calculated using cubic splines.

Results

The composition, inclusion criteria, and adjudication of ground truth for each dataset are described in **Table 2**. In each dataset, there were between 295 and 340 notes containing goals-of-care discussions, and between 1,832 and 4,302 notes without goals-of-care discussions. In the Trial 1 test set, 59 of 160 patients (37%) had at least one goals-of-care discussion between index admission and 30 days post-randomization. In the Trial 2 held-out test set, 163 of 617 patients (26%) had at least one goals-of-care discussion between index admission and 30 days post-randomization.

In comparing Llama 3.3 and BERT NLP predictions against manual or NLP-screened manual abstraction in the Trial 1 test set and Trial 2 held-out test set, the zero-shot Llama model consistently demonstrated highly comparable performance to the reference BERT model that

had been trained on the pilot set. **Figure 1** shows ROC and precision-recall curves for note- (panel A) and patient-level (panel B) classifiers across models (Llama vs. BERT) and datasets, and **Table 3** shows areas under the ROC and precision-recall curves as well as maximum observed F_1 scores for each model and dataset. In the Trial 2 held-out test set, the area under the ROC curve for note-level analysis was 0.979 for Llama, and 0.981 for BERT; and, the area under the precision-recall curve for note-level analysis was 0.873 for Llama, and 0.874 for BERT. The maximum observed note-level F_1 score in the held-out test set was 0.83 for both Llama and BERT.

Notably, we observed higher performance in the Trial 2 held-out test set for both Llama and BERT models than in the Trial 1 test set. The difference in performance between datasets (**Figure 1**, red vs. blue lines) exceeded differences in performance between models (**Figure 1**, dark vs. light lines).

Discussion

In this report, we demonstrated that a zero-shot LLM with no task-specific training can achieve comparable performance to a task-specific supervised learning BERT model trained on thousands of manually labeled EHR notes to identify documented goals-of-care discussions.

Although it hardly seems noteworthy that a 70-billion-parameter LLM can achieve comparable performance to a 110-million parameter BERT model, the ability to measure linguistically complex outcomes using a zero-shot prompting strategy—which is only possible with LLMs—is of great practical importance to clinical researchers. The training data alone for the referent BERT model was acquired over the course of 287 human abstractor-hours over a 4-month period—a ~$20,000 investment to measure a single linguistic construct that must be repeated for any related constructs of interest, or even for different operational definitions of the same construct.[19] This cost barrier poses significant challenges for researchers interested in measuring multidimensional constructs for which definitions are likely to differ appropriately from study to study.[37,38,35] In contrast, iterative development of an LLM prompt is a faster process that may be quickly replicated or adapted to different constructs, allowing investigators to pivot toward different measures and outcomes between studies while incurring minimal repeated costs. This is especially important for linguistically complex outcomes such as goals-of-care discussions, which are difficult to measure using rule-based methods.[12] The ability to identify linguistically complex constructs at scale also has implications for mixed-methods researchers, in that measures that were previously unquantifiable may be quantified.[39–42]

Despite their advantages, zero-shot methods have their disadvantages as well. First, zero-shot prompting relies heavily on the model's pre-training from its developers. This dependency may introduce instability and irreproducibility as changes between model versions introduced by the developer may alter the model's response to a given prompt in unpredictable ways. Second, LLMs are computationally resource-intensive, often requiring upgraded hardware to deploy at scale. The inefficiency of these models may limit their use cases, at least until LLM-capable hardware or institutionally-approved secure cloud environments for their use are widely available. Active efforts in the machine learning community to shrink and streamline LLMs may also help improve their efficiency.[43] Third, LLMs suffer from a "black box" phenomenon, in which the reasons for a model's predictions are opaque to their operators. However, this limitation—which is shared by nearly all ML models—is potentially easier to overcome with LLMs than with previous-generation models such as BERT, as prompting strategies that ask LLMs to follow specific patterns of logical reasoning have been shown to both be feasible to implement and also a means to generate more robust model output.[44] LLMs also contain embedded biases that

arise from their training data.[45] However, prompting strategies that guide LLMs through cognitive processes are an intriguing potential solution to reduce biases in LLM responses, a strategy that is not available to previous-generation models.[46]

Performance of both models was similar within each dataset, but differed noticeably between the Trial 1 test set and the Trial 2 held-out test set (**Figure 1**). Although BERT was itself used to screen records for human ground-truth adjudication in the Trial 2 dataset, the screening threshold was exceedingly permissive (i.e., sensitive), and additionally, any such bias should not affect the performance of the zero-shot Llama 3.3 model. We suspect that observed differences in performance between datasets are probably attributable to differences between the datasets such as label quality and data curation procedures, differences in patient population, or secular trends in documentation. We believe this observation highlights the importance of evaluating NLP performance within specific use contexts, as both goals-of-care communication practices and their associated documentation practices are likely to differ between patient populations and clinical environments.

LLMs have many potential applications in serious illness communication. Beyond supporting clinical research, LLMs may assist clinicians by reviewing prior goals-of-care discussions, summarizing illness trajectories, identifying patients who may need serious illness conversations, supporting documentation, and enabling quality monitoring. Looking ahead, more advanced models with patient-specific inputs could even serve as adjunctive communication tools tailored to individual patients. However, realizing these potential applications will require overcoming both technical and implementation barriers. On the technical side, LLMs demand computational resources that often exceed current clinical hardware, making real-time chart-wide analysis impractical at the point of care. These challenges are likely surmountable through deeper EHR-cloud integration, advances in model efficiency, and adaptive AI systems that route tasks between lightweight and resource-intensive models. In contrast, implementation barriers will be far harder to address. Clinical workflows are already burdened by high cognitive load, and serious illness communication is further strained by high stakes, strong emotions, deficits in trust and trustworthiness, inequities in care delivery, and resistance to impersonal technologies. Successfully applying LLMs in this context will therefore require disciplined, methodical development that includes patient- and clinician-centered design, robust attention to safety and equity, and meaningful engagement of all stakeholders—practices that are too often overlooked in EHR development. The *potential* of LLMs in this domain is very real, but translating promises into practice hinges less on computational power and more on human factors, workflow design, and stakeholder trust.

Our study has important limitations. First, we evaluated a single operational definition of a single linguistically complex outcome. While much of the promise of this work lies in measuring related constructs, our findings and implications may not apply to all linguistic constructs. Even within the community of investigators interested in measuring goals-of-care content, other operational definitions of the same construct are more amenable to measurement using rule-based and other methods.[15,16,20] Differences in the performance of rule-based methods across these studies[12,20] illustrates how minor differences in operational definitions (e.g., the inclusion vs. exclusion of routine code status discussions) can lead to important differences in the choice of best measurement tools. Second, all data were collected from a series of randomized trials of related interventions performed at a single multi-hospital health system in Washington State, which limits generalizability outside of this setting. Although this study does not account for the trial interventions, our understanding of the interventions (which are all low-touch communication prompting interventions), as well as qualitative work in progress, suggest that the interventions had very little systematic effect on the location or linguistic structure of

clinicians' goals-of-care documentation. Third, our held-out test set was incompletely adjudicated by humans due to its sheer size. However, we feel that the high expected sensitivity of our adjudication method for the held-out test set lends advantages in generalizability that far outweigh limitations in fidelity. This tension between scale and fidelity has long been known to the NLP community, and we anticipate that navigating this tension will continue to be an area of great interest in the intersection of NLP and epidemiology.[47–49]

Conclusion

In measuring documented goals-of-care discussions for hospitalized adults with serious illness, a zero-shot Llama 3.3 large language model with no task-specific training achieved comparable classification performance to a task-specific supervised-learning BERT model that had been trained on thousands of manually labeled EHR notes. The advent of large language models and their facilitation of zero-shot methods have important implications for clinical and health services researchers interested in leveraging the EHR to measure linguistically complex exposures and outcomes.

TABLES

Table 1. Comparison of supervised, few-shot, and zero-shot machine learning.

Learning approach	Description [a]	Examples and analogies
Supervised learning	Model is trained ("fine-tuned") on a labeled dataset, mapping input data to the correct output by minimizing error between predicted and actual labels.	▪ ML model to distinguish between dogs from cats trained on 1,000 pictures of dogs and 1,000 pictures of cats, each labeled accordingly ▪ ML model to identify documented GOC discussions trained on 5,000 EHR notes, of which 300 are labeled as containing GOC discussions
Few-shot learning	Pre-trained model is fine-tuned to, or prompted with, a small number of labeled examples, relying on the model's pre-training to apply previously learned patterns to a new task. Examples need not cover full range of variability or noise expected in the target task, as the model leverages prior learning to generalize effectively.	▪ ML model with generalized pre-training fine-tuned on just 6 pictures of dogs and cats to distinguish between them ▪ A child learning to distinguish airplanes from helicopters after being shown 5 examples from a picture book ▪ LLM prompted to identify documented GOC discussions after being given a small number of examples
Zero-shot prompting	Model is prompted to make predictions without any task-specific examples, relying entirely on the model's pre-training to generalize to the new task.	▪ Vision-language model with generalized pre-training asked to distinguish between dogs and cats, with no fine-tuning or task-specific examples ▪ Resident physician asked to characterize an unusual chest x-ray finding ▪ LLM prompted to identify documented GOC discussions by definition alone, with no provided examples

Abbreviations: ML, machine learning; LLM, large language model; GOC, goals of care.

[a] In machine learning, *pre-training* refers to the initial process of training a model on a large, general dataset to learn patterns, representations and features that are broadly applicable across tasks. Pre-training is often performed by the model's creators. Whereas, *fine-tuning* refers to subsequent adjustment of a pre-trained model on a smaller, task-specific dataset, often performed by those adapting the model for a particular use.

Table 2. Datasets used in this study.

	Pilot set	Trial 1 test set	Trial 2 held-out test set
Description of patients, and enrollment procedure	All 150 patients enrolled in a 2-hospital pilot trial with informed consent over Nov. 2018 to Feb. 2020 [28]	Sample of 160 patients enrolled in a 2,512-person 3-hospital pragmatic trial under a waiver of informed consent, enriched for ADRD (80 of 160 [50%]), over Apr. 2020 to Mar. 2021 [21,29]	All 617 patients enrolled in a 3-hospital comparative-effectiveness trial with informed consent over Jul. 2021 to Nov. 2023 [29] (trial results not yet published)
Description of EHR notes	4,642 notes from index admission to discharge	2,974 notes from index admission to 30 days post-randomization	11,574 notes from randomization to 30 days post-randomization
Adjudication of ground truth	Manual whole-chart abstraction by human reviewers, with regular quality assurance applied at the passage level	Manual whole-chart abstraction by human reviewers, with regular quality assurance applied at the passage level	BERT NLP-screened human abstraction of passages scoring over 98.5th percentile, [a] with principal investigator co-review of all positive abstractions by patient
Prevalence of GOC discussions	0.2% of BERT segments; 340 / 4,642 (7.3%) notes; 34 / 150 (23%) patients	0.4% of BERT segments; 295 / 2,974 (9.9%) notes; 59 / 160 (37%) patients	304 / 2,136 notes in case-control sample; [b] 163 / 617 (26%) patients
Role of dataset for each NLP model [c]			
Llama 3.3 LLM (zero-shot prompt)	Development	Testing	Held-out testing
BERT (supervised ML)	Training and validation	Testing	Held-out testing

Abbreviations: ADRD, Alzheimer disease and related dementias; EHR, electronic health record; BERT, Bidirectional Encoder Representations from Transformers; GOC, goals of care; NLP, natural language processing; LLM, large language model; ML, machine learning.

[a] In the Trial 1 test set, this screening threshold corresponded with 99.3% note-level sensitivity (95%CI 97.6%, 99.9%) and 100% patient-level sensitivity (one-sided 97.5%CI 97.7%, 100%).

[b] Case-control sample of notes in the Trial 2 held-out test: cases, 304 notes with human-confirmed GOC content; controls, up to three randomly sampled GOC-negative notes per patient. Prevalence in sample (14%) is not expected to represent source data.

[c] In supervised machine learning, “training” refers to fitting a model to labeled data; “validation” refers to tuning model hyperparameters during development; and, “testing” refers to evaluating the performance of the final model after all training and development is complete. A “held-out test set” is a special test set that is kept completely separate from all model development, minimizing possibility of indirect leakage or bias. Because zero-shot prompting does not fit a model to any labeled data, the pilot set is referred to as a “development” set for the Llama model.

Table 3. Performance metrics of zero-shot Llama 3.3 vs. supervised-learning BERT models in identifying documented goals-of-care discussions for hospitalized patients with serious illness.

Dataset	NLP model (machine learning approach)	AUC	AUPRC	Maximal F_1
Note-level classification				
Trial 1 test set (n=2,974)	Llama 3.3 (zero-shot ML)	0.961	0.795	0.74
	BERT (supervised ML)	0.962	0.789	0.74
Trial 2 held-out test set (sample of n=2,136) [a]	Llama 3.3 (zero-shot ML)	0.979	0.873	0.83
	BERT (supervised ML)	0.981	0.874	0.83
Patient-level classification over ~30-day period				
Trial 1 test set (n=160)	Llama 3.3 (zero-shot ML)	0.912	0.885	0.80
	BERT (supervised ML)	0.898	0.833	0.80
Trial 2 held-out test set (n=617) [a]	Llama 3.3 (zero-shot ML)	0.977	0.955	0.89
	BERT (supervised ML)	0.981	0.952	0.89

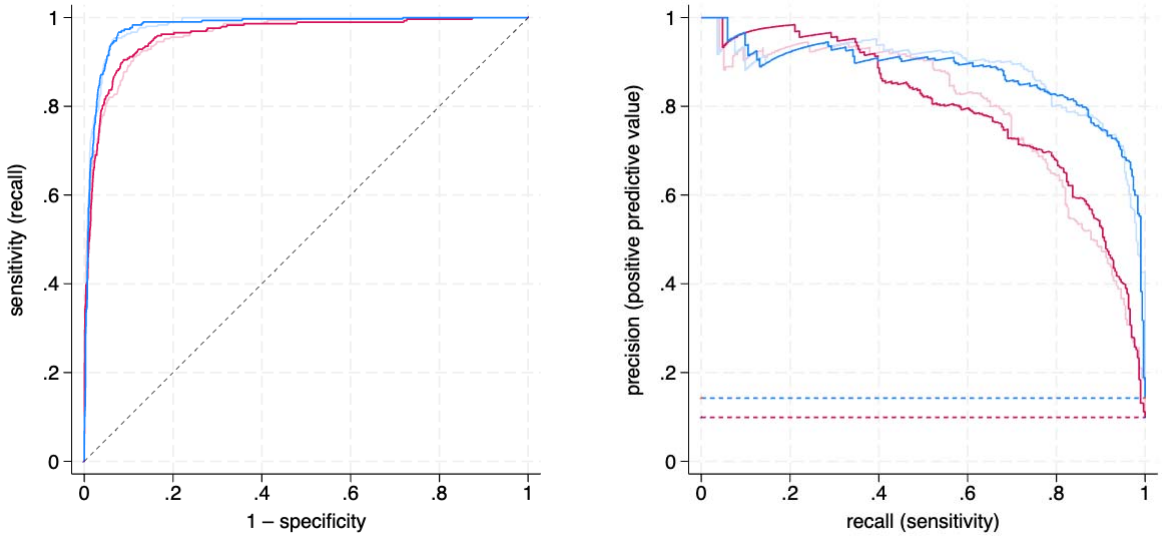
Abbreviations: AUC, area under the receiver operating characteristic [ROC] curve; AUPRC, area under the precision-recall curve; BERT, Bidirectional Encoder Representations from Transformers; ML, machine learning; NLP, natural language processing. The F_1 score is defined as the harmonic mean of positive predictive value (precision) and sensitivity (recall); the maximal F_1 is the highest observed F_1 across all discrimination thresholds.

[a] The Trial 2 held-out test set contained 11,574 notes, of which 2,136 with known true values were sampled for evaluation of note-level performance. Patient-level performance was evaluated by comparing NLP analysis of all 11,574 notes against the NLP-screened human-abstracted gold standard for all 617 patients.

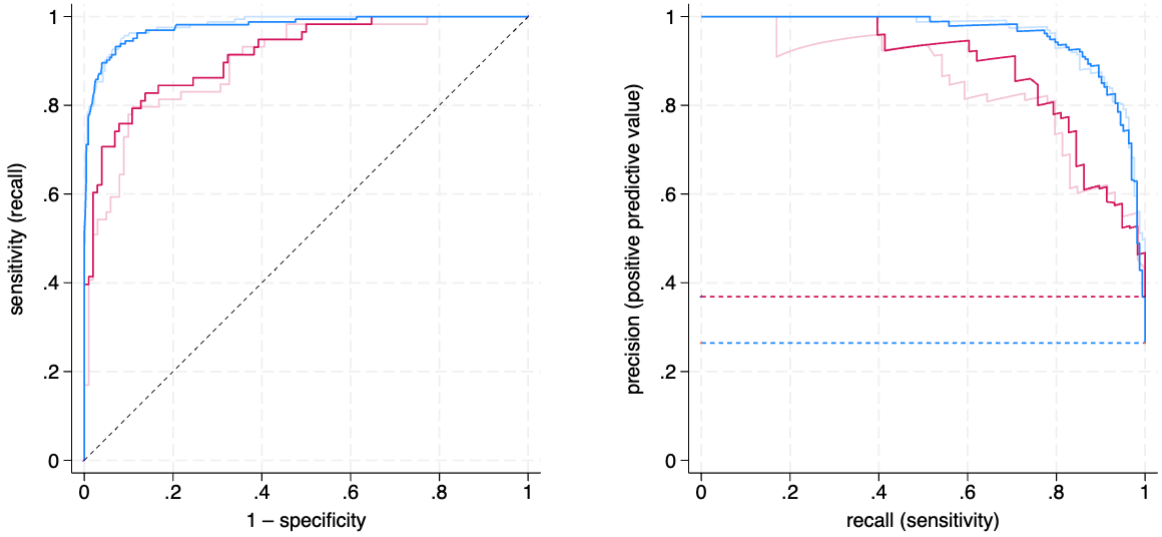
FIGURES

Figure 1. Performance of zero-shot Llama 3.3 vs. supervised-learning BERT models in detecting documented goals-of-care discussions for hospitalized patients with serious illness.

A. Note-level classification



B. Patient-level classification over ~30-day period



Trial 1 test set, Llama 3.3 (zero-shot ML)
Trial 1 test set, BERT (supervised ML)
Trial 2 held-out test set, Llama 3.3 (zero-shot ML)
Trial 2 held-out test set, BERT (supervised ML)
- - - Non-discriminating classifiers (colored when dataset-specific)

Acknowledgements

Author contributions: Dr. Lee had full access to all of the data in the study and takes responsibility for the integrity of the data and the accuracy of the data analysis.

Funding/Support: This work was supported by the National Institute on Aging (R01AG062441), the National Heart, Lung, and Blood Institute (K23HL161503, K12HL137940, T32HL125195), the National Palliative Care Research Center, and the Cambia Health Foundation. Infrastructure support was provided by the Institute of Translational Health Science (National Center for Advancing Translational Sciences, UL1TR002319).

Role of the Funder/Support: The funding sources had no role in the design and conduct of the study; collection, management, analysis, and interpretation of the data; preparation, review, or approval of the manuscript; and decision to submit the manuscript for publication.

Acknowledgement: The authors are grateful for the contributions of the late J. Randall Curtis, MD, MPH, who was a founding principal investigator of this research program.

References

1. Rijcken E, Zervanou K, Mosteiro P, Scheepers F, Spruit M, Kaymak U. Machine learning vs. rule-based methods for document classification of electronic health records within mental health care—A systematic literature review. *Nat Lang Process J*. 2025;10:100129. doi:10.1016/j.nlp.2025.100129

2. Spasic I, Nenadic G. Clinical Text Data in Machine Learning: Systematic Review. *JMIR Med Inform*. 2020;8(3):e17984. doi:10.2196/17984

3. Wang Z, Chu Z, Doan TV, Ni S, Yang M, Zhang W. History, development, and principles of large language models: an introductory survey. *AI Ethics*. 2025;5:1955-1971. doi:10.1007/s43681-024-00583-7

4. Mitchell M. The Turing Test and our shifting conceptions of intelligence. *Science*. 2024;385(6710):eadq9356. doi:10.1126/science.adq9356

5. Dave Bergmann. What is zero-shot learning? Explainers: Become an expert on emerging and fundamental tech topics. January 24, 2024. Accessed September 10, 2025. https://www.ibm.com/think/topics/zero-shot-learning

6. Bernacki RE, Block SD, American College of Physicians High Value Care Task F. Communication about serious illness care goals: a review and synthesis of best practices. *JAMA Intern Med*. 2014;174(12):1994-2003. doi:10.1001/jamainternmed.2014.5271

7. Curtis JR, Patrick DL, Shannon SE, Treece PD, Engelberg RA, Rubenfeld GD. The family conference as a focus to improve communication about end-of-life care in the intensive care unit: opportunities for improvement. *Crit Care Med*. 2001;29(2 Suppl):N26-33. doi:10.1097/00003246-200102001-00006

8. Seaman JB, Arnold RM, Scheunemann LP, White DB. An Integrated Framework for Effective and Efficient Communication with Families in the Adult Intensive Care Unit. *Ann Am Thorac Soc*. 2017;14(6):1015-1020. doi:10.1513/AnnalsATS.201612-965OI

9. Heyland DK, Barwich D, Pichora D, et al. Failure to engage hospitalized elderly patients and their families in advance care planning. *JAMA Intern Med*. 2013;173(9):778-787. doi:10.1001/jamainternmed.2013.180

10. Shah K, Swinton M, You JJ. Barriers and facilitators for goals of care discussions between residents and hospitalised patients. *Postgrad Med J*. 2017;93(1097):127-132. doi:10.1136/postgradmedj-2016-133951

11. Kruser JM, Benjamin BT, Gordon EJ, et al. Patient and Family Engagement During Treatment Decisions in an ICU: A Discourse Analysis of the Electronic Health Record. *Crit Care Med*. 2019;47(6):784-791. doi:10.1097/CCM.0000000000003711

12. Lee RY, Uyeda AM, Li KS, et al. Performance of Predefined Search Patterns for Identifying Documented Goals-of-Care Discussions in Inpatient Electronic Health Records. *J Palliat Med*. 2025;28(7):952-959. doi:10.1089/jpm.2024.0496

13. Wilson CJ, Newman J, Tapper S, et al. Multiple locations of advance care planning documentation in an electronic health record: are they easy to find? *J Palliat Med*. 2013;16(9):1089-1094. doi:10.1089/jpm.2012.0472

14. Sinuff T, Dodek P, You JJ, et al. Improving End-of-Life Communication and Decision Making: The Development of a Conceptual Framework and Quality Indicators. *J Pain Symptom Manage*. 2015;49(6):1070-1080. doi:10.1016/j.jpainsymman.2014.12.007

15. Chien I, Shi A, Chan A, Lindvall C. Identification of Serious Illness Conversations in Unstructured Clinical Notes Using Deep Neural Networks. In: Fernando Koch, Andrew Koster, Isabelle Bichindaritz, Pau Herrero, et al, eds. *Artificial Intelligence in Health: First International Workshop, AIH 2018, Stockholm, Sweden, July 13-14, 2018, Revised Selected Papers*. Lecture Notes in Artificial Intelligence. Springer Nature Switzerland; 2019:199-212.

16. Chan A, Chien I, Moseley E, et al. Deep learning algorithms to identify documentation of serious illness conversations during intensive care unit admissions. *Palliat Med*. 2019;33(2):187-196. doi:10.1177/0269216318810421

17. Lee RY, Brumback LC, Lober WB, et al. Identifying Goals of Care Conversations in the Electronic Health Record Using Natural Language Processing and Machine Learning. *J Pain Symptom Manage*. 2021;61(1):136-142 e2. doi:10.1016/j.jpainsymman.2020.08.024

18. Agaronnik ND, Davis J, Manz CR, Tulsky JA, Lindvall C. Feasibility Study for Using Large Language Models to Identify Goals-of-Care Documentation at Scale in Patients With Advanced Cancer. *JCO Oncol Pract*. Published online April 10, 2025:OP2400992. doi:10.1200/OP-24-00992

19. Lee RY, Kross EK, Torrence J, et al. Assessment of Natural Language Processing of Electronic Health Records to Measure Goals-of-Care Discussions as a Clinical Trial Outcome. *JAMA Netw Open*. 2023;6(3):e231204. doi:10.1001/jamanetworkopen.2023.1204

20. Lindvall C, Deng CY, Moseley E, et al. Natural Language Processing to Identify Advance Care Planning Documentation in a Multisite Pragmatic Clinical Trial. *J Pain Symptom Manage*. 2022;63(1):e29-e36. doi:10.1016/j.jpainsymman.2021.06.025

21. Curtis JR, Lee RY, Brumback LC, et al. Intervention to Promote Communication About Goals of Care for Hospitalized Patients With Serious Illness: A Randomized Clinical Trial. *JAMA*. 2023;329(23):2028-2037. doi:10.1001/jama.2023.8812

22. Volandes AE, Chang Y, Lakin JR, et al. An Intervention to Increase Advance Care Planning Among Older Adults With Advanced Cancer: A Randomized Clinical Trial. *JAMA Netw Open*. 2025;8(5):e259150. doi:10.1001/jamanetworkopen.2025.9150

23. Casarett D, Lakis K, Ma JE, et al. Goal-Concordant Care: End-of-Life Planning Conversations for All Seriously Ill Patients. *NEJM Catal*. 2022;3(12). doi:10.1056/CAT.22.0271

24. Taylor SP, Kowalkowski MA, Courtright KR, et al. Deficits in Identification of Goals and Goal-Concordant Care After Sepsis Hospitalization. *J Hosp Med*. 2021;16(11):667-670. doi:10.12788/jhm.3714

25. Devlin J, Chang MW, Lee K, Toutanova K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *arXiv*. Preprint posted online May 24, 2019. Accessed August 29, 2024. http://arxiv.org/abs/1810.04805

26. Alsentzer E, Murphy JR, Boag W, et al. Publicly Available Clinical BERT Embeddings. *arXiv*. Preprint posted online June 20, 2019. Accessed August 29, 2024. https://arxiv.org/abs/1904.03323

27. Agaronnik ND, Davis J, Manz CR, Tulsky JA, Lindvall C. Large Language Models to Identify Advance Care Planning in Patients With Advanced Cancer. *J Pain Symptom Manage*. 2025;69(3):243-250.e1. doi:10.1016/j.jpainsymman.2024.11.016

28. Lee RY, Kross EK, Downey L, et al. Efficacy of a Communication-Priming Intervention on Documented Goals-of-Care Discussions in Hospitalized Patients With Serious Illness: A Randomized Clinical Trial. *JAMA Netw Open*. 2022;5(4):e225088. doi:10.1001/jamanetworkopen.2022.5088

29. Curtis JR, Lee RY, Brumback LC, et al. Improving communication about goals of care for hospitalized patients with serious illness: Study protocol for two complementary randomized trials. *Contemp Clin Trials*. 2022;120:106879. doi:10.1016/j.cct.2022.106879

30. Jack Bandy, Nicholas Vincent. Addressing “Documentation Debt” in Machine Learning: A Retrospective Datasheet for BookCorpus. In: J. Vanschoren, S. Yeung, eds. *35th Conference on Neural Information Processing Systems Datasets and Benchmarks Track*. 2021.

31. Johnson AE, Pollard TJ, Shen L, et al. MIMIC-III, a freely accessible critical care database. *Sci Data*. 2016;3:160035. doi:10.1038/sdata.2016.35

32. Alsentzer E. BioClinicalBERT. 2019. https://huggingface.co/emilyalsentzer/Bio_ClinicalBERT

33. Jacob Devlin, Ming-Wei Chang. Open Sourcing BERT: State-of-the-Art Pre-training for Natural Language Processing. 2018. https://research.google/blog/open-sourcing-bert-state-of-the-art-pre-training-for-natural-language-processing/

34. Secunda K, Wirpsa MJ, Neely KJ, et al. Use and Meaning of “Goals of Care” in the Healthcare Literature: a Systematic Review and Qualitative Discourse Analysis. *J Gen Intern Med*. 2019;35(5):1559-1566. doi:10.1007/s11606-019-05446-0

35. Uyeda AM, Curtis JR, Engelberg RA, et al. Mixed-methods evaluation of three natural language processing modeling approaches for measuring documented goals-of-care discussions in the electronic health record. *J Pain Symptom Manage*. 2022;63(6):e713-e723. doi:10.1016/j.jpainsymman.2022.02.006

36. Cook J, Ramadas V. When to consult precision-recall curves. *Stata J*. 2020;20(1):131-148. doi:10.1177/1536867x20909693

37. Edmonds KP, Ajayi TA. Do We Know What We Mean? An Examination of the Use of the Phrase “Goals of Care” in the Literature. *J Palliat Med*. 2019;22(12):1546-1552. doi:10.1089/jpm.2019.0059

38. Myers J, Kim GL, Bytautas J, Webster F. Differing Conceptualizations of the Goals of Care Discussion: A Critical Discourse Analysis. *J Pain Symptom Manage*. 2022;63(4):495-502. doi:10.1016/j.jpainsymman.2021.12.020

39. Bhattarai K, Oh IY, Sierra JM, et al. Leveraging GPT-4 for identifying cancer phenotypes in electronic health records: a performance comparison between GPT-4, GPT-3.5-turbo, Flan-T5, Llama-3-8B, and spaCy’s rule-based and machine learning-based methods. *JAMIA Open*. 2024;7(3):ooae060. doi:10.1093/jamiaopen/ooae060

40. Chiang CC, Luo M, Dumkrieger G, et al. A large language model-based generative natural language processing framework fine-tuned on clinical notes accurately extracts headache frequency from electronic health records. *Headache*. 2024;64(4):400-409. doi:10.1111/head.14702

41. Guevara M, Chen S, Thomas S, et al. Large language models to identify social determinants of health in electronic health records. *NPJ Digit Med*. 2024;7(1):6. doi:10.1038/s41746-023-00970-0

42. Patra BG, Lepow LA, Kasi Reddy Jagadeesh Kumar P, et al. Extracting social support and social isolation information from clinical psychiatry notes: comparing a rule-based natural language processing system and a large language model. *J Am Med Inform Assoc JAMIA*. 2025;32(1):218-226. doi:10.1093/jamia/ocae260

43. Zhu X, Li J, Liu Y, Ma C, Wang W. A Survey on Model Compression for Large Language Models. *Trans Assoc Comput Linguist*. 2024;12:1556-1577. doi:10.1162/tacl_a_00704

44. Wei J, Wang X, Schuurmans D, et al. Chain-of-Thought Prompting Elicits Reasoning in Large Language Models. In: Koyejo S, Mohamed S, Agarwal A, Belgrave D, Cho K, Oh A, eds. *Advances in Neural Information Processing Systems 35: 36th Conference on Neural Information Processing Systems (NeurIPS 2022)*. Vol 35. Curran Associates, Inc.; 2022:24824-24837.

45. Gianfrancesco MA, Tamang S, Yazdany J, Schmajuk G. Potential Biases in Machine Learning Algorithms Using Electronic Health Record Data. *JAMA Intern Med*. 2018;178(11):1544-1547. doi:10.1001/jamainternmed.2018.3763

46. Kamruzzaman M, Kim GL. Prompting Techniques for Reducing Social Bias in LLMs through System 1 and System 2 Cognitive Processes. *arXiv*. Preprint posted online 2024. doi:10.48550/ARXIV.2404.17218

47. Poibeau T, Messiant C. Do we Still Need Gold Standards for Evaluation? In: Calzolari N, Choukri K, Maegaard B, et al., eds. *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC ’08)*. European Language Resources Association (ELRA); 2008. http://www.lrec-conf.org/proceedings/lrec2008/pdf/144_paper.pdf

48. Himmi A, Irurozki E, Noiry N, Clemencon S, Colombo P. Towards More Robust NLP System Evaluation: Handling Missing Scores in Benchmarks. *arXiv*. Preprint posted online 2023. doi:10.48550/ARXIV.2305.10284

49. Hoffman K, Salerno S, Afiaz A, Leek JT, McCormick TH. Do We Really Even Need Data? *arXiv*. Preprint posted online February 2, 2024. Accessed August 29, 2024. http://arxiv.org/abs/2401.08702

Appendix 1. Prompt and prompt development for zero-shot Llama 3.3 language model.

Prompt development

Our goal was to develop a zero-shot LLM prompt to measure an outcome previously annotated by human abstractors in our pilot and test sets (codebook published in Lee RY et al, *JAMA Network Open* 2023;6(3):e231204, Supplement 1, eAppendix 2). Beginning with a simple, expert-composed prompt, the authors met weekly over 9 months to qualitatively review LLM outputs on exemplar pilot-set records. Prompt iterations prompted for both binary adjudications and machine-generated reasoning. Based on qualitative review of false positives and negatives and published approaches to prompt engineering, we proposed and tested prompt revisions designed to improve alignment between the LLM's apparent interpretation of the task and the outcome definition used for annotation. Exemplar records were initially sampled randomly across positive and negative strata; and, in later runs, enriched for edge cases to probe specific topical misunderstandings.

This iterative process produced a definition-informed prompt that defined key concepts and instructed the LLM to identify a composite outcome comprising three constructs: goals-of-care discussions (adapted from Secunda et al, *J Gen Int Med* 2023;35(5):1559-1566), documentation about advance care planning documents, and designation of surrogate decision-makers. The prompt also specifically excluded code-status-only discussions. Once the core definitional content of the prompt was finalized, instructions to generate reasoning were replaced with requests for a binary ("YES/NO") adjudication and an "impact score." Although the LLM was asked to consider the impact score in forming judgements, it was not asked to (and did not) output a numeric value. The final prompt was then quantitatively evaluated on the Trial 1 test set and validated on the Trial 2 held-out test set, as described in this report.

Prompt text

```
Before we begin, let's define a few key concepts:

- GOALS-OF-CARE DISCUSSIONS are defined as discussions between clinicians and
patients/families about the patient's aims of medical care, which are ideally informed
by their underlying values and priorities, established within the current clinical
context, and used to guide decisions about the use of or limitations on life-
sustaining treatments.

- CODE STATUS refers to inpatient orders for a patient that specify what is to be done
in the event of cardiac or respiratory arrest--i.e., whether to perform CPR or not
("DNR"), whether to intubate or not ("DNI"). Code status discussions are one type of
goals-of-care discussion, albeit one of narrow scope.

- "Advance care planning documents" are structured documents such as advance
directives and POLST forms that may provide insight into a patient's goals of care or
treatment preferences for future health states, as well as durable power of attorney
(DPOA) forms. One component of advance care planning, the designation of a "surrogate
decision-maker" or "durable power of attorney", refers to designation by a patient of
a proxy who can make healthcare-related decisions on their behalf if they are
incapacitated.


Our research outcome, COMPOSITE_JS, is defined as ANY of three constructs: (1) goals-
of-care discussions that extend -beyond- decisions relating to code status alone; (2)
incident designation or review of surrogate decision-maker or durable power of
attorney (not just citations of past documentation); and, (3) incident completion or
review of advance care planning documents such as advance directives or POLSTs (not
```

just citation of prior documents). COMPOSITE_JS should -not- include text that only references statements aspiring to future goals-of-care discussions, treatment-specific therapeutic goals, or documentation focused on discharge or disposition planning (e.g., plans for post-discharge placement).

Using the medical record excerpts provided in INPUT, do the excerpts in INPUT reflect documentation of a discussion meeting criteria for COMPOSITE_JS? Consider the evidence and assign an impact score adding up to 10, where higher scores indicate stronger influence on the final decision. If key elements are missing (e.g., no documented goals-of-care discussion, no incident designation/review of a surrogate, or no review/completion of advance care planning documents), assign higher scores to reflect their critical absence.

Please only answer YES or NO.

Use the following example format:

ANSWER: [YES/NO]

INPUT = [*insert candidate electronic health record excerpt*]