npj | digital medicine

Article

Published in partnership with Seoul National University Bundang Hospital

https://doi.org/10.1038/s41746-024-01375-3

Hospitalization prediction from the emergency department using computer vision AI with short patient video clips

Check for updates

Wui Ip[1,14] ✉, Maria Xenochristou[2,14], Elaine Sui[3,14], Elyse Ruan[4], Ryan Ribeira[5], Debadutta Dash[5], Malathi Srinivasan[6], Maja Artandi[6], Jesutofunmi A. Omiye[2,7], Nicholas Scoulios[6], Hayden L. Hofmann[2], Ali Mottaghi[8], Zhenzhen Weng[9], Abhinav Kumar[3], Ananya Ganesh[3], Jason Fries[10], Serena Yeung-Levy[2,3,8,11,12] & Lawrence V. Hofmann[4,13]

In this study, we investigate the performance of computer vision AI algorithms in predicting patient disposition from the emergency department (ED) using short video clips. Clinicians often use "eye-balling" or clinical gestalt to aid in triage, based on brief observations. We hypothesize that AI can similarly use patient appearance for disposition prediction. Data were collected from adult patients at an academic ED, with mobile phone videos capturing patients performing simple tasks. Our AI algorithm, using video alone, showed better performance in predicting hospital admissions (AUROC = 0.693 [95% CI 0.689, 0.696]) compared to models using triage clinical data (AUROC = 0.678 [95% CI 0.668, 0.687]). Combining video and triage data achieved the highest predictive performance (AUROC = 0.714 [95% CI 0.709, 0.719]). This study demonstrates the potential of video AI algorithms to support ED triage and alleviate healthcare capacity strains during periods of high demand.

In recent years, computer vision research has made significant strides in various medical applications[1,2]. Leveraging AI algorithms, medical videos have been utilized for patient monitoring, encompassing clinical mobilization[3], gait analysis[4,5], evaluation of pediatric head injuries resulting from falls[6], assessment of Parkinsonian hand movements and disease severity[7,8], and detection of cognitive impairment[9]. Moreover, medical video technology shows promise in aiding medical staff with diverse tasks, including hand hygiene detection [10] and surgical assessment through video analysis[11,12].

Despite these advancements, one important area that remains largely unexplored is the use of medical video AI for patient triage. During triage and ED encounters, physicians determine patient disposition—such as hospital admission or discharge—based on the presenting problem, vital signs, expected clinical course, patient resources, and their observations of the patient's condition, often referred to as "clinical gestalt"[13–15]. Notably, a study has demonstrated that ED physicians can achieve an 80% accuracy rate in predicting disposition based on a 30-s observation of a patient, complemented by routinely available triage information such as vital signs, mode of arrival (e.g., ambulance), and chief complaints[13].

While physician judgment is valuable, it can be limited in a busy ED where physicians may not be available at the time of triage. Existing systems like the Emergency Severity Index (ESI) offer insights into patient acuity but are not designed to predict patient disposition at triage. An AI-based predictive model could provide a more consistent, objective, cost-efficient, and scalable solution to optimize patient flow and resource allocation in real-time.

We hypothesize that mobile phone video AI algorithms can similarly extract valuable insights from patients' clinical appearances to predict their disposition, effectively mimicking clinical gestalt. Successful implementation of such a triage algorithm could mitigate the challenges of overcrowded emergency rooms, particularly during peak viral seasons and pandemics[16]. Moreover, it could enable healthcare systems to allocate finite resources

[1]Department of Pediatrics, Stanford University School of Medicine, Palo Alto, CA, USA. [2]Department of Biomedical Data Science, Stanford University School of Medicine, Palo Alto, CA, USA. [3]Department of Computer Science, Stanford University, Palo Alto, CA, USA. [4]Digital Health Care Integration, Stanford Health Care, Palo Alto, CA, USA. [5]Department of Emergency Medicine, Stanford University School of Medicine, Palo Alto, CA, USA. [6]Department of Medicine, Stanford University School of Medicine, Palo Alto, CA, USA. [7]Department of Dermatology, Stanford University School of Medicine, Palo Alto, CA, USA. [8]Department of Electrical Engineering, Stanford University, Palo Alto, CA, USA. [9]Institute for Computational & Mathematical Engineering, Stanford University, Palo Alto, CA, USA. [10]Stanford Center for Biomedical Informatics Research, Palo Alto, CA, USA. [11]Clinical Excellence Research Center, Stanford University School of Medicine, Palo Alto, CA, USA. [12]Chan Zuckerberg Biohub—San Francisco, San Francisco, CA, USA. [13]Department of Radiology, Stanford University School of Medicine, Palo Alto, CA, USA. [14]These authors contributed equally: Wui Ip, Maria Xenochristou, Elaine Sui. ✉e-mail: wui@stanford.edu

more efficiently and, in the future, prioritize first available appointments for urgent patient telehealth visits[17]. Managing hospital capacity and staffing effectively is a significant challenge in healthcare systems. Imbalances between bed availability and patient demand, along with staffing requirements, can impact hospital access, wait times, care quality, and overall satisfaction for both patients and staff[18]. When bed supply exceeds demand, it results in unnecessary costs due to underutilized resources. On the other hand, when demand surpasses available beds, it leads to longer wait times—especially for ED patients awaiting admission—lower care quality, higher risks of errors, and decreased satisfaction for both patients and staff, and in extreme cases diversion of patients to other hospitals. Predicting disposition early could help balance these issues, improving patient flow and resource use.

Studies have shown that machine learning models using data collected at triage and from electronic health records (EHR) can effectively predict patient disposition from the ED, including hospitalization and the need for critical care[19–27]. Various algorithms, such as logistic regression, random forests, gradient boosting, and deep neural networks, have been employed, with gradient boosting and deep neural networks often yielding better performance[19,27].

However, many of these models rely on historical patient data from the EHR, such as past ED visits or hospitalizations[20,23,24], or past diagnosis[19–21,23,24,26]. This dependence on historical data poses challenges, particularly for patients new to a health system or those transitioning between systems with incomplete or limited data. Additionally, the use of EHR-based models may limit applicability in developing countries where EHR systems are not yet widely available.

Some existing models also incorporate the ESI, a 5-point triage scoring system commonly used in EDs[28,29], as a predictor[22–24,27]. While ESI scores are useful, they still require manual input from triage staff based on a patient's presenting symptoms, vital signs, and clinical judgment regarding severity and resource needs. This reliance on human input limits the potential of algorithms to ease the triage burden.

The aim of this study is to assess the predictive performance of a multimodal AI algorithm for patient disposition from the ED, using only a short mobile phone video clip capturing the patient's clinical appearance (Fig. 1) and a limited set of clinical data (age, sex, vital signs, pain level, and chief complaint), without relying on historical patient data from the EHR. We compare the performance of our multimodal model (Fig. 2) to a reference model based on logistic regression using ESI triage scores, as well as to ablated versions of our model that utilize only video data or only triage data. This is the first step towards our long-term goal of developing a video-based AI triage tool in pre-hospital settings that could be adopted without the need for EHR integration.

Results

Patient characteristics

We approached a total of 843 adult patients and enrolled 723 of them at the Stanford Health Care ED between August 2021 and September 2022. In the enrolled patients, the median age was 52 years (interquartile range [IQR] 33–76), and 51% were female (Table 1). Nearly all patients (over 95%) received a triage ESI score of 2 or 3, and 40.9% were subsequently admitted to the hospital (inpatient or hospital observation). Pain level data were missing for 97 patients, temperature was missing for two patients, and one patient was missing all vital signs. The most frequently reported chief complaints during triage were abdominal pain (16.5%), followed by chest pain (8.0%), shortness of breath (5.5%), and dizziness (3.9%) (Table 1).

Model performance

The model utilizing video data alone yielded better performance in hospitalization prediction compared to the one utilizing triage clinical data alone, across multiple metrics including AUROC (0.693 vs 0.678), PPV (0.563 vs

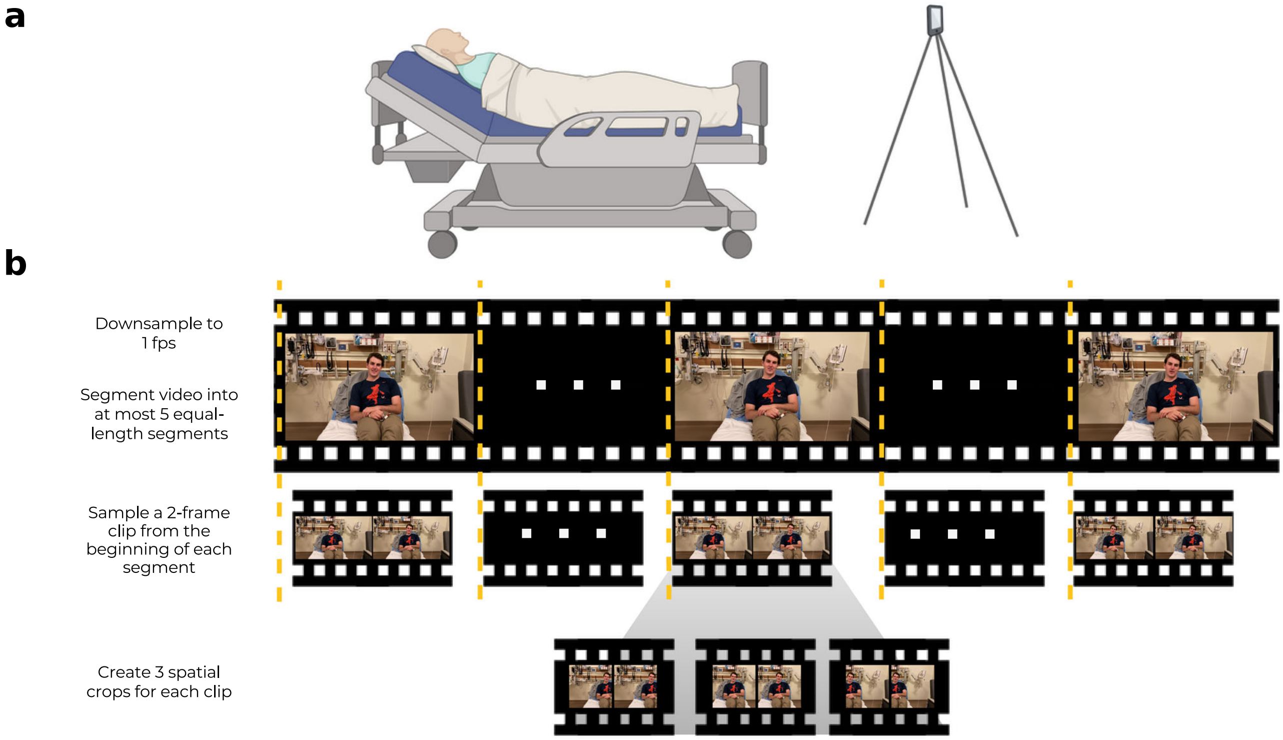
a
b
Downsample to 1 fps
Segment video into at most 5 equal-length segments
Sample a 2-frame clip from the beginning of each segment
Create 3 spatial crops for each clip

Fig. 1 | Patient video recording and processing. **a** Video recording was conducted using secured mobile devices and lasted ~5 min, with the camera at the base of the patient's bed, and the patient as upright as tolerated. Video recording did not interfere with patient care, and was paused if the clinical team needed to interact with the patient. Videos were spot-checked regularly to ensure adherence to study protocol. The figure was created with BioRender.com. **b** The video recording (did not include audio) was then processed via the ImageBind video processing pipeline, which involves uniformly sampling up to five 2-s clips at 1 frame per second and taking three spatial crops (left, middle, right) per clip. Then, these fifteen spatially cropped clips are passed into the vision encoder and the resulting output is the mean of the clips' 1024-dimensional representations. (Subject in the figure is a co-author, not patient).

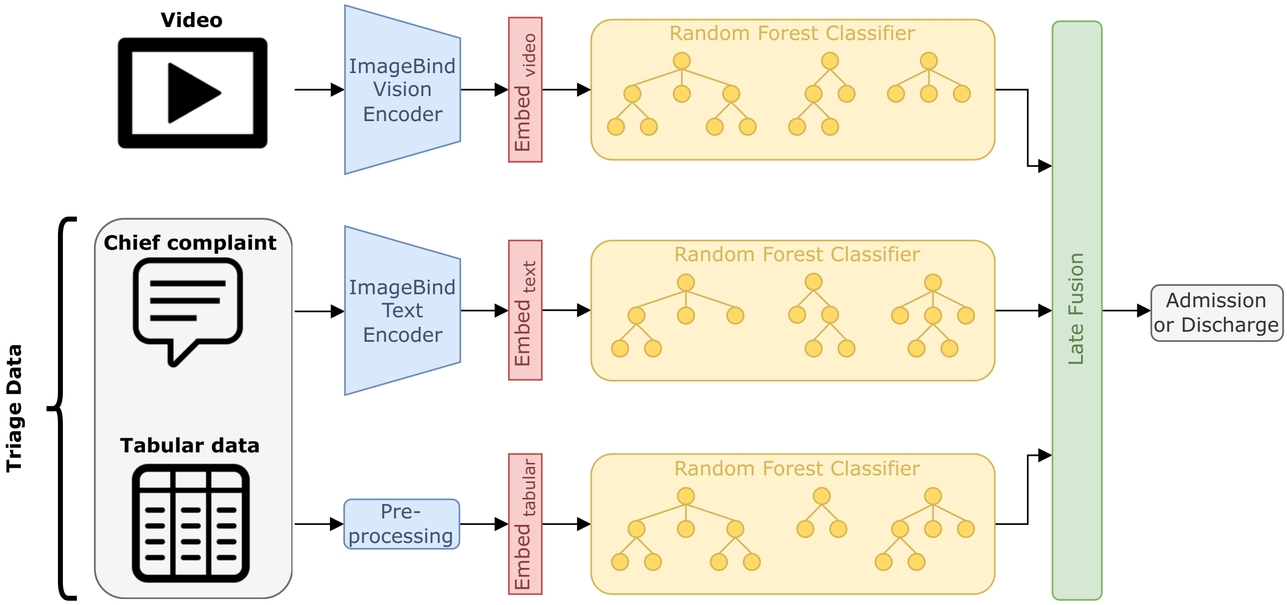

Fig. 2 | Schematic representation of the predictive model. We illustrate our multimodal late fusion method. 1) Collect triage data and a short video of the patient. 2) Pre-process the tabular data and encode the chief complaint and video with ImageBind pre-trained text and vision encoders, to obtain embeddings for each data modality (tabular, text, video). 3) Independently train random forest classifiers for each data modality. 4) Fuse the predictions of each of the trained Random Forests to get the final model prediction for patient disposition.

0.529), and specificity (0.658 vs 0.587). The video-data-only model and the triage-data-only model were comparable for AUPRC (0.608 vs 0.603), NPV (0.721 vs 0.716), while sensitivity (0.632 vs 0.658) is better in the triage-data only model.

Combining both video and triage data resulted in the highest performance, achieving an AUROC of 0.714 (95% CI: 0.709–0.719) and an AUPRC of 0.642 (95% CI: 0.636–0.649). Additional performance details are illustrated in Fig. 3, Fig. 4 and Supplementary Table 1. Notably, all models demonstrated improved performance with the exception of specificity when compared to the reference ESI model.

Furthermore, we conduct a Shapley[30] value analysis (Supplementary Fig. 1) to determine the contribution of adding triage data and video to the overall model performance. These values quantify contribution by taking the average marginal contribution of each of these data types over all possible subsets. We find that video data contributes +0.111 and triage data contributes +0.096 points increase in average model AUROC.

Discussion

Physicians instinctively use visual cues from their interactions with patients to assess the severity of illness. This is commonly referred to as “the eye-ball test” or “the foot of the bed test”, which involves observing a patient from the foot of their hospital bed. Similarly, using visual AI, hospital admissions can be predicted from the video signal alone when a patient is asked to perform simple tasks. Our findings suggest that mobile phone video data contains valuable information for hospitalization prediction, with an AUROC of 0.693. Interestingly, triage data alone, which included an assessment by a medical professional using medical equipment (such as pulse oximetry and sphygmomanometer) has an AUROC of 0.678. When video AI is combined with triage data, the AUROC increases to 0.714. Our results suggest that the short video clips may capture “clinical gestalt” detected and leveraged by the AI algorithm.

Our study demonstrates the potential of utilizing short video clips of patients to predict their ED disposition. It is unexpected to see that video-only model has better performance compared to the triage-date-only model considering the prominence of biometric data, such as vital signs, in clinical risk assessment, as indicated by previous studies[15,19]. One possible explanation is that video data implicitly encodes certain biometric parameters, such as respiratory rate and heart rate[31,32], and also includes markers of patient distress and alterations in breathing patterns and uncomfortable movements.

Other machine learning models have been developed to predict hospitalization from the emergency department[19–27], but none have leveraged patient video data. With the widespread availability of smart mobile devices equipped with cameras, video data may become increasingly accessible to both patients and providers in healthcare settings. Unlike many models that rely on historical EHR data, our multimodal AI algorithm uses only routinely available triage data (age, sex, vital signs, pain level, and chief complaints). As a result, this predictive model does not require integration with existing EHR systems, making it easier to deploy. It could also be useful in countries or regions where EHR systems are not available. Unlike studies such as Raita et al.[19] or Hong et al.[24], we did not include the mode of ED arrival as a covariate, despite it being a known strong predictor of hospital admission[33], because our goal was to develop an algorithm that could facilitate patient disposition decisions before hospital arrival.

This study has several limitations. Our outcome variable was admission decisions made by ED physicians for each patient. ED physicians may vary in their admission decisions based on factors such as patient severity of illness, patient resources (e.g., availability of a caregiver at home, homelessness, or access to transportation), practice style variation, and subjective risk assessment of likely clinical deterioration[34–36]. Such variability inherently sets an upper bound on how well an AI algorithm can perform in predicting hospitalization. Additionally, the videos were sourced from a single academic institution, resulting in a relatively small dataset from an AI training perspective. We anticipate that predictive performance will improve with larger and more diverse datasets from various healthcare settings. Next, the patient population in this study has higher acuity (only 3.6% with ESI 4 or 5), and we limited enrollment to English- or Spanish-speaking patients as well as those able to provide informed consent, which may not generalize to other ED populations. The admission rate of our enrolled cohort was 40.9%, which is slightly higher than the overall admission rate at the Stanford ED of 33.1%. This discrepancy is likely due to some low-acuity patients being treated and discharged quickly for simple issues (e.g., suture removal) before our research assistants could obtain consent. Apart from this, our cohort remains representative of the overall Stanford ED population, with a similar ESI distribution, where the majority of patients were categorized as ESI 2 and 3.

Table 1 | Characteristics of enrolled patients

Variable	Enrolled ED Patients [*n* (%)]	Admission [*n* (%)]
Total	723 (100%)	296 (40.9%)
Sex		
Female	370 (51.2%)	137 (37.0%)
Male	353 (48.8%)	159 (45.0%)
Race and Ethnicity		
Asian	113 (15.6%)	43 (38.1%)
Black or African American	45 (6.2%)	14 (31.1%)
Hispanic or Latino	140 (19.4%)	44 (31.4%)
Native American or Pacific Islander	9 (1.2%)	5 (55.6%)
Non-Hispanic White	362 (50.1%)	169 (46.7%)
Other	54 (7.5%)	21 (38.9%)
Triage ESI score		
1 (highest acuity)	1 (0.14%)	1 (100%)
2	175 (24.2%)	93 (53.1%)
3	521 (72.1%)	198 (38.0%)
4	26 (3.6%)	4 (15.4%)
5 (lowest acuity)	0 (0.0%)	0 (0.0%)
Top chief complaints		
Abdominal Pain	119 (16.5%)	48 (40.3%)
Chest pain	58 (8.0%)	26 (44.8%)
Shortness of Breath	40 (5.5%)	30 (75.0%)
Dizziness	28 (3.9%)	8 (28.6%)
Abnormal Lab	27 (3.7%)	17 (63.0%)
Back pain	21 (2.9%)	5 (23.8%)
Fever	20 (2.8%)	14 (70.0%)
Fall	19 (2.6%)	5 (26.3%)
Flank pain	18 (2.5%)	3 (16.7%)
Leg pain	13 (1.8%)	3 (23.1%)

Furthermore, our algorithm did not incorporate audio or additional EHR data elements, such as past medical history, prior hospitalizations, or other social determinants of health (e.g., health literacy, homelessness), which may further enhance performance[24]. Our experiment represents a scenario where patients are new to the healthcare system, providing a baseline for algorithm performance using only routinely collected triage information (i.e., vital signs, pain level, and chief complaints). Lastly, we did not directly compare physician gestalt with our AI algorithm given the added study design complexity and coordination with more than 90 attending physicians, but it would be a valuable direction for future research.

Since routine recording of patient clinical appearance is not standard practice, this study uncovered several logistical and ethical challenges. Logistically, the research team had to secure approval from the institution's information technology office to ensure all recording devices and the infrastructure for secure video data storage complied with institutional standards. In addition to obtaining IRB approval, we also sought review from the hospital's privacy office to ensure full compliance with institutional policies. Early in the study, we learned the importance of avoiding the unintentional recording of hospital staff in the patient's room to protect their privacy. To address this, our research assistants received extensive training not only on how to operate the recording devices but also on minimizing disruptions to clinical workflows.

Ethically, a key concern is the potential for video recordings to capture sensitive situations, such as patient distress or unexpected clinical events. Obtaining informed consent from patients is critical, as they must understand how their video data will be used and the risks of sharing such sensitive information. Patients are also given the option to request the removal of their data from the study after their ED encounter, recognizing that the stress of the situation may lead to second thoughts. Future studies should address these challenges by establishing clear protocols that prioritize patient privacy and safety, while adhering to institutional and regulatory standards.

Despite these challenges, advances in video AI research offer promising opportunities for future triage workflows. In a future scenario, patients could check in at kiosks equipped with mobile devices, where a brief video recording, combined with vital signs measurements, could aid in triage. Alternatively, patients could remotely provide similar information using their own mobile devices, potentially streamlining the triage process and diverting low-risk patients to telemedicine encounters for final disposition. Increasing appropriate healthcare access is particularly critical during periods of overcrowding, disease outbreaks, and in underserved areas both nationally and internationally. Video AI innovations have the potential to alleviate healthcare capacity strains—delivering the right care, at the right time, and in the right place for patients.

Methods

Data acquisition—video and triage data

We enrolled adult patients (aged 18 years and older), who were English and/or Spanish speaking, undergoing evaluation at the Stanford Health Care ED between August 2021 and September 2022. Stanford Health Care ED is a suburban tertiary academic medical center in Palo Alto, California, with an annual patient volume of over 110,000. The distribution of patients by ESI is as follows: ESI 1—0.7%, ESI 2—27.6%, ESI 3—60.8%, ESI 4—10.2%, and ESI 5—0.8%. The overall admission rate for adult patients is 33.1%. The ED is staffed by 93 attending physicians, 60 resident physicians, 15 advanced practice providers, and 190 registered nurses. Patients were excluded if they presented with psychiatric emergencies, major trauma, altered mental status (e.g., significant delirium), clinical deterioration requiring immediate intervention (such as intubation), or were otherwise unable to provide consent. Due to privacy concerns related to video recording, we excluded patients who were placed in the hallway. However, these cases were uncommon, as our facility, which opened in 2019, has expanded capacity to accommodate a high volume of patients. Prior to participation, all patients provided informed consent for video recording and study participation.

Research assistants (RAs) received 10 h of training, which included observing patient interactions by faculty members to ensure video fidelity and ethical patient engagement. RAs collaborated with ED attending physicians and charge nurses to identify eligible patients during morning, afternoon, and evening shifts on weekdays and weekends throughout the study period. On average, an RA enrolled 2–4 eligible patients per shift.

Video recording was conducted using secured mobile devices (iPhone 12, Apple Inc) and lasted ~5 min. The camera was positioned at the base of the patient's bed, with the patient in an upright position as tolerated (Fig. 1). Video recording did not interfere with patient care and was paused if clinical staff needed to interact with the patient. Videos were regularly spot-checked to ensure adherence to study protocol.

During video recording, patients were verbally instructed to perform seven tasks (Supplementary Table 2) designed to capture their clinical appearance and behaviors. The tasks were designed to be simple for patients to perform and intended to mimic a range of observations typically made during a clinical encounter. These tasks included assessment of general appearance (looking into the camera at the beginning and end), respiration (taking a single deep breath; counting after taking a deep breath), orientation (answering three simple questions: "What is your name?", "Where are you now?", and "What is today's date?"), and standardized movements (performing a modified finger-to-nose test and covering the left eye with the right hand). Although audio was recorded, it was not included in our models.

We extracted 14 clinical data elements from the electronic health record (EHR): age (in years), sex as recorded in the EHR (female or male),

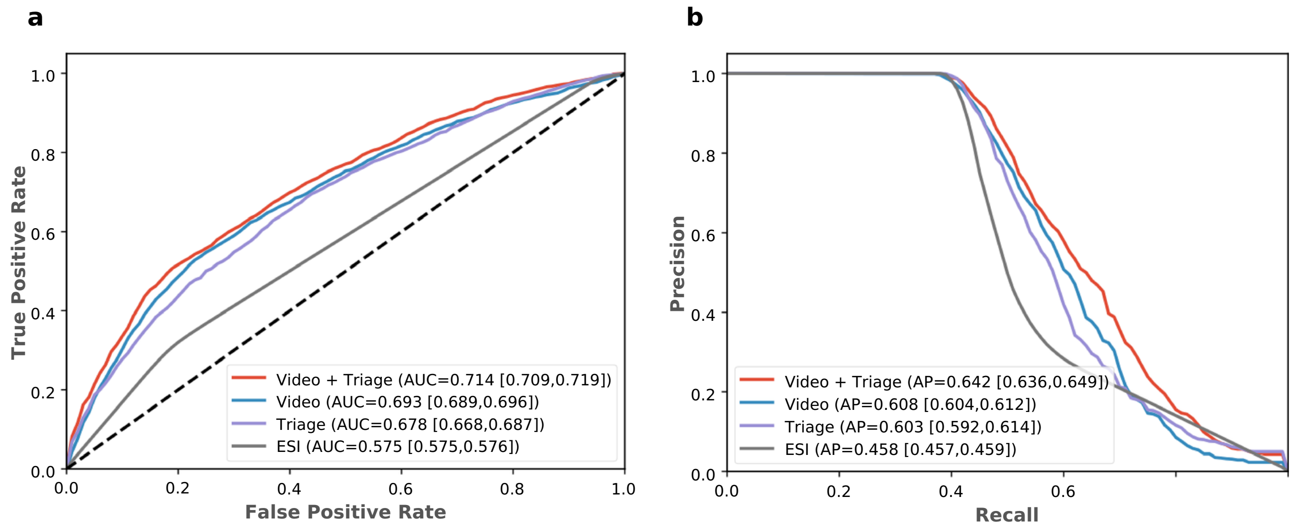

Fig. 3 | Comparison of receiver operating characteristic curve and precision-recall curve. Comparison of **a** receiver operating characteristic curves and **b** precision-recall curves for the baseline ESI model, triage data-only model, video-only model, and late fusion video and triage data model. We observe that each uni-modal model is able to achieve performance over the baseline ($p < 0.01$), with the late fusion multi-modal model achieving higher performance than the uni-modal ones ($p < 0.01$). The square bracket indicates 95% confidence intervals.

self-reported racial and ethnicity group (Asian, Black or African American, Hispanic or Latino, Native American or Pacific Islander, Non-Hispanic White, Other), and primary reasons for the ED visit (free text). Additionally, six triage vital signs were collected: body temperature, heart rate (beats per minute), respiratory rate (respirations per minute), pulse oximetry (peripheral hemoglobin oxygen saturation %), and systolic and diastolic blood pressure (mmHg). The nurse-generated triage score was recorded using the ESI[28,29] where 1 indicates the most urgent and 5 the least urgent. We also collected the initial pain level reported at triage (on a scale of 0 = no pain to 10 = severe pain) and the final disposition (discharge or hospital admission).

The primary reasons for ED visits, known as chief complaints (e.g., chest pain, shortness of breath, dizziness) were entered as free text by nursing staff. The ESI is a commonly used triage tool that reflects the triage nurse's assessment of the severity of the patient's presenting illness and downstream resource utilization[28]. Hospital admission, our dependent variable, included patients admitted to the ED observation unit or an inpatient clinical service (e.g., medicine, cardiology, neurology), as well as those transferred to another hospital

All patient data was securely stored on encrypted devices, following security best practices. The study team completed a data risk assessment and privacy review through Stanford Health Care. The research protocol was approved by the Stanford University Institutional Review Board (IRB protocol #61666).

Data processing—video and triage data

Our tabular data include nine elements (patient's age, sex, triage vital signs including body temperature, heart rate, respiratory rate, oxygen saturation, systolic blood pressure, and triage pain level). We use these nine data elements to create a 9-dimensional vector representation of the data. Missing feature values were imputed using the mean value from the training set, with the exception of pain level, which was imputed with zero. Our assumption is that if the triage nurse did not record pain level, the patient was likely not experiencing pain.

We used frozen pre-trained ImageBind[37] text and vision encoders to encode the patient chief complaint-free text and the patient videos. ImageBind is a method that learns a shared embedding space for six different data modalities (text, image/video, audio, depth, thermal, inertial measurement unit) through contrastive learning[38–41]. These multi-modal contrastive trained models have been shown to perform well on many modality-specific downstream tasks, even out-performing their uni-modal trained counterparts[42].

Specifically, the free text is tokenized into a byte-pair encoding and encoded using a transformer-based architecture to obtain a 1024-dimensional embedding. The ImageBind video processing pipeline involves uniformly sampling up to five 2-s clips at 1 frame per second and taking three spatial crops (left, middle, right) per clip. Then, these fifteen spatially cropped clips are passed into the vision encoder and the resulting output is the mean of the clips' 1024-dimensional representations (Fig. 1).

Modeling

We developed a late fusion model (Fig. 2) to classify whether a patient should be admitted or discharged from the ED. Late fusion is a technique that aggregates the predictions of multiple models to make a final classification. In contrast, early fusion models concatenate embeddings and train a single classifier[43]. We chose late fusion for its flexibility in handling varying availability of input data.

Since the individual classifiers are trained independently, this design allows for flexibility in the training set, as it does not require all data modalities to be present for each patient. At inference time, the model can easily manage missing data modalities by excluding their contribution in the final weighted average. In contrast, an early fusion model would require the input size of the concatenated embeddings to remain consistent during both training and inference. We also found experimentally that late fusion outperformed early fusion (Supplementary Tables 3, 4). In our approach, we trained independent Random Forest[44] classifiers for each data modality: triage tabular data, triage chief complaint text, and patient video. A random forest classifier is a classical machine-learning algorithm that combines the outputs of multiple decision trees. Each tree is built by learning the optimal decision rules to classify the data based on its features. We also experimented with other classifiers such as XGBoost and AdaBoost, but found that Random Forest models outperformed the others (Supplementary Tables 3–5).

For the video data, the model was trained on a single task. At inference time, we combined the predicted probabilities from the video data with triage data using the following weighted average:

$$\hat{y} = \alpha\hat{y}_{video} + (1-\alpha)\hat{y}_{triage} \tag{1}$$

$$\hat{y}_{triage} = 0.5\hat{y}_{tabular} + 0.5\hat{y}_{chief\ complaint} \tag{2}$$

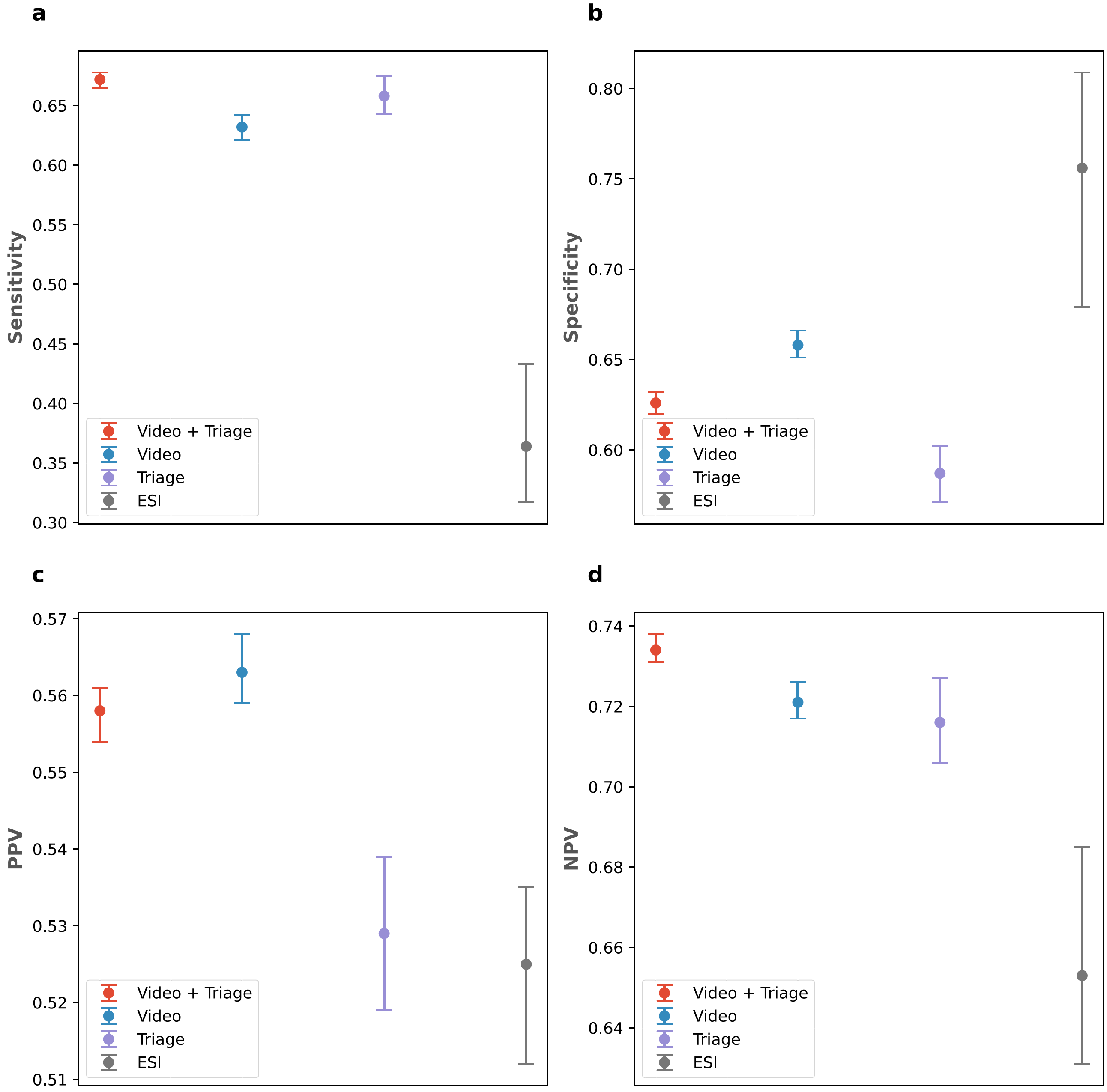

Fig. 4 | Comparison of sensitivity, specificity, PPV, and NPV of different models. Comparison of **a** sensitivity, **b** specificity, **c** PPV, and **d** NPV (with 95% confidence interval) for the baseline ESI model, triage data-only model, video-only model, and late fusion video and triage data model. We observe that the video-based models (video-only, and the late fusion model) have better PPV ($p < 0.01$). The late fusion video and triage data model is able to achieve the best performance in sensitivity ($p < 0.01$ compared to ESI and video-only models; $p < 0.05$ compared to triage data-only model) and NPV ($p < 0.01$). The ESI baseline model has the best specificity ($p < 0.01$ compared to triage data-only and late fusion models; $p < 0.02$ compared to video-only model).

Here in Eqs. (1) and (2), $\hat{y}$ represent the multimodal model's predicted probability of admission, and $\hat{y}_{modality}$ denotes the modality-specific model's prediction, where modality is either video or triage data (tabular or chief complaint free text) The scalar mixing coefficient α lies between 0 and 1.

Upon examining all single-task models, our sensitivity analysis, which compared different video segments where patients performed different tasks (Supplementary Tables 3–5), revealed the Orientation segment to have an AUROC and an AUPRC significantly greater than most other segments. Therefore, we selected the orientation video segment for use in the algorithm.

Experimental details

We compared the predictive performance of our multi-modal model to a reference model that uses logistic regression on the ESI triage severity scores[19], as well as ablated versions of our model using video-data only and triage-data only.

The following metrics were used to assess the performance of the models:

- **Area Under the Receiver Operating Characteristic Curve (AUROC):** This metric quantifies the overall ability of the model to discriminate between positive (i.e., hospitalization) and negative (i.e., discharge) cases across all possible threshold values. A higher AUROC indicates better model performance.
- **Area Under the Precision-Recall Curve (AUPRC):** This metric emphasizes the model's performance on the positive class, particularly useful for imbalanced datasets. A higher AUPRC indicates better precision and recall trade-off.

- **Sensitivity (Recall):** The proportion of actual positives correctly identified by the model (true positive rate). High sensitivity means the model effectively identifies positive cases.
- **Specificity:** The proportion of actual negatives correctly identified by the model (true negative rate). High specificity indicates the model effectively identifies negative cases.
- **Positive Predictive Value (PPV or Precision):** The proportion of positive predictions that are truly positive. High PPV indicates a low false positive rate.
- **Negative Predictive Value (NPV):** The proportion of negative predictions that are truly negative. High NPV indicates a low false negative rate.

For all models, we performed nested stratified threefold cross-validation, repeated over ten random seeds, to conduct our experiments and hyperparameter tuning. We stratified the folds by the patient to prevent data leakage between training and testing. Nested cross-validation involves performing cross-validation on the training set for each outer cross-validation fold and tuning hyperparameters to maximize the average performance on the nested validation sets. This common technique tunes hyperparameters to reduce the bias in performance from tuning hyperparameters on each fold's outer validation set.

For the logistic regression baseline, we selected the inverse of regularization strength from the following values: {0.1, 0.5, 1, 2, 5, 10}. For each Random Forest classifier in our multi-modal model, we selected the number of estimators from {10, 50, 100} and the maximum tree depth from {1, 2, 3, None}. The "None" option allows nodes to be expanded until all leaves are pure or until all leaves contain fewer than two samples. We stratified our dataset by the admission rate, ensuring even distribution between the training and validation sets within each fold.

Given the limited size of our data, we opted for repeated k-fold cross-validation to obtain a more reliable estimate of model performance, choosing a lower value of $k = 3$ for a better estimate of model generalization.

For simplicity, our experiments used a mixing coefficient $\alpha = 0.5$. We report the results of our multi-modal late fusion model with a binary classification threshold of 0.4. This threshold was selected through a grid search of values between 0.1 and 0.9 (at intervals of 0.1), identifying the value that maximized the average Youden's index across all training folds.

Data availability

The video and clinical data used in this study cannot be shared publicly due to the presence of protected health information of patients.

Code availability

The underlying code for this study is not publicly available but may be made available to qualified researchers on reasonable request from the corresponding author.

Received: 24 August 2024; Accepted: 8 December 2024;
Published online: 19 December 2024

References

1. Esteva, A. et al. Deep learning-enabled medical computer vision. *Npj Digit. Med.* **4**, 1–9 (2021).
2. Debnath, B., O'Brien, M., Yamaguchi, M. & Behera, A. A review of computer vision-based approaches for physical rehabilitation and assessment. *Multimed. Syst.* **28**, 209–239 (2022).
3. Yeung, S. et al. A computer vision system for deep learning-based detection of patient mobilization activities in the ICU. *Npj Digit. Med.* **2**, 1–5 (2019).
4. Ino, T. et al. Validity of AI-based gait analysis for simultaneous measurement of bilateral lower limb kinematics using a single video camera. *Sensors* **23**, 9799 (2023).
5. Ramesh, S. H., Lemaire, E. D., Tu, A., Cheung, K. & Baddour, N. Automated implementation of the Edinburgh Visual Gait Score (EVGS) using OpenPose and Handheld smartphone video. *Sensors* **23**, 4839 (2023).
6. Yang, Z., Tsui, B. & Wu, Z. Assessment system for child head injury from falls based on neural network learning. *Sensors* **23**, 7896 (2023).
7. Güney, G. et al. Video-based hand movement analysis of Parkinson patients before and after medication using high-frame-rate videos and mediapipe. *Sensors* **22**, 7992 (2022).
8. Islam, M. S. et al. Using AI to measure Parkinson's disease severity at home. *Npj Digit. Med.* **6**, 1–14 (2023).
9. Chu, C.-S. et al. Automated video analysis of audio-visual approaches to predict and detect mild cognitive impairment and dementia in older adults. *J. Alzheimers Dis.* **92**, 875–886 (2023).
10. Singh, A. et al. Automatic detection of hand hygiene using computer vision technology. *J. Am. Med. Inform. Assoc.* **27**, 1316–1320 (2020).
11. Aklilu, J. G. et al. Artificial Intelligence identifies factors associated with blood loss and surgical experience in cholecystectomy. *NEJM AI* **1**, AIoa2300088 (2024).
12. Goodman, E. D. et al. Analyzing surgical technique in diverse open surgical videos with multitask machine learning. *JAMA Surg*. https://jamanetwork.com/journals/jamasurgery/fullarticle/2812760 (2023).
13. Cabrera, D. et al. Accuracy of 'My Gut Feeling:' comparing system 1 to system 2 decision-making for acuity prediction, disposition and diagnosis in an academic emergency department. *West. J. Emerg. Med.* **16**, 653–657 (2015).
14. Wiswell, J., Tsao, K., Bellolio, M. F., Hess, E. P. & Cabrera, D. Sick' or 'not-sick': accuracy of System 1 diagnostic reasoning for the prediction of disposition and acuity in patients presenting to an academic ED. *Am. J. Emerg. Med.* **31**, 1448–1452 (2013).
15. Barak-Corren, Y. et al. Prediction of patient disposition: comparison of computer and human approaches and a proposed synthesis. *J. Am. Med. Inform. Assoc.* **28**, 1736–1745 (2021).
16. Savioli, G. et al. Emergency department overcrowding: understanding the factors to find. *Corresp. Solut. J. Pers. Med.* **12**, 279 (2022).
17. Kobeissi, M. M. & Ruppert, S. D. Remote patient triage: shifting toward safer telehealth practice. *J. Am. Assoc. Nurse Pract.* **34**, 444–451 (2021).
18. Tello, M. et al. Machine learning based forecast for the prediction of inpatient bed demand. *BMC Med. Inform. Decis. Mak.* **22**, 55 (2022).
19. Raita, Y. et al. Emergency department triage prediction of clinical outcomes using machine learning models. *Crit. Care* **23**, 1–13 (2019).
20. Lee, S.-Y., Chinnam, R. B., Dalkiran, E., Krupp, S. & Nauss, M. Prediction of emergency department patient disposition decision for proactive resource allocation for admission. *Health Care Manag. Sci.* **23**, 339–359 (2020).
21. Goto, T., Camargo, C. A., Faridi, M. K., Yun, B. J. & Hasegawa, K. Machine learning approaches for predicting disposition of asthma and COPD exacerbations in the ED. *Am. J. Emerg. Med.* **36**, 1650–1654 (2018).
22. Graham, B., Bond, R., Quinn, M. & Mulvenna, M. Using data mining to predict hospital admissions from the emergency department. *IEEE Access* **6**, 10458–10469 (2018).
23. Fenn, A. et al. Development and validation of machine learning models to predict admission from emergency department to inpatient and intensive care units. *Ann. Emerg. Med.* **78**, 290–302 (2021).
24. Hong, W. S., Haimovich, A. D. & Taylor, R. A. Predicting hospital admission at emergency department triage using machine learning. *PLOS ONE* **13**, e0201016 (2018).
25. Joseph, J. W. et al. Deep-learning approaches to identify critically Ill patients at emergency department triage using limited information. *J. Am. Coll. Emerg. Physicians Open* **1**, 773–781 (2020).
26. Levin, S. et al. Machine-learning-based electronic triage more accurately differentiates patients with respect to clinical outcomes compared with the emergency severity index. *Ann. Emerg. Med.* **71**, 565–574.e2 (2018).
27. Patel, D. et al. Predicting adult hospital admission from emergency department using machine learning: an inclusie gradient boosting model. *J. Clin. Med.* **11**, 6888 (2022).

28. Völk, S. et al. Patient disposition using the Emergency Severity Index: a retrospective observational study at an interdisciplinary emergency department. *BMJ Open* **12**, e057684 (2022).
29. Sax, D. R. et al. Evaluation of Version 4 of the emergency severity index in us emergency departments for the rate of mistriage. *JAMA Netw. Open* **6**, e233404 (2023).
30. Štrumbelj, E. & Kononenko, I. Explaining prediction models and individual predictions with feature contributions. *Knowl. Inf. Syst.* **41**, 647–665 (2014).
31. Alnaggar, M., Siam, A. I., Handosa, M., Medhat, T. & Rashad, M. Z. Video-based real-time monitoring for heart rate and respiration rate. *Expert Syst. Appl.* **225**, 120135 (2023).
32. Bae, S. et al. Prospective validation of smartphone-based heart rate and respiratory rate measurement algorithms. *Commun. Med.* **2**, 1–10 (2022).
33. Harjola, P. et al. The emergency department arrival mode and its relations to ED management and 30-day mortality in acute heart failure: an ancillary analysis from the EURODEM study. *BMC Emerg. Med.* **22**, 27 (2022).
34. Smulowitz, P. B., O'Malley, A. J., Zaborski, L., McWilliams, J. M. & Landon, B. E. Variation in emergency department admission rates among medicare patients: does the physician matter? *Health Aff.* **40**, 251–257 (2021).
35. Dean, N. C. et al. Hospital admission decision for patients with community-acquired pneumonia: variability among physicians in an emergency department. *Ann. Emerg. Med*. **59**, https://doi.org/10.1016/j.annemergmed.2011.07.032 (2012).
36. Smulowitz, P. B. et al. Physician variability in management of emergency department patients with chest pain. *West. J. Emerg. Med.* **18**, 592–600 (2017).
37. Girdhar, R. et al. ImageBind one embedding space to bind them all. In *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* 15180–15190 (IEEE, Vancouver, BC, Canada). https://ieeexplore.ieee.org/document/10203733 (IEEE, 2023).
38. Chen, T., Kornblith, S., Norouzi, M. & Hinton, G. A Simple framework for contrastive learning of visual representations. In *Proc. 37th International Conference on Machine Learning* 1597–1607 (PMLR, 2020).
39. Oord, A. van den, Li, Y. & Vinyals, O. Representation learning with contrastive predictive coding. Preprint at https://doi.org/10.48550/arXiv.1807.03748 (2019).
40. Zhang, Y., Jiang, H., Miura, Y., Manning, C. D. & Langlotz, C. P. Contrastive learning of medical visual representations from paired images and text. In *Proc. 7th Machine Learning for Healthcare Conference* 2–25 (PMLR, 2022).
41. Chen, T., Kornblith, S., Swersky, K., Norouzi, M. & Hinton, G. E. Big Self-Supervised Models are Strong Semi-Supervised Learners. in *Adv. Neural Inf. Process. Syst.* Vol. **33**, 22243–22255 (Curran Associates, Inc., 2020).
42. Radford, A. et al. Learning transferable visual models from natural language supervision. In *Proc. 38th International Conference on Machine Learning* 8748–8763 (PMLR, 2021).
43. Soenksen, L. R. et al. Integrated multimodal artificial intelligence framework for healthcare applications. *Npj Digit. Med.* **5**, 1–10 (2022).
44. Breiman, L. Random forests. *Mach. Learn.* **45**, 5–32 (2001).

Acknowledgements

This study was funded by Stanford Medicine Catalyst, Stanford University Human-Centered Artificial Intelligence, and Stanford University Sarafan ChEM-H. The funders played no role in study design, data collection, and interpretation of data, or the writing of this manuscript. We would like to thank Dr. Nigam Shah, Chief Data Scientist at Stanford Health Care, for his advice and guidance in this work. Additionally, we would like to thank Dr. Anna Romanowska-Pawliczek for her assistance in preparing Fig. 1, as well as Ms. Daniela Sarellano and Ms. Brittney Williams for their support in project administration. Finally, we would like to express our gratitude to the Stanford University Department of Emergency Medicine for their support of this research project.

Author contributions

W.I., S.Y., and L.H. conceptualized and designed the study, and supervised data analysis. M.X. and E.S. wrote code for model development and conducted data analysis. W.I., M.X., and E.S. wrote the first draft of the manuscript. W.I. and E.R. coordinated the study operations. R.R., D.D., M.S., and M.A. supervised data acquisition and analysis. J.O., N.S., H.H., A.M., Z.W., A.K., A.G., and J.F. assisted in data acquisition and analysis. All authors provided critical revisions, and contributed to the study design, and data interpretation. All authors have read and approved the manuscript.

Competing interests

W.I. is an employee of *nference* and holds stock options in the company. Otherwise the authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41746-024-01375-3.

Correspondence and requests for materials should be addressed to Wui Ip.

Reprints and permissions information is available at http://www.nature.com/reprints

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.