ED-Copilot: Reduce Emergency Department Wait Time with Language Model Diagnostic Assistance

Liwen Sun [1] **Abhineet Agarwal** [2] **Aaron Kornblith** [3] **Bin Yu** [2 4] **Chenyan Xiong** [1]

Abstract

In the emergency department (ED), patients undergo triage and multiple laboratory tests before diagnosis. This time-consuming process causes ED crowding which impacts patient mortality, medical errors, staff burnout, etc. This work proposes (time) *cost-effective diagnostic assistance* that leverages artificial intelligence systems to help ED clinicians make efficient and accurate diagnoses. In collaboration with ED clinicians, we use public patient data to curate MIMIC-ED-Assist, a benchmark for AI systems to suggest laboratory tests that minimize wait time while accurately predicting critical outcomes such as death. With MIMIC-ED-Assist, we develop ED-Copilot which sequentially suggests patient-specific laboratory tests and makes diagnostic predictions. ED-Copilot employs a pre-trained bio-medical language model to encode patient information and uses reinforcement learning to minimize ED wait time and maximize prediction accuracy. On MIMIC-ED-Assist, ED-Copilot improves prediction accuracy over baselines while halving average wait time from four hours to two hours. ED-Copilot can also effectively personalize treatment recommendations based on patient severity, further highlighting its potential as a diagnostic assistant. Since MIMIC-ED-Assist is a retrospective benchmark, ED-Copilot is restricted to recommend only observed tests. We show ED-Copilot achieves competitive performance without this restriction as the maximum allowed time increases. Our code is available at https://github.com/cxcscmu/ED-Copilot.

[1]Language Technologies Institute, Carnegie Mellon University [2]Department of Statistics, University of California, Berkeley [3]Department of Emergency Medicine & Pediatrics, University of California, San Francisco [4]Department of EECS, University of California, Berkeley. Correspondence to: Chenyan Xiong <cx@cs.cmu.edu>.

Proceedings of the 41st International Conference on Machine Learning, Vienna, Austria. PMLR 235, 2024. Copyright 2024 by the author(s).

1. Introduction

Emergency Department (ED) crowding represents a critical challenge in healthcare, significantly impacting morbidity, mortality, medical error, staff burnout, and incurring excessive costs (Sartini et al., 2022). Despite the documented effects of ED crowding, this issue remains inadequately addressed in healthcare systems. An efficient and effective ED is vital for providing timely care to severely ill or injured patients (Savioli et al., 2022).

One key area to address ED crowding, as identified by the American College of Emergency Physicians, is to enhance *throughput*—the efficacy and efficiency of care delivery in the ED (Jarvis, 2016; DeAnda, 2018). A crucial factor affecting throughput is the laboratory testing process, where patients often face lengthy waits for tests to be ordered and completed, delaying diagnoses and treatment decisions (Li et al., 2015). Studies also show that 40 to 60% of ED laboratory tests are unnecessary (Miyakis et al., 2006), further exacerbating wait times.

This paper proposes an artificial intelligence "Co-Pilot" system intended to offer (time) cost-effective diagnostic assistance in the ED. This system should aid diagnosis and minimize ED length of stay (LOS), i.e., wait times, by suggesting laboratory tests after patient triage. Further, it should help with resource management and planning by identifying severely ill patients who require rapid intervention. That is, by selecting informative tests, the system streamlines the diagnostic process, reducing LOS while improving outcomes, particularly for high-risk patients.

To support the machine learning (ML) community in developing a time-cost-effective diagnostic assistant, we collaborate with ED clinicians to curate a benchmark, called MIMIC-ED-Assist, that is derived from MIMIC-IV (Johnson et al., 2023b) and related datasets (Xie et al., 2022). MIMIC-ED-Assist is designed to test the ability of AI systems to provide both accurate and time-cost saving laboratory recommendations. Our benchmark consists of two prediction targets identified by our clinical collaborators to reflect patient risk: critical outcomes which includes patient death and ICU transfer (Levin et al., 2018), and lengthened ED stay, defined as ED LOS exceeding 24 hours. Accu-

rately identifying patients at high risks of these outcomes reduces time-cost by allowing clinicians to perform timely interventions and efficiently allocate resources. MIMIC-ED-Assist mirrors real-world ED practices by grouping individual laboratory tests into commonly performed groups, e.g., complete blood count (CBC). MIMIC-ED-Assist then tests AI systems on their ability to recommend the most informative groups to make accurate diagnostic suggestions while minimizing the total time required to perform these tests, thereby reducing LOS.

With MIMIC-ED-Assist, we propose ED-Copilot which suggests a series of laboratory groups to flag patients at high risks on our prediction targets while minimizing total time-cost. ED-Copilot first linearizes (i.e., converts to text) patient information, including demographic, triage, and laboratory test results into a text sequence. It then fine-tunes a bio-medical pre-trained language model BioGPT (Luo et al., 2022) to suggest future groups and predict our two defined targets. Next, we use a reinforcement learning (RL) framework (Yu et al., 2023) to teach BioGPT to dynamically recommend the subsequent, most informative laboratory group based on prior laboratory and triage information. Unlike baselines, ED-Copilot serves as a personalized diagnostic assistant since it uses past patient information to recommend future medically relevant laboratory groups.

Experiments on MIMIC-ED-Assist show that ED-Copilot outperforms state-of-the-art tree models while halving time-costs of laboratory testing from four hours to two hours. Reducing the number of laboratory tests also has the benefit of reducing financial cost. We also perform ablation studies to confirm the benefits of our feature linearization technique and the bio-medical pre-trained language model backbone. Our ablation studies also investigate the effect of size of the language model backbone on prediction accuracy which indicates larger models can lead to further gain in performance.

Our analyses also confirm the benefit of ED-Copilot's personalized modeling approach. We show ED-Copilot can adapt its recommendations based on patient severity, thereby providing more accurate diagnostic suggestions for severely ill patients as compared to non-personalized baselines. Further, ED-Copilot achieves consistent performance across various subgroups such as age and sex. Lastly, since MIMIC-ED-Assist is a retrospective offline benchmark, we restrict ED-Copilot to only select laboratory tests a patient actually receives. We perform simulations without this restriction to approximate online performance, and show ED-Copilot is still able to make medically appropriate recommendations.

The rest of this paper is organized as follows. In Section 2, we review related work. We discuss MIMIC-ED-Assist and ED-Copilot in Sections 3 and 4 respectively. Sections 5 and 6 discuss our experimental set-up and results.

2. Related Work

Healthcare Benchmarks. Researchers have spent considerable effort in converting raw electronic health records (EHRs) into large-scale open-source datasets to ensure easy access to high-quality medical data. A notable example is the Medical Information Mart for Intensive Care (MIMIC) database (Johnson et al., 2023b) which provides patient information such as measurements, laboratory orders, and treatments, ranging from the ED to inpatient care, including the intensive care unit (ICU). MIMIC has led to the development of a range of related prediction benchmarks and models (Purushotham et al., 2018; Harutyunyan et al., 2019; Wang et al., 2020) focused on the ICU. Xie et al. (2022) took a step towards filling this gap by using the MIMIC-IV-ED (Johnson et al., 2023a) database to build a ED-focused benchmark. Their dataset includes ED triage information, and various clinical outcomes such as hospitalization that interest clinicians, and impact ED LOS.

AI Models for Healthcare. There has been significant effort to apply ML to accurately predict clinical outcomes. Traditional methods (e.g., random forests, gradient boosting, and their variants (Breiman, 2001; Chen & Guestrin, 2016; Agarwal et al., 2022; 2023)), along with deep learning (DL) have been used to predict pneumonia (Kang et al., 2020), and septic shock in the ICU (Wardi et al., 2021). Other works use interpretable models to provide diagnostic assistance in the ED such as identifying traumatic brain injury (Kornblith et al., 2022; Tan et al., 2022). Another closely related line of work is AI for cost-effective medicine. For instance, Bejnordi et al. (2017) showed DL led to faster analysis of pathology laboratory results; Komorowski et al. (2018) proposed an "AI clinician" to learn optimal dosing strategies for treating sepsis; Yu et al. (2023) focused on minimizing financial costs associated with laboratory testing while maximizing prediction accuracy. Specifically, Yu et al. (2023) used RL to sequentially select laboratory groups based on a patient's observed test results to optimize this (financial) cost-accuracy trade-off. They validated the accuracy and cost-effectiveness of their approach on multiple clinical tasks such as diagnosing kidney injury. Researchers have also begun to explore the use of large language models (LLMs) in medical applications. LLMs have been used to extract clinical concepts (Yang et al., 2021; Luo et al., 2022; Yang et al., 2022), and facilitate medical question answering (Singhal et al., 2023; Yagnik et al., 2024).

3. MIMIC-ED-Assist Benchmark

In this section, we discuss the curation of MIMIC-ED-Assist in collaboration with ED clinicians.

Task Description. We consider the following two tasks relevant to reducing ED LOS. 1) Flagging patients at high risks

of critical outcome (i.e., death and ICU transfer), and ED LOS exceeding 24 hours. 2) Providing time-cost-effective diagnostic assistance by recommending the next most medically informative laboratory group while simultaneously minimizing the time-cost of these groups.

Data Pre-processing. Since laboratory results are only available for admitted patients, we filter out non-hospitalized ED patients from MIMIC-IV-ED. We only focus on adults and remove patients younger than 18 years old. We also exclude patients that miss triage information. This step is necessary since clinicians order laboratory groups depending on triage information. To simplify the task, we remove patients who receive the same test multiple times, approximately 1.5% of all patients.

Clinically Relevant Outcomes. In collaboration with ED clinicians, we chose the two following prediction targets. (1) Critical outcome, which refers to death during hospitalization or transfer to an ICU within 12 hours. Identifying patients at high risks of critical outcome allows clinicians to prioritize treatment and resources for them. (2) Lengthened ED stay, indicating if ED LOS exceeding 24 hours. Lengthened ED stay is typically correlated with the complexity of a patient's case. Flagging patients at high risks of lengthened ED stay can enable timely intervention, and reduce ED LOS. The proportion of patients with these outcomes is described in Table 1. While these two outcomes are correlated, they also cover different aspects of patient care. For example, patients at high risks of critical outcomes often should be hospitalized quickly, while patients with complications do not necessarily have severe cases. As such, healthcare providers often require different diagnostics and resource managements for these two tasks.

Triage Feature Selection. Triage features are measurements that are available for every patient before laboratory tests are ordered. We select a number of triage features in collaboration with ED clinicians. Specifically, we chose 9 triage variables available at the beginning of patient encounters, which include patient demographics, medical history, vital signs, and chief complaints (i.e., natural language description of symptoms).

Laboratory Test Selection. We only include laboratory tests performed in the ED. For simplicity, we exclude tests received by less than 5% of patients and leave examination of rare tests to future research. This process results in a total of 68 available laboratory tests. While there are 68 tests, ED clinicians rarely order individual tests for a patient. Typically, they order groups of tests (e.g., complete blood count) based on a patient's signs, symptoms, and risk factors. To reflect this clinical practice, our clinical collaborators categorized these 68 tests into 12 distinct groups. See Appendix A for all 68 tests, and their assigned groupings. On average each patient receives 4.7 laboratory groups. Consequently, MIMIC-ED-Assist contains numerous missing values, each representing a laboratory test not administered to a specific patient.

Table 1. Statistics of MIMIC-ED-Assist. It includes information from patient triage and laboratory tests. Laboratory tests are grouped into 12 groups by ED clinicians based on how they are commonly ordered. An ED visit has a critical outcome if the patient is transferred to ICU or there is an inpatient mortality. ED stay is lengthened if the length of stay (LOS) exceeds 24 hours. The positive rate is shown in parentheses.

Variable/Label	Count
# of ED visits	32356
# of patients	25714
# of triage variables	9
# of laboratory variables	67
# of laboratory groups	12
Avg. # of laboratory groups per patient	4.7
# of Inpatient mortality	467 (1.44%)
# of ICU transfer in 12h	2894 (8.94%)
# of Critical outcome	3129 (9.67%)
# of ED LOS $>$ 24h	2232 (6.90%)

ED LOS. Our benchmark records when each group is ordered, and assigns its average completion time as its 'time-cost'. ED LOS depends on the time-costs of the administered groups and can be modeled in different ways. For example, sequential tests result in ED LOS being the sum of time-costs, whereas parallel tests imply ED LOS is equal to the group with the largest time-cost. MIMIC-ED-Assist does not specify how to approximate ED LOS from these time-costs, and instead provides this flexibility to researchers and practitioners.

Data Availability. Our pipeline to create MIMIC-ED-Assist from the MIMIC-IV dataset can be found at `https://github.com/cxcscmu/ED-Copilot`. After completing a training course and signing a data use agreement regarding patient information privacy, individuals will gain access to MIMIC-IV and can utilize our pipeline to create MIMIC-ED-Assist.

Limitations. Since MIMIC-ED-Assist is derived from public patient data, it suffers from some downsides due to a lack of data availability. (1) MIMIC-ED-Assist is derived from MIMIC-IV which only has laboratory results for hospitalized ED patients. Thus, our dataset is not reflective of the entire population that visits the ED, but instead is biased towards those with more severe issues (hospitalized). (2) As an offline benchmark derived from past patient records, all data is retrospective. As a result when developing diagnostic assistant, one can only use laboratory tests patients actually received, otherwise no testing results are available. This leads to common challenges of offline benchmarks and may

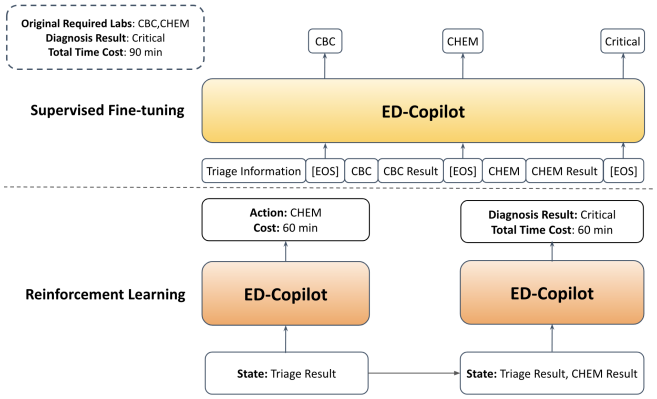


Figure 1. Overview of ED-Copilot procedure on one ED visit.

result in models learning sub-optimal recommendations, and unobserved confounding. We note that measuring 'online' performance of ED-Copilot and AI systems in general will require clinical trials which is out of the scope of this paper. (3) MIMIC-IV is collected at the Beth Israel Deaconess Medical Center, and may not reflect the distribution of other healthcare systems. Caution is needed when generalizing models and insights derived from MIMIC-ED-Assist to new healthcare systems.

4. ED-Copilot for Diagnostic Assistance

We use MIMIC-ED-Assist to develop ED-Copilot, which offers cost-effective diagnostic assistance by minimizing the number of laboratory groups required to identify high-risk patients. This section is organized as follows. First, we provide a high-level overview of ED-Copilot, see Figure 1 for a visualization. Then, we detail its training process which consists of two stages: supervised fine-tuning to adapt the pre-trained language model (PLM) (e.g., BioGPT) to our prediction task, followed by RL to select laboratory groups that reduce time-costs. Finally, we discuss how ED-Copilot conducts inference.

4.1. Problem Formulation

Let a patient has triage features denoted by x_0 and n laboratory groups $[x_i]_{i=1}^n$with their associated results $[r_i]_{i=0}^n$ in order $[x_0, r_0, x_1, r_1, x_2, r_2, ..., x_n, r_n]$. Additionally, let y denotes our prediction targets defined in Section 3. ED-Copilot first linearizes this patient's triage and laboratory results into a text sequence, which is then inputted into the PLM backbone to suggest the next group of tests or predict the outcome. We describe these steps in detail as follows.

Linearization. As laboratory results and triage information are stored in a tabular format, we first convert this information to text for a PLM to use. Specifically, we linearize laboratory group results via the following text template (Hegselmann et al., 2023) r = *test name* : *test value*|*test name* : *test value*... *Test name/value* refers to the name and recorded result of the ordered laboratory test respectively. Each *test name/value* pair is separated by a pipe (|).

Language Model Backbone. We apply PLM G_θ to the text sequence constructed above to obtain hidden representations $\boldsymbol{H}$ for each [EOS] token. We then use a MLP $p_\phi(x_i|\boldsymbol{h}_{<i})$ to predict the next laboratory group using hidden representations $\boldsymbol{h}_{<i}$ for tokens $\texttt{[EOS]}_{<i}$. Lastly, we predict outcomes y using $p_\psi(y|\boldsymbol{h}_{\leq n})$ where $\boldsymbol{h}_n$ is the hidden state for token $\texttt{[EOS]}_n$. This process is reflected below.

$$[x_0, r_0, \texttt{[EOS]}_0, ..., x_n, r_n, \texttt{[EOS]}_n, y] \xrightarrow{G_\theta} \boldsymbol{H}, \quad (1)$$

$$p_\phi(x_i|\boldsymbol{h}_{<i}) = \text{MLP}_\phi(\boldsymbol{h}_{i-1}), \quad (2)$$

$$p_\psi(y|\boldsymbol{h}_{\leq n}) = \text{MLP}_\psi(\boldsymbol{h}_n). \quad (3)$$

Note that we only predict laboratory groups, not their associated laboratory tests' results, which should be determined by conducting the actual laboratory tests. Next, we describe our training procedure.

4.2. Supervised Fine-tuning

In this section, we perform supervised fine-tuning of our PLM to suggest the next laboratory group and predict outcomes. To predict the next laboratory group, we use a standard auto-regressive loss function,

$$\mathcal{L}_{\text{lab}} = -\frac{1}{n}\sum_{i=1}^{n} \log p_\phi(x_i|\boldsymbol{h}_{<i}). \quad (4)$$

For outcome prediction, we use the following loss,

$$\mathcal{L}_y = -\log p_\psi(y|\boldsymbol{h}_{\leq n}). \quad (5)$$

To fine-tune, we minimize the loss,

$$\theta^*, \psi^*, \phi^* = \min_{\theta,\psi,\phi}(\mathcal{L}_{\text{lab}} + \mathcal{L}_y), \quad (6)$$

where θ, ψ, ϕ are parameters in the PLM G_θ and two MLPs p_ψ, p_ϕ. ED clinicians can use the fine-tuned PLM to suggest a sequence of laboratory groups and predict outcomes.

4.3. Reinforcement Learning

Choosing informative laboratory groups is a key factor in reducing laboratory testing, and thereby reducing ED LOS. In this section, we employ RL to introduce the notion of time-cost effectiveness to the fine-tuned PLM to select laboratory groups that maximize predictive accuracy while minimizing time-cost.

Markov Decision Process. The RL process can be viewed as a Markov Decision Process (MDP), represented by $\{\mathcal{S}, \mathcal{A}, P, R\}$. Let $\mathcal{S}$ denote the state space, and state $s \in \mathcal{S}$

corresponds to a patient's observed triage information, and laboratory group results. Specifically, for a patient with observed information $[x_0, r_0, x_1, r_1, x_2, r_2, ..., x_n, r_n]$, let $s_{\leq i} = [x_0, r_0, x_1, r_1, x_2, r_2, ..., x_i, r_i]$ for $0 \leq i \leq n$.

Action Space. Denote the action space as $\mathcal{A} = \{x_1, x_2, ..., x_K\} \cup \{y^+, y^-\}$. Action $a \in \{x_1, x_2, ..., x_K\}$ corresponding to ordering a group of laboratory tests with associated time-cost $c(a)$. Action $a \in \{y^+, y^-\}$ refers to predicting an outcome and terminating the MDP.

Policy. The policy $\pi_\eta : \mathcal{S} \to \mathcal{A}$ maps from states to actions, and is parameterized by η. Specifically, $\pi_\eta(a|\boldsymbol{h}_{<i})$ outputs probability of an action a using the hidden representation $\boldsymbol{h}_{<i}$ corresponding to `[EOS]` token for state s_i.

RL Training. We train the policy π_η to follow two objectives: maximize F1-score and minimize time-cost. We measure the time-cost of policy π_η, as follows:

$$\text{Cost}(\pi_\eta) = \mathbb{E}_{\pi_\eta}[\sum_{t \geq 0} \sum_{j \in [K]} c(j) \cdot \mathbf{1}\{a_t = x_j\}], \quad (7)$$

where $\mathbb{E}_{\pi_\eta}$ is the expectation under policy π_η. Using (7), let $\pi^*_\eta(\alpha, \beta)$ represent the policy that maximizes F1-score while minimizing time-cost for hyper-parameters α, β to be defined. Then, $\pi^*_{\alpha,\beta}$ is equivalent to the solution of the following program:

$$\pi^*_\eta(\alpha, \beta) = \text{argmax}_{\pi_\eta}\{\text{TN}(\pi_\eta) + \alpha\text{TP}(\pi_\eta) + \beta\text{Cost}(\pi_\eta)\}. \quad (8)$$

$\text{TN}(\pi_\eta)$ and $\text{TP}(\pi_\eta)$ are the numbers of true negatives and true positives under policy π_η. Hyper-parameters α, β control the trade-off between F1-score and time-cost. We train the MLP via proximal policy optimization (Schulman et al., 2017).

Measuring Time-cost. We measure time-cost via the total time taken to run all laboratory groups. In the ED, laboratory groups are often ordered both in parallel, and sequentially since the decision on new tests to order depends on previous tests. Additionally, the number of tests an ED clinician orders depends on other factors outside of the patient's health record, such as insurance policy. As the first step towards reducing ED LOS with AI, we use the sum of laboratory group time-costs as the total cost, which serves as an approximation to the effect of testing on ED LOS. Better ways to model ED LOS are future research for both the AI and healthcare communities.

4.4. Inference

During inference, ED-Copilot assists clinicians to optimize their workflow by suggesting the next most informative laboratory group, and also by flagging patients at high risks of critical outcome and lengthened ED stay. Specifically, given triage information and previous test results, ED-Copilot recommends additional tests. The results of these tests are fed back to ED-Copilot to suggest additional laboratory groups or to flag one of two possible outcomes. Further, ED-Copilot can be used in more others ways depending on clinical needs. For example, multiple tests can be ordered using multiple suggestions from ED-Copilot.

5. Experimental Set-up

Dataset Split. We randomly split the dataset using 80% for training, 10% for validation, and 10% for testing, while ensuring each split has the same class distribution. The validation set is used to tune hyper-parameters. During inference, the initial state of a patient is set to their triage information.

Evaluation Metrics. We use four evaluation metrics: F1-score, area under the receiver operating characteristic (AUC), sensitivity (true positive rate), and specificity (true negative rate) which are all standard in healthcare tasks prediction tasks (Harutyunyan et al., 2019; Xie et al., 2022).

Baselines. For prediction tasks only, ED-Copilot is compared to three tree-based methods: random forests (Breiman, 2001), XGBoost (Chen & Guestrin, 2016) and LightGBM (Ke et al., 2017), all of which are known to achieve strong performance on tabular data. We also compare ED-Copilot to a 3-layer dense deep neural network (DNN).

Since each patient is not observed under all laboratory groups, there are many missing values in MIMIC-ED-Assist. Since baselines cannot naturally handle missing values, we consider a variety of imputation methods: (1) mean imputation, (2) median imputation, and (3) zero-imputation (i.e., replacing missing values by 0), (4) using a dummy indicator to encode missing values. Results for the best imputation method are presented in the main manuscript. The rest are in Appendix C.

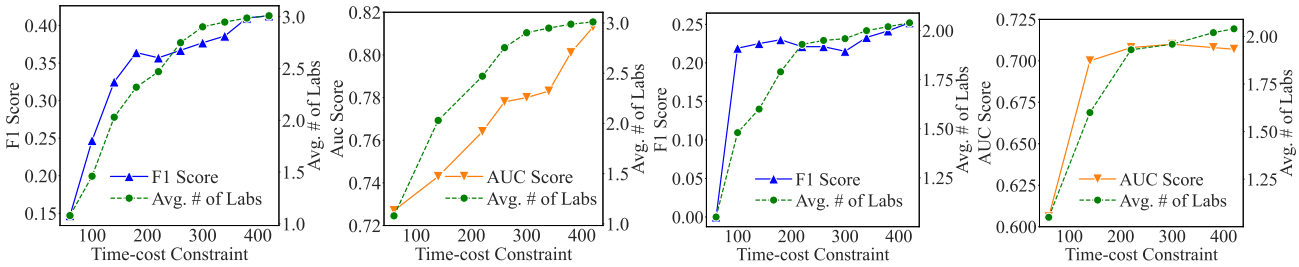
We compare ED-Copilot to another cost-effective baseline, SM-DDPO (Yu et al., 2023), which selects laboratory groups to minimize time-cost while maximizing F1-score. SM-DDPO is a non-personalized method, i.e., it selects the same laboratory groups for all patients.

Time-cost. As discussed, each group of tests is assigned a time-cost by observing time-stamps in the MIMIC-IV database. For all methods, we measure the amount of time taken by averaging the time required for an algorithm to make a prediction across patients.

Implementation Details. BioGPT (Luo et al., 2022) is used as our backbone language model, which is a generative pre-trained transformer (347 million parameters) for biomedical text generation and representation. In the RL phase of training ED-Copilot, for a given patient, our ac-

Table 2. Overall performance of ED-Copilot and baselines in predicting *Critical Outcome* (transfer to ICU or mortality) and *Lengthened ED Stay* (exceeding 24 hours in the ED), as well as the *Average Time-cost* (in minutes) to perform laboratory tests. Sensitivity and specificity are true positive and true negative rates. We report results averaged over three random seeds alongside standard deviations.

	Critical Outcome					**Lengthened ED Stay**				
Model	F1	AUC	Sensitivity	Specificity	Avg. Time-cost	F1	AUC	Sensitivity	Specificity	Avg. Time-cost
Random Forest	0.377 (0.015)	0.807 (0.011)	0.754 (0.012)	0.748 (0.005)	265 Min	0.206 (0.014)	0.698 (0.011)	0.693 (0.016)	0.616 (0.024)	265 Min
XGBoost	0.379 (0.019)	0.807 (0.009)	0.731 (0.017)	0.744 (0.006)	265 Min	0.212 (0.010)	0.679 (0.007)	0.619 (0.020)	0.661 (0.020)	265 Min
LightGBM	0.394 (0.016)	0.813 (0.008)	0.725 (0.012)	0.769 (0.004)	265 Min	0.217 (0.015)	0.705 (0.011)	0.706 (0.017)	0.605 (0.014)	265 Min
3-layer DNN	0.339 (0.032)	0.743 (0.021)	0.676 (0.024)	0.683 (0.011)	265 Min	0.194 (0.031)	0.637 (0.013)	0.649 (0.015)	0.593 (0.014)	265 Min
SM-DDPO	0.353 (0.031)	0.780 (0.020)	0.685 (0.023)	0.763 (0.022)	182 (32) Min	0.183 (0.028)	0.619 (0.012)	0.472 (0.012)	0.739 (0.011)	177 (60) Min
ED-Copilot	**0.413 (0.028)**	**0.820 (0.021)**	**0.750 (0.018)**	**0.779 (0.011)**	**125 (21)** Min	**0.232 (0.023)**	**0.707 (0.015)**	**0.725 (0.018)**	**0.606 (0.015)**	**154 (33)** Min



(a) Critical Outcome F1 (b) Critical Outcome AUC (c) Lengthened ED Stay F1 (d) Lengthened ED Stay AUC

Figure 2. Prediction accuracy and average number of laboratory groups of ED-Copilot with different maximum allowed time to perform laboratory tests. Each point reflects ED-Copilot's F1/AUC (y-axes) at different time upper-bounds.

tion space is restricted to laboratory groups that the patient actually performed. Experiments when ED-Copilot is allowed to select unobserved laboratory groups can be found in Section 6.5.

6. Evaluation Results

In this section, we present our experimental results. Section 6.1 compares the predictive and time-cost performance of ED-Copilot with baselines. Section 6.2 describes our ablation studies. Sections 6.3 and 6.4 establish ED-Copilot's ability to personalize recommendations based on patient severity and achieve consistent performance across subgroups such as age and sex. Lastly, in Section 6.5, we investigate ED-Copilot's performance when it is not restricted to select administered groups.

6.1. Prediction Accuracy and Time-cost

Overall Performance. The prediction performance and time-costs of all methods are presented in Table 2. For critical outcome, ED-Copilot outperforms the next best baseline by 1.9%, 0.7%, 2.5%, and 1% for F1-score, AUC, sensitivity, and specificity, respectively. For lengthened ED stay, ED-Copilot outperforms the next best model by 3.5%, 0.2%, 1.9%, and 0.1% for the four metrics. ED-Copilot halves average time-costs from roughly 4 hours to 2 hours.

In comparison to the other cost-effective baseline, SM-DDPO, ED-Copilot also has significantly better accuracy and lower time-costs. Both methods utilize the same RL algorithm to minimize time-cost, but ED-Copilot benefits from the strong capability of PLM backbone (Sec 6.2) and the personalized diagnostic assistance it enabled (Sec 6.3).

Prediction Accuracy at Different Time-costs. In Figure 2, we investigate how ED-Copilot's prediction performance changes as we vary the maximum time allowed to perform laboratory tests. Performance increases with maximum allowed time and number of laboratory groups selected. ED-Copilot requires relatively few tests to achieve peak performance for lengthened ED stay, while prediction accuracy for critical outcome steadily increases as more tests are allowed. These curves illustrate different requirements for different outcomes and shed light as to how clinicians should allocate resources across tasks.

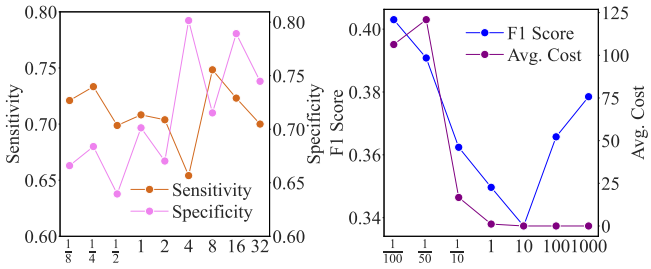
6.2. Ablation Studies

Linearization Technique. Following Hegselmann et al. (2023) and Manikandan et al. (2023), we replace the true laboratory test name with a standard feature name while keeping the result for that laboratory group unchanged. That is, we linearize laboratory results as follows (*feature1 : value1* | *feature2 : value2* ...). Table 3 shows using standard feature names leads to better F1 and AUC score as compared to using the raw laboratory test names. One possible reason is that using raw laboratory test names biases ED-Copilot to pick names over informativeness for prediction.

Feature Importance. We examine the change in ED-Copilot performance when training with and without a given

Table 3. Ablation study on components of ED-Copilot, including linearization techniques, feature importance, and PLM backbones.

Group	Variations	Critical Outcome					Lengthened ED Stay				
		F1	AUC	Sensitivity	Specificity	Avg. Time-cost	F1	AUC	Sensitivity	Specificity	Avg. Time-cost
	ED-Copilot (345M)	**0.413**	**0.820**	**0.750**	**0.779**	**125 Min**	**0.252**	**0.707**	**0.725**	**0.606**	**154 Min**
Linearization	Raw Lab Test Name	0.397	0.777	0.768	0.677	134 Min	0.241	0.695	0.611	0.701	144 Min
Features	w/o. Triage	0.277	0.704	0.679	0.649	—	0.145	0.593	0.532	0.606	—
	w/o. CBC	0.385	0.803	0.692	0.777	—	0.224	0.686	0.696	0.596	—
	w/o. CHEM	0.420	0.827	0.788	0.746	—	0.234	0.702	0.656	0.606	—
Backbone	BioGPT (345M) w/o. RL	0.381	0.810	0.725	0.765	265 Min	0.236	0.718	0.710	0.620	265 Min
	Llama (7B LORA) w/o. RL	0.397	0.798	0.692	0.767	265 Min	0.232	0.701	0.705	0.610	265 Min
	Pythia (70M) w. RL	0.290	0.698	0.574	0.702	166 Min	0.178	0.596	0.555	0.619	126 Min
	GPT2-Medium (345M) w. RL	0.358	0.757	0.621	0.747	133 Min	0.166	0.539	0.498	0.584	96 Min



(a) Sensitivity-Specificity α (b) F1-Cost β

Figure 3. Impact of Hyper-parameters on Sensitivity-Specificity (α) and F1-Cost (β) trade-off when predicting critical outcome.

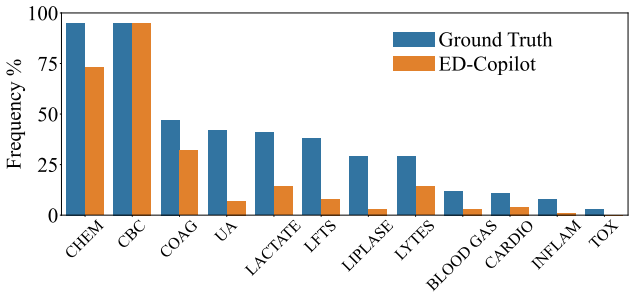


Figure 4. Fraction of patients performing laboratory groups and predicted by ED-Copilot. On average each patient performed 4.7 groups and cost-effective ED-Copilot suggested 2.4 groups.

feature set from top three popular feature sets: triage information, complete blood count (CBC), and chemical group (CHEM). Table 3 shows removing triage features leads to a significant drop in performance. Our ED collaborators confirms this is medically plausible since initial laboratory tests are based on triage information. Our results also show all laboratory groups are not equally important, for example, removing CBC causes a larger drop in performance than removing the chemical laboratory group.

PLM Backbone. We compare BioGPT to GPT2 which is the same size, and a smaller model Pythia (Biderman et al., 2023) that has 70 million parameters. BioGPT is based on GPT2 (Radford et al., 2019), but is further pre-trained on a bio-medical corpus. Table 3 shows both model scale and bio-medical pre-training lead to stronger performance. We also compare Llama (7B) (Touvron et al., 2023) without RL for this task to BioGPT (345M) without RL, and ED-Copilot. Llama is fine-tuned using Low-Rank Adaptation (LORA) (Hu et al., 2022) for efficiency. Llama 7B (LORA) without RL performs comparably to BioGPT without RL. This indicates that scaling up the backbone parameters is promising avenue to boost absolute performance. However, Llama 7B (LORA) underperforms BioGPT with RL (i.e., ED-Copilot), particularly when it comes to time-cost. This is expected as the PLM backbone and the RL serve different roles in the system. PLM is used to predict outcomes while RL minimizes time-cost. Fine-tuning does not address time-costs. These two aspects are both necessary.

Hyper-parameters. Hyper-parameters (α, β) in Equation (8) control the trade-off between prediction accuracy and time-cost in training. Increasing α trades off sensitivity over specificity, while increasing β trades off F1-score over time-cost. Figure 3a shows increasing α leads to trade off sensitivity over specificity. Figure 3b shows increasing β causes ED-Copilot to trade off F1-score over time-cost. The hyper-parameters (α, β) provide a control for ED clinicians to change ED-Copilot's behavior to fit their preferences.

6.3. Analysis on Personalized Diagnostic Assistance

Personalized treatment strategies are often required since patients have diverse needs and present differently. ED-Copilot learns a personalized representation for each patient and naturally supports personalized diagnostic assistance. Previous cost-effective methods based on traditional ML algorithms such as tree-based methods optimize cost-effectiveness at the full population level, e.g., via feature selection. Unlike ED-Copilot, this does not lead to personalized recommendations that are based on triage and previous test results. The following experiments highlight the benefit of personalization via ED-Copilot.

We plot both the fraction of patients receiving each group of tests as well as the fraction of patients predicted by ED-

Table 4. Critical Outcome prediction accuracy at different cohorts, by the category of laboratory groups they performed (Top: Performed at least one of the top two most frequent laboratory groups; Middle: Performed at least one of the middle six laboratory groups; Rare: Performed one or more rare tests in the last four laboratory groups). The total number of positive (critical)/negative cases and positive rate is shown in parentheses.

	W. Top 2 Lab Groups (302/2823,9.6%)			**W. Middle 6 Lab Groups** (299/2603, 10.3%)			**W. Last 4 Lab Groups** (141/817, 14.7%)		
Model	F1	Sensitivity	Specificity	F1	Sensitivity	Specificity	F1	Sensitivity	Specificity
Random Forest (Top 2 groups)	0.330	0.735	0.752	0.335	0.746	0.750	0.405	0.730	0.739
XGBoost (Top 2 groups)	0.361	0.788	0.680	0.374	0.732	0.728	0.433	0.738	0.718
LightGBM (Top 2 groups)	0.401	0.788	0.705	0.409	0.806	0.690	0.462	0.738	0.742
SM-DDPO	0.364	0.760	0.715	0.373	0.762	0.724	0.435	0.732	0.734
ED-Copilot	0.414	0.701	0.788	0.431	0.731	0.774	0.461	0.767	0.720

Table 5. Model performance using only triage information under supervised fine-tuning training.

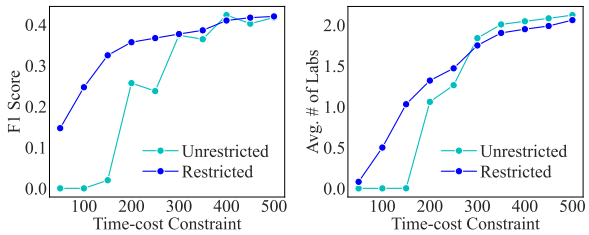
Model	**Critical Outcome**				**Lengthened ED Stay**			
	F1	AUC	Sensitivity	Specificity	F1	AUC	Sensitivity	Specificity
Random Forest	0.346	0.767	0.689	0.731	0.181	0.617	0.597	0.571
XGBoost	0.324	0.736	0.673	0.668	0.165	0.603	0.580	0.560
LightGBM	0.341	0.743	0.624	0.738	0.171	0.615	0.554	0.628
3-layer DNN	0.275	0.712	0.611	0.702	0.155	0.553	0.485	0.597
ED-Copilot (SFT)	0.392	0.792	0.708	0.762	0.199	0.675	0.679	0.557

Copilot to undergo each of these groups in Figure 4. This plot shows significant variation in tests received by ED patients. After the two most common groups (CHEM and CBC), more than half of the patients performed some other tests. ED-Copilot reflects this diversity by recommending a variety of tests based on patients' condition. On the other hand, non-personalized cost-effective methods often withdraw to picking the most frequent groups.

To further understand the benefit of personalized diagnostic assistance, we partition patients in the test set into three cohorts based on the rarity of laboratory groups they were administered. A precise definition of these cohorts can be found in Table 4. The middle and rare cohorts have a larger proportion of positive cases, indicating higher severity. We train tree-based baselines using triage information and results from the two most frequent groups (CHEM and CBC), and compare it to ED-Copilot. Results in Table 6 show ED-Copilot and baselines perform comparably on patients in the low severity cohorts (i.e., patients that present typically). For the middle cohort, ED-Copilot greatly improves F1-score. In the rare cohort which has the highest proportion of critical outcomes (15% versus 10% in middle cohort), ED-Copilot achieves higher significantly higher sensitivity than other methods. This indicates ED-Copilot's ability to provide personalized assistance which results in more equitable care, particularly for high-risk patients.

6.4. Sub-group Analysis

We compare the performance of ED-Copilot to tree-based baselines across sub-groups including gender and age to study if the algorithm is working fairly across all sub-groups of interest in Table 6. ED-Copilot displays consistent performance, and outperforms baselines across subgroups on both prediction tasks.



(a) Accuracy (b) Avg. # of Labs

Figure 5. Comparison of prediction performance on critical outcome of ED-Copilot when restricted or not restricted to tests patients performed in its suggestion at different time-cost constraints.

6.5. Performance of Unrestricted Lab Group Suggestion

Since MIMIC-ED-Assist is an offline retrospective benchmark, we restrict ED-Copilot during training to only select laboratory groups that patients have received. However, this restriction may lead to sub-optimal group suggestion. We investigate ED-Copilot's performance without this restriction through the following experiment.

If ED-Copilot (unrestricted) selects a laboratory group not received by a patient, we impute those results by 0, but add to the time-cost. ED-Copilot (unrestricted) achieves an accuracy of 77% when predicting the next laboratory group, i.e., the group actually received by a given patient, indicating that ED-Copilot (unrestricted) is reasonably good at replicating clinical decisions but there are still differences between model suggestions and historical data.

Next, we compare ED-Copilot (unrestricted) to ED-Copilot (restricted) in terms of F1-score for critical outcome when varying the maximum allowed time. Figure 5 shows that when maximum allowed time is small, ED-Copilot (unrestricted) performs worse than ED-Copilot. This is because if ED-Copilot (unrestricted) picks a laboratory test that was not administered to a patient, this selection does not result in new information being added. However, as the number of allowed laboratory groups increases, ED-Copilot (unre-

Table 6. Sub-group Analysis for Critical Outcome and Lengthened ED Stay.

Model	**Critical Outcome**					**Lengthened ED Stay**				
	F1	AUC	Sensitivity	Specificity	Avg. Time-cost	F1	AUC	Sensitivity	Specificity	Avg. Time-cost
Group: Male, **Positive/Negative (Ratio)**: 155/1329 (10.44%) on critical outcome, 104/1380 (7.01%) on lengthened ED stay										
Random Forest	0.387	0.793	0.785	0.692	265	0.207	0.723	0.644	0.704	265
XGBoost	0.379	0.804	0.726	0.732	265	0.258	0.722	0.692	0.659	265
LightGBM	0.391	0.807	0.762	0.742	265	0.223	0.724	0.769	0.597	265
ED-Copilot	0.377	0.780	0.671	0.775	121	0.242	0.670	0.574	0.702	151
Group: Female, **Positive/Negative (Ratio)**: 143/1608 (8.17%) on critical outcome, 127/1624 (7.25%) on lengthened ED stay										
Random Forest	0.385	0.819	0.730	0.794	265	0.193	0.677	0.646	0.616	265
XGBoost	0.368	0.826	0.766	0.740	265	0.171	0.645	0.661	0.576	265
LightGBM	0.394	0.813	0.766	0.730	265	0.198	0.683	0.669	0.636	265
ED-Copilot	0.413	0.823	0.727	0.819	129	0.215	0.694	0.713	0.621	160
Group: Age 18-30, **Positive/Negative (Ratio)**: 22/341 (6.06%) on critical outcome, 23/340 (6.34%) on lengthened ED stay										
Random Forest	0.277	0.871	0.889	0.838	265	0.197	0.742	0.739	0.738	265
XGBoost	0.413	0.908	0.833	0.884	265	0.306	0.764	0.652	0.824	265
LightGBM	0.390	0.895	0.889	0.913	265	0.259	0.767	0.783	0.629	265
ED-Copilot	0.208	0.849	0.864	0.708	106	0.304	0.823	0.956	0.705	159
Group: Age 31-60, **Positive/Negative (Ratio)**: 97/1169 (7.66%) on critical outcome, 92/1174 (7.27%) on lengthened ED stay										
Random Forest	0.367	0.824	0.739	0.794	265	0.196	0.675	0.620	0.670	265
XGBoost	0.367	0.835	0.750	0.775	265	0.228	0.685	0.674	0.658	265
LightGBM	0.400	0.835	0.802	0.756	265	0.213	0.692	0.739	0.611	265
ED-Copilot	0.409	0.800	0.711	0.823	124	0.209	0.697	0.612	0.664	153
Group: Age 61-90, **Positive/Negative (Ratio)**: 189/1261 (13.03%) on critical outcome, 108/1342 (7.45%) on lengthened ED stay										
Random Forest	0.411	0.780	0.710	0.753	265	0.208	0.707	0.685	0.621	265
XGBoost	0.383	0.785	0.659	0.789	265	0.188	0.655	0.704	0.552	265
LightGBM	0.405	0.779	0.688	0.757	265	0.192	0.698	0.667	0.622	265
ED-Copilot	0.400	0.749	0.667	0.694	129	0.222	0.695	0.731	0.671	156

stricted) matches the performance of ED-Copilot. Restriction to observed laboratory groups is necessary for offline evaluation. In real-world settings, ED-Copilot should serve as a "Co-Pilot" and suggest laboratory groups from the entire set without restriction, and assist ED clinicians in optimizing their workflows.

We also evaluate ED-Copilot on new patients only using triage data, and without any laboratory results which are the information available on patient arrival. We use ED-Copilot only with supervised fine-tuning (SFT), and without RL. Results are displayed in Table 5. Without laboratory results, all models suffer, but ED-Copilot (SFT) outperforms baselines, showing its utility at an early treatment stage.

7. Conclusion

In this paper, we aim at reducing emergency department wait time through time-cost-effective diagnostic assistance. In collaboration with ED clinicians, we curate MIMIC-ED-Assist, a benchmark that contains comprehensive laboratory information annotated with medically relevant groupings alongside key clinical outcomes that serve as useful approximations to ED wait-time. Using MIMIC-ED-Assist, we propose ED-Copilot, an AI system that provides time-cost-effective diagnostic assistance for ED clinicians, recommending informative laboratory groups and flagging patients at high risks of critical outcome and lengthened ED stay. ED-Copilot outperforms baselines by significantly improved prediction accuracy and reduced laboratory time-costs. Its language model backbone allows for personalized diagnostic assistance to better address the needs of severe patients. We believe this work takes a step towards AI-driven assistance in the ED and hope that MIMIC-ED-Assist spurs interest in applying advancements in AI to tackle this critical healthcare challenge.

Acknowledgments

We gratefully acknowledge partial support from NSF grants DMS-2209975, 2015341, NSF grant 2023505 on Collaborative Research: Foundations of Data Science Institute (FODSI), the NSF and the Simons Foundation for the Collaboration on the Theoretical Foundations of Deep Learning through awards DMS-2031883 and 814639, and NIH grant R01GM152718. We also acknowledge support by the Eunice Kennedy Shriver National Institute of Child Health and Human Development of the National Institutes of Health under Award Number K23HD110716. We also thank Yikuan Li from Northwestern University to share their pre-processed sepsis dataset.

Impact Statement

The development of ED-Copilot marks an advancement in medical technology, offering an AI-driven diagnostic tool to improve patient care in ED. By expediting diagnoses, ED-Copilot has the potential to increase efficiency of ED operations and enhance patient care.

From an ethical perspective, the deployment of ED-Copilot carries a significant responsibility on the privacy and security of patient data. As the system will handle sensitive health information, strict measures, for example, internal-accessible only systems and compliant data protection procedures, must be in place to protect against breaches and misuse, ensuring patient confidentiality.

In the broader societal context, the implementation of ED-Copilot aims to address the issue of ED congestion. It is vital to ensure that the benefits of such technology are accessible to all segments of society, regardless of socioeconomic status. Equity in healthcare technology means that all patients, irrespective of their background, should have the opportunity to benefit from advancements like ED-Copilot. Moreover, the development of ED-Copilot should support healthcare professionals, not replace them, and should be viewed as a tool to assist medical staff, allowing them to focus on the more nuanced aspects of patient care.

References

Agarwal, A., Tan, Y. S., Ronen, O., Singh, C., and Yu, B. Hierarchical shrinkage: Improving the accuracy and interpretability of tree-based models. In Chaudhuri, K., Jegelka, S., Song, L., Szepesvari, C., Niu, G., and Sabato, S. (eds.), *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 111–135. PMLR, 17–23 Jul 2022. URL https://proceedings.mlr.press/v162/agarwal22b.html.

Agarwal, A., Kenney, A. M., Tan, Y. S., Tang, T. M., and Yu, B. MDI+: A Flexible Random Forest-Based Feature Importance Framework, 2023.

Bejnordi, B. E., Veta, M., Van Diest, P. J., Van Ginneken, B., Karssemeijer, N., Litjens, G., Van Der Laak, J. A., Hermsen, M., Manson, Q. F., Balkenhol, M., et al. Diagnostic assessment of deep learning algorithms for detection of lymph node metastases in women with breast cancer. *Jama*, 318(22):2199–2210, 2017.

Biderman, S., Schoelkopf, H., Anthony, Q. G., Bradley, H., O'Brien, K., Hallahan, E., Khan, M. A., Purohit, S., Prashanth, U. S., Raff, E., et al. Pythia: A suite for analyzing large language models across training and scaling. In *International Conference on Machine Learning*, pp. 2397–2430. PMLR, 2023.

Breiman, L. Random forests. *Machine learning*, 45:5–32, 2001.

Chen, T. and Guestrin, C. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, pp. 785–794, 2016.

DeAnda, R. Stop the Bottleneck: Improving Patient Throughput in the Emergency Department. *Journal of Emergency Nursing*, 44(6):582–588, 2018. ISSN 0099-1767. doi: https://doi.org/10.1016/j.jen.2018.05.002. URL https://www.sciencedirect.com/science/article/pii/S0099176717305962.

Harutyunyan, H., Khachatrian, H., Kale, D. C., Ver Steeg, G., and Galstyan, A. Multitask learning and benchmarking with clinical time series data. *Scientific Data*, 6(1), June 2019. ISSN 2052-4463. doi: 10.1038/s41597-019-0103-9. URL http://dx.doi.org/10.1038/s41597-019-0103-9.

Hegselmann, S., Buendia, A., Lang, H., Agrawal, M., Jiang, X., and Sontag, D. Tabllm: Few-shot classification of tabular data with large language models. In *International Conference on Artificial Intelligence and Statistics*, pp. 5549–5581. PMLR, 2023.

Hu, E. J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., and Chen, W. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=nZeVKeeFYf9.

Jarvis, P. R. Improving emergency department patient flow. *Clinical and Experimental Emergency Medicine*, 3(2):63–68, 2016. doi: 10.15441/ceem.16.127. URL https://doi.org/10.15441/ceem.16.127.

Johnson, A., Bulgarelli, L., Pollard, T., Celi, L. A., Mark, R., and Horng, S. MIMIC-IV-ED (version 2.2). PhysioNet, Jan 2023a. URL https://physionet.org/content/mimic-iv-ed/2.2/. Version: 2.2.

Johnson, A. E., Bulgarelli, L., Shen, L., Gayles, A., Shammout, A., Horng, S., Pollard, T. J., Hao, S., Moody, B., Gow, B., et al. MIMIC-IV, a freely accessible electronic health record dataset. *Scientific data*, 10(1):1, 2023b.

Kang, S. Y., Cha, W. C., Yoo, J., Kim, T., Park, J. H., Yoon, H., Hwang, S. Y., Sim, M. S., Jo, I. J., and Shin, T. G. Predicting 30-day mortality of patients with pneumonia in an emergency department setting using machine-learning models. *Clinical and Experimental Emergency Medicine*, 7(3):197, 2020.

Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q., and Liu, T.-Y. Lightgbm: A highly efficient gradient boosting decision tree. *Advances in neural information processing systems*, 30, 2017.

Komorowski, M., Celi, L. A., Badawi, O., Gordon, A. C., and Faisal, A. A. The artificial intelligence clinician learns optimal treatment strategies for sepsis in intensive care. *Nature medicine*, 24(11):1716–1720, 2018.

Kornblith, A. E., Singh, C., Devlin, G., Addo, N., Streck, C. J., Holmes, J. F., Kuppermann, N., Grupp-Phelan, J., Fineman, J., Butte, A. J., et al. Predictability and stability testing to assess clinical decision instrument performance for children after blunt torso trauma. *PLOS Digital Health*, 1(8):e0000076, 2022.

Levin, S., Toerper, M., Hamrock, E., Hinson, J. S., Barnes, S., Gardner, H., Dugas, A., Linton, B., Kirsch, T., and Kelen, G. Machine-learning-based electronic triage more accurately differentiates patients with respect to clinical outcomes compared with the emergency severity index. *Annals of Emergency Medicine*, 71(5):565–574.e2, 2018. ISSN 0196-0644. doi: https://doi.org/10.1016/j.annemergmed.2017.08.005. URL https://www.sciencedirect.com/science/article/pii/S0196064417314427.

Li, L., Georgiou, A., Vecellio, E., Eigenstetter, A., Toouli, G., Wilson, R., and Westbrook, J. I. The effect of laboratory testing on emergency department length of stay: a multihospital longitudinal study applying a cross-classified random-effect modeling approach. *Academic Emergency Medicine*, 22(1):38–46, 2015.

Luo, R., Sun, L., Xia, Y., Qin, T., Zhang, S., Poon, H., and Liu, T.-Y. BioGPT: generative pre-trained transformer for biomedical text generation and mining. *Briefings in Bioinformatics*, 23(6):bbac409, 2022.

Manikandan, H., Jiang, Y., and Kolter, J. Z. Language models are weak learners. In Oh, A., Naumann, T., Globerson, A., Saenko, K., Hardt, M., and Levine, S. (eds.), *Advances in Neural Information Processing Systems*, volume 36, pp. 50907–50931. Curran Associates, Inc., 2023.

Miyakis, S., Karamanof, G., Liontos, M., and Mountokalakis, T. D. Factors contributing to inappropriate ordering of tests in an academic medical department and the effect of an educational feedback strategy. *Postgraduate medical journal*, 82(974):823–829, 2006.

Purushotham, S., Meng, C., Che, Z., and Liu, Y. Benchmarking deep learning models on large healthcare datasets. *Journal of Biomedical Informatics*, 83:112–134, 2018. ISSN 1532-0464. doi: https://doi.org/10.1016/j.jbi.2018.04.007. URL https://www.sciencedirect.com/science/article/pii/S1532046418300716.

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., and Sutskever, I. Language models are unsupervised multitask learners. 2019.

Raffin, A., Hill, A., Gleave, A., Kanervisto, A., Ernestus, M., and Dormann, N. Stable-baselines3: Reliable reinforcement learning implementations. *Journal of Machine Learning Research*, 22(268):1–8, 2021. URL http://jmlr.org/papers/v22/20-1364.html.

Sartini, M., Carbone, A., Demartini, A., Giribone, L., Oliva, M., Spagnolo, A. M., Cremonesi, P., Canale, F., and Cristina, M. L. Overcrowding in emergency department: Causes, consequences, and solutions-a narrative review. *Healthcare (Basel, Switzerland)*, 10(9):1625, 2022. doi: 10.3390/healthcare10091625. URL https://doi.org/10.3390/healthcare10091625.

Savioli, G., Ceresa, I. F., Gri, N., Bavestrello Piccini, G., Longhitano, Y., Zanza, C., Piccioni, A., Esposito, C., Ricevuti, G., and Bressan, M. A. Emergency department overcrowding: Understanding the factors to find corresponding solutions. *Journal of Personalized Medicine*, 12(2):279, 2022. doi: 10.3390/jpm12020279. URL https://doi.org/10.3390/jpm12020279.

Schulman, J., Wolski, F., Dhariwal, P., Radford, A., and Klimov, O. Proximal policy optimization algorithms, 2017.

Singhal, K., Tu, T., Gottweis, J., Sayres, R., Wulczyn, E., Hou, L., Clark, K., Pfohl, S., Cole-Lewis, H., Neal, D., Schaekermann, M., Wang, A., Amin, M., Lachgar, S., Mansfield, P., Prakash, S., Green, B., Dominowska, E., y Arcas, B. A., Tomasev, N., Liu, Y., Wong, R., Semturs, C., Mahdavi, S. S., Barral, J., Webster, D., Corrado, G. S., Matias, Y., Azizi, S., Karthikesalingam, A., and Natarajan, V. Towards Expert-Level Medical Question Answering with Large Language Models, 2023.

Tan, Y. S., Singh, C., Nasseri, K., Agarwal, A., and Yu, B. Fast interpretable greedy-tree sums (FIGS). *arXiv preprint arXiv:2201.11931*, 2022.

Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.-A., Lacroix, T., Rozière, B., Goyal, N., Hambro, E., Azhar, F., et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023.

Wang, S., McDermott, M. B. A., Chauhan, G., Ghassemi, M., Hughes, M. C., and Naumann, T. MIMIC-Extract: a data extraction, preprocessing, and representation pipeline for MIMIC-III. In *Proceedings of the ACM Conference on Health, Inference, and Learning*, ACM

CHIL '20. ACM, April 2020. doi: 10.1145/3368555.3384469. URL http://dx.doi.org/10.1145/3368555.3384469.

Wardi, G., Carlile, M., Holder, A., Shashikumar, S., Hayden, S. R., and Nemati, S. Predicting progression to septic shock in the emergency department using an externally generalizable machine-learning algorithm. *Annals of emergency medicine*, 77(4):395–406, 2021.

Xie, F., Zhou, J., Lee, J. W., Tan, M., Li, S., Rajnthern, L. S., Chee, M. L., Chakraborty, B., Wong, A.-K. I., Dagan, A., et al. Benchmarking emergency department prediction models with machine learning and public electronic health records. *Scientific Data*, 9(1):658, 2022.

Yagnik, N., Jhaveri, J., Sharma, V., Pila, G., Ben, A., and Shang, J. MedLM: Exploring Language Models for Medical Question Answering Systems, 2024.

Yang, X., Yu, Z., Guo, Y., Bian, J., and Wu, Y. Clinical Relation Extraction Using Transformer-based Models, 2021.

Yang, X., Chen, A., PourNejatian, N., Shin, H. C., Smith, K. E., Parisien, C., Compas, C., Martin, C., Costa, A. B., Flores, M. G., Zhang, Y., Magoc, T., Harle, C. A., Lipori, G., Mitchell, D. A., Hogan, W. R., Shenkman, E. A., Bian, J., and Wu, Y. A large language model for electronic health records. *NPJ Digital Medicine*, 5(1):194, 2022. doi: 10.1038/s41746-022-00742-2.

Yu, Z., Li, Y., Kim, J., Huang, K., Luo, Y., and Wang, M. Deep Reinforcement Learning for Cost-Effective Medical Diagnosis. *arXiv preprint arXiv:2302.10261*, 2023.

A. Triage and Laboratory Test Group

In Table 7, we include triage assessments and 12 laboratory groups with 68 tests and their estimated time-costs.

Table 7. Triage and Lab Test Group

Triage	Age Gender Heart rate Respiratory rate Systolic blood pressure Temperature Past ICU/ED/hospital visiting frequency	Diastolic blood pressure Self-reported pain Emergency severity index acuity Chief complaint Oxygen saturation Past commodities
Complete Blood Count (CBC) (30 min)	Hematocrit White Blood Cells Hemoglobin Red Blood Cells Mean Corpuscular Volume Mean Corpuscular Hemoglobin Mean Corpuscular Hemoglobin Concentration Red Blood Cell Distribution Width Platelet Count Basophils Eosinophils Lymphocytes	Neutrophils Red Cell Distribution Width (Standard Deviation) Absolute Lymphocyte Count Absolute Basophil Count Absolute Eosinophil Count Absolute Monocyte Count Absolute Neutrophil Count Bands Atypical Lymphocytes Nucleated Red Cells Monocytes
Chemistry (CHEM) (60 min)	Urea Nitrogen Creatinine Sodium Chloride Bicarbonate	Glucose (Chemistry) Potassium Anion Gap Calcium, Total
Coagulation (COAG)(48 min)	Prothrombin Time International Normalised Ratio	Partial thromboplastin time
Urinalysis (UA) (40 min)	PH (Urine) Specific Gravity Red Blood Count (Urine) White Blood Count (Urine) Epithelial Cells	Protein Hyaline Casts Ketone Urobilinogen Glucose (Urine)
Lactate (4 min)	Lactate	
Liver Function (LFTs)(104 min)	Alkaline Phosphatase Asparate Aminotransferase (AST) Alanine Aminotransferase (ALT)	Bilirubin, Total Albumin
Lipase (100 min)	Lipase	
Electrolyte (LYTES)(89 min)	Magnesium	Phosphate
Cardiovascular (CARDIO) (122 min)	NT-proBNP	Troponin T
Blood Gas (12 min)	Potassium, Whole Blood PH (Blood Gas) Calculated Total CO2 Base Excess	PO2 PCO2 Glucose (Blood Gas) Sodium, Whole Blood
Toxicology (TOX) (70 min)	Ethanol	
Inflammation (INFLAM) (178 min)	Creatine Kinase (CK)	C-Reactive Protein

B. Training Details

We use a 3-layer neural network for all MLPs in ED-Copilot. Due to large class imbalance, we use class weights when training the diagnostic predictor p_ψ. All experiments, training and hyper-parameter tuning are conducted on one NVIDIA RTX A6000 GPU. During the RL phase of training ED-Copilot, we restrict the action space to be on laboratory groups that were administered to patients. That is, we do not select laboratory groups that patients have not received. To ensure sufficient policy sampling on experience replay buffer and obtain a large batch size, we freeze the weights of our PLM. To train the policy π_η, we use the PPO algorithm (Schulman et al., 2017). The loss function employed is as the following for

Algorithm 1 Proximal Policy Optimization (PPO)

for iteration $= 0, 1, \ldots$ **do**
  for actor $= 1, 2, \ldots, N_{\text{actor}}$ **do**
    Run policy $\pi_{\eta_{\text{old}}}$ in environment for T timesteps and save all observations in experience buffer
    Compute estimated advantage $\hat{A}_1, \ldots, \hat{A}_i$
  end for
  Optimize surrogate $\mathcal{L}_{\text{rl}}$ w.r.t η, with D epochs and minibatch size $M \leq N_{\text{actor}} T$
  $\eta_{\text{old}} \leftarrow \eta$
end for

Table 8. Hyperparameters Configurations for ED-Copilot

Hyperparameter	**Supervised Fine-tuning**	**Reinforcement Learning**
Learning Rate	1e-5	
DNN Hidden size (3-layer)	1024	
Optimizer	AdamW	
Adam ϵ	1e-8	
Adam Betas (β_1, β_2)	(0.9, 0.999)	
Weight decay	0.01	
Batch Size	32	128
Epochs	15	10
Max sequence length	656	656
Class Weight	10	-
Warmup percentage	0.1	-
Buffer Steps	-	2048
Timesteps PPO trained per epoch	-	20000
The penalty ratio between false positive and false negative α	-	15
The penalty ratio between false positive and laboratory cost β	-	$\frac{1}{100}$

hyper-parameters c_1 and c_2:

$$\mathcal{L}_{\text{rl}}(\eta; \eta^{old}) = \mathcal{L}_{\text{clip}}(\eta; \eta^{old}) - c_1 \mathcal{L}_{\text{value}}(\eta) + c_2 \text{Entropy}[\pi_\eta],$$
$$\mathcal{L}_{\text{clip}}(\eta; \eta^{old}) = \hat{\mathbb{E}}_i \left[\min \left(\frac{\pi_\eta(a_i|\boldsymbol{h}_{<i})}{\pi_{\eta^{old}}(a_i|\boldsymbol{h}_{<i})} \hat{A}_i, \text{clip} \left(\frac{\pi_\eta(a_i|\boldsymbol{h}_{<i})}{\pi_{\eta^{old}}(a_i|\boldsymbol{h}_{<i})}, 1-\epsilon, 1+\epsilon \right) \hat{A}_i \right) \right],$$
$$\mathcal{L}_{\text{value}}(\eta) = \hat{\mathbb{E}}_i \left[(V_\eta(\boldsymbol{h}_{<i}) - V_i)^2 \right].$$

Here $\boldsymbol{h}_{<i}$ is the hidden representation from PLM of `[EOS]`$_{<i}$ in this patient's linearized sequence in state s_i. $\mathcal{L}_{\text{clip}}$ is clipped surrogate loss and $\hat{A}_i$ is estimated advantages which are regularized by value function $\mathcal{L}_{\text{value}}$. Entropy$[\pi_\eta]$ denotes an entropy term over the states, and $\hat{\mathbb{E}}_i$ is the empirical average over the collected dataset. We used a masked actor-critic network with package stable-baseline3 (Raffin et al., 2021). See algorithm 1 for a high-level description of this method.

We list the hyper-parameters in Table 8, including the supervised fine-tuning and reinforcement learning stage. In the RL stage, we use grid-search to tune α and β to balance the trade-off between accuracy and cost. The search scope for $\alpha \in \{\frac{1}{8}, \frac{1}{4}, \frac{1}{2}, 1, 2, 4, 8, 15, 16, 32, 64, 256\}$ and $\beta \in \{\frac{1}{100}, \frac{1}{50}, \frac{1}{20}, \frac{1}{10}, 1, 10, 100, 1000\}$. The parameters in PPO that are not specified are assigned the default values found in the Python package (Schulman et al., 2017).

C. More Baseline Results

In Table 9, we present tree models' results with different imputation methods: (1) Mean imputation. (i.e. replacing missing values with the mean value of the non-missing data for the particular feature). (2) Median imputation. (i.e. replacing missing values with the median value of the non-missing data for the particular feature) (3) Zero imputation. (i.e. replacing missing values by 0). (4) A dummy indicator to encode missing values, which is used in XGBoost and LightGBM.

Table 9. Results with Different Imputation Methods

Model	**Method**	**Critical Outcome**				**Lengthened ED Stay**			
		F1	AUC	Sensitivity	Specificity	F1	AUC	Sensitivity	Specificity
Random Forest	Mean	0.377	0.807	0.754	0.748	0.198	0.701	0.736	0.578
	Median	0.355	0.807	0.793	0.709	0.206	0.698	0.693	0.616
	Zero	0.367	0.808	0.738	0.748	0.194	0.702	0.628	0.674
XGBoost	Mean	0.328	0.768	0.715	0.697	0.212	0.678	0.593	0.683
	Median	0.358	0.783	0.731	0.708	0.197	0.674	0.658	0.601
	Zero	0.374	0.818	0.735	0.754	0.212	0.680	0.619	0.662
	Dummy	0.379	0.807	0.731	0.744	0.190	0.656	0.680	0.557
LightGBM	Mean	0.390	0.812	0.764	0.732	0.217	0.705	0.707	0.606
	Median	0.394	0.813	0.725	0.769	0.190	0.681	0.688	0.599
	Zero	0.393	0.814	0.735	0.774	0.209	0.702	0.697	0.623
	Dummy	0.379	0.807	0.731	0.744	0.192	0.671	0.654	0.629