PORTER: PORTABLE REPRESENTATIONS FOR TRANSFERABLE ELECTRONIC HEALTH RECORDS

Anonymous authors
Paper under double-blind review

ABSTRACT

Clinical prediction models often degrade across sites because differences in clinical practice and documentation induce site-specific joint distributions over observed records and outcomes. These shifts arise from the underlying data-generating process and therefore cannot be fully addressed by increasing downstream model capacity alone. We instead focus on constructing representations that support transferable prediction across sites. We propose Portable Representations for Transferable Electronic Health Records (PORTER), a summarize-then-embed pipeline built using a clinical schema designed to reduce site-specific variation, alongside a fine-tuning method to addresses variation unresolvable by schema alone. We design the schema to guide the LLM to preserve predictive information while ignoring site-specific noise. However, a site-independent representation alone cannot capture how predictions should be adapted to a new site. To address this limitation, we build on the site-invariant representation and further adapt its predictions by distilling target-site policies from a small support set. Across three ICU datasets, six transfer directions, and five clinical prediction tasks, schema-only PORTER achieves the best average rank among all 13 methods in zero-shot transfer. Its target-site distillation procedure also achieves the best average rank in the few-shot setting, with the largest gains on tasks that transfer most poorly under zero-shot setting. We also experiment with PORTER on longitudinal EHR data where we find PORTER improves with increasing context length. Our work demonstrates that LLMs can be used not only to encode clinical text, but to generate representations explicitly designed to improve cross-site clinical prediction.

1 INTRODUCTION

Machine learning for healthcare has developed rapidly, with strong performance reported across clinical prediction tasks (Rajpurkar et al., 2022; Harutyunyan et al., 2019). This progress has been enabled in part by large EHR datasets such as MIMIC-IV (Johnson et al., 2023). But clinical prediction models often degrade under shifts across sites (Futoma et al., 2020; Zhang et al., 2021). This is in large part because neural networks often learn site-specific shortcuts rather than learning site-agnostic and generalizable insights (Nestor et al., 2019).

Similar failures have been demonstrated across independent clinical trials and patient cohorts (Chekroud et al., 2024). Much of clinical machine learning has focused on improving predictive performance through more expressive models and increasingly large datasets (Shickel et al., 2018; Xiao et al., 2018; Rajkomar et al., 2018). But site-specific measurement, documentation, and treatment policies can change the relationship between the observed EHR and downstream outcomes (Finlayson et al., 2021; Subbaswamy & Saria, 2020).

EHRs depend on patient state and the process through which that state is observed. For example, whether a test is ordered and how frequently it is measured depend on patient condition and care policies, making patterns of observation informative rather than ignorable missingness (Agniel et al., 2018; Sisk et al., 2021). These site-dependent observation processes can make associations learned at one hospital unreliable at another.

We believe there is a different approach. Instead of asking *how do we train a transferable model?*, we ask *how do we create representations that enable models to transfer across sites?* This motivates a new class of methods for building transferable models by building transferable representations.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

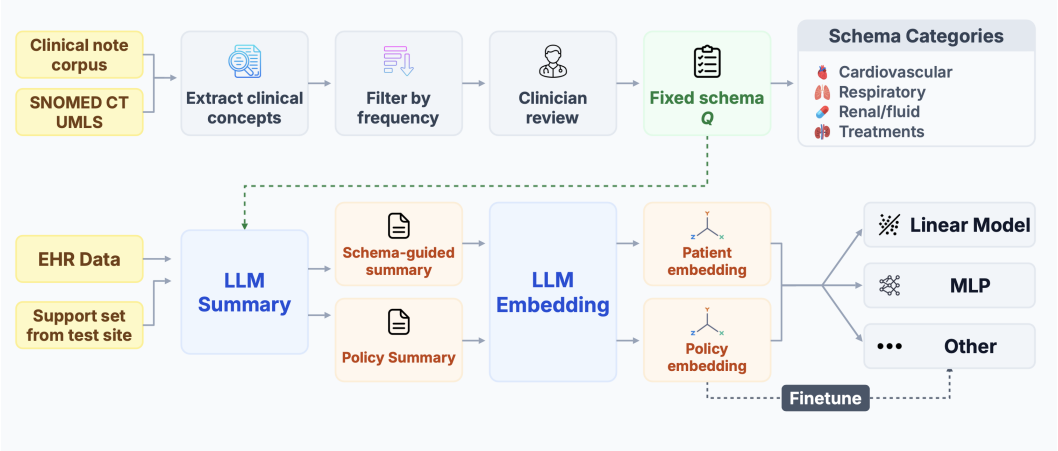

Figure 1: **Overview of PORTER**. **Top.** A fixed clinical schema Q is constructed by extracting common clinical concepts from a clinical note corpus and SNOMED CT/UMLS (standardized biomedical vocabularies and concept relations), filtering concepts by frequency, and performing clinician review. The resulting schema defines a shared set of clinical categories used across sites. **Bottom.** PORTER uses Q to guide an LLM in transforming site-specific EHRs into schema-aligned summaries, which are embedded to obtain a patient representation Z. This representation reduces site-dependent variation while preserving predictive information and can be reused across downstream models and clinical prediction tasks. A target-site support set may additionally be used to distill site-specific policy context and fine-tune the downstream prediction model.

Prior work falls into two groups. The first serializes structured records as text and applies a language model directly. GenHPF uses hierarchical text to bridge EHR schemas (Hur et al., 2024), TabLLM serializes tabular records using natural-language templates (Hegselmann et al., 2023), and Hegselmann et al. (2026) replace medical codes with descriptions and embed the resulting text. These methods largely preserve the structure and fields of the source record. The second group uses an LLM to transform the record before downstream modeling, building on clinical summarization (Pivovarov & Elhadad, 2015; Wang et al., 2021; Van Veen et al., 2024). Record2Vec summarizes ICU records before embedding them (Ji et al., 2026), while Demirel et al. (2026) use LLM-generated rubrics to extract and organize task-relevant information.

We propose **Portable Representations for Transferable Electronic Health Records (PORTER)**, a summarize-then-embed pipeline for creating transferable representations. Different from the first group, PORTER *summarizes* the record and embeds the summary, instead of embedding the record directly. Compared to the second group, PORTER also transforms records semantically, but uses a schema designed based on the DGP of EHRs that aims to create transferable representations.

Our contributions are as follows:

- We present an approximate graph of the data generating process (DGP) of EHR data that captures three common causes of distribution shift across sites: patient case-mix shift, care policies, and outcome policies. We use this graph to show when transferable representations can be obtained.
- We introduce PORTER, a summarize-then-embed pipeline that creates transferable representations from raw EHR. PORTER addresses patient case-mix shift and care policies via *schema-guided semantic normalization* and outcome policies via *target-site policy distillation*.
- We conduct empirical evaluations on ICU datasets across representation generation methods, downstream models, prediction tasks, and prompt ablations, and show that PORTER improves transferability, highlighting the importance of representation quality, prompt design, and target-site adaptation. On longitudinal EHR datasets, we evaluate the effect of input context window length on the transferability of PORTER and other methods, and show that more context generally improves performance on prediction tasks.

2 PROBLEM

A clinical prediction problem typically requires a map ϕ from raw EHR X to some outcome of interest Y. EHRs contain measurements M and treaments T, so $X = (M, T)$. A clinical state C is an unobserved variable that represents patient physiology. For a fixed C, sites S (i.e. hospitals) have different policies that affects observed measurements, administered treatments, and the outcomes. We partition these policies into I_X and I_Y which directly affects X and Y, respectively. In terms of probability distributions, I_X changes $p(X \mid C)$ and I_Y changes $p(Y \mid X, C)$. For an example of I_X, hospitals differ in blood tests frequency and whether they provide life-saving treatment to extremely premature infants. For I_Y, hospitals differ in nurse staffing levels, which can affect inpatient mortality. We consider the clinical prediction settings compatible with Figure 2, which allows variations in I_X, I_Y and patient case-mix. Although not all clinical prediction problems follow this graph, it captures common sources of variation across sites (Sisk et al., 2021; Futoma et al., 2021; Rysavy et al., 2015; Needleman et al., 2011). Under Figure 2, generally $p(Y \mid X, S = s) \neq p(Y \mid X)$. Learning $p(Y \mid X)$ directly requires marginalizing over s, which can be difficult due to the limited number of training sites available. One way to solve this problem is to construct a *transferable* representation Z such that $p(Y \mid Z, S = s)$ is the same for all sites, i.e., $Y \perp S \mid Z$. We identify two sufficient conditions for transferability (proof in Appendix B):

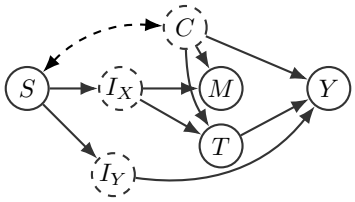

Figure 2: **Assumed data-generating process.** S denotes site, C latent clinical state, M measurments, T treatments, and Y patient outcome. I_X denotes site-specific measurement and treatment policies (e.g. how often to order blood tests, whether to perform invasive interventions), while I_Y denotes site-specific outcome policies (e.g. nurse staffing level which affects inpatient mortality). Dashed nodes represent latent variables, and the dashed bidirected edge represents unmodeled common causes of S and C.

Condition 1 (Invariance)**.** *A representation Z is **invariant** if $Z \perp S \mid (C, T, I_Y)$.*

Condition 2 (Sufficiency)**.** *A representation Z is **sufficient** if $Y \perp (C, T, I_Y) \mid (Z, S)$.*

Information about C, T, I_Y are available on different levels. C and T are expressed in the EHR of an individual patient. In contrast, I_Y cannot generally be inferred from a single patient and must be estimated from a sample of patients at the site. This motivates decomposing the representation as $Z = (Z_X, Z_Y)$, where Z_X is constructed from a single patient and Z_Y from a sample of patients. Our problem is therefore to construct these two components so that their combination is approximately invariant and sufficient.

3 METHOD

PORTER constructs two representations, Z_X and Z_Y, from different sources of information. Subsection 3.1 constructs Z_X from a patient's EHR to approximate the relevant information in clinical state C and treatments T. Subsection 3.2 constructs Z_Y from a small labeled support set from the target site to approximate the relevant information in its outcome policies I_Y. A downstream predictor uses the concatenation $Z = (Z_X, Z_Y)$ as input.

3.1 SCHEMA-GUIDED SEMANTIC ENCODING

PORTER uses an LLM to summarize each EHR X conditioned on a clinical schema Q, then embeds the resulting summary: $Z_X \coloneqq \mathrm{Embed}(\mathrm{LLM}(Q, X))$. Without the schema, a general summary may omit information needed to predict Y or retain differences in how sites record the same information. We therefore provide a schema Q as context at the start of the input to condition the format and content of the LLM summary (Arora et al., 2025; Tu et al., 2025).

We construct the schema Q by extracting SNOMED CT concepts from a clinical notes corpus using MedCAT. We take the components from SNOMED CT (Donnelly, 2006), a clinical terminology mapped to UMLS (Bodenreider, 2004) and ICD-10 (Fung & Xu, 2012). We use MedCAT (Kraljevic et al., 2021) to identify SNOMED CT concepts in a corpus of clinical notes and retain high-level

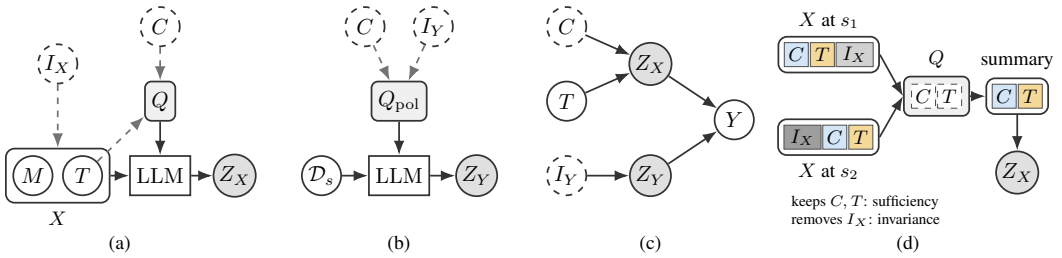

Figure 3: **(a) Construction of Z_X in Subsection 3.1.** PORTER summarizes each patient's EHR X using the shared clinical schema Q, then embeds the summary to obtain Z_X. We need Z_X to encode two types of information: C and T. For C, the schema contains components corresponding to different organ systems and diagnoses. For T, the schema contains components for treatments and device support. **(b) Construction of Z_Y in Subsection 3.2.** PORTER uses a small labeled support set from the target site to summarize site-level factors that affect the outcome, then embeds this summary to obtain Z_Y. Intuitively Z_Y approximates I_Y. **(c)** Y depends on C, T and I_Y only through Z_X and Z_Y. **(d) How Q aims at invariance and sufficiency.** Suppose patients with the same C and T are recorded at two sites, differing only in I_X. Q asks only for C and T, so the summary keeps them and should be the same at both sites. These are sufficiency and invariance. Dashed circles are latent variables, and dashed arrows and boxes show what a prompt asks for.

concepts that clinicians record frequently. Attending physicians then remove concepts that are redundant, poorly defined, or clinically uninformative. The prompt follows how clinicians write notes, first requesting one line describing the overall clinical course. Under each component, it specifies the variables to consider and asks the LLM to summarize their values and changes over time when supported by the record. Prompt details are provided in Appendix C.

The schema Q is designed to guide the LLM to generate summaries that satisfy invariance and sufficiency. For invariance, the schema specifies a set of clinical concepts that emphasize the larger clinical state over details that are more likely to differ across sites. For sufficiency, the schema covers major dimensions of clinical state, treatments, and support devices to preserve outcome-relevant information in C and T. Figure 3(d) illustrates both properties. The prompt additionally asks the LLM to summarize changes and trends on invariance and sufficiency when supported by the record, capturing temporal information relevant to the patient's clinical state. In contrast, text-alignment methods such as Markdown (Hegselmann et al., 2026) do not satisfy invariance, while methods such as SOFA scoring (Vincent et al., 1996) and Record2Vec (Ji et al., 2026) do not satisfy sufficiency.

3.2 Target-site policy distillation

A patient-level representation alone can be insufficient for transfer. The same patient can have different outcomes at different sites due to different outcome policies I_Y. For example, discharge policies affect length-of-stay. I_Y is a property of the site. Thus in general it cannot be inferred from a single patient and must be estimated from a sample of patients at the site.

PORTER explicitly estimates the outcome policies. Given a labeled support set from the target site $\mathcal{D}_s = \{x^{(i)}, y^{(i)}\}_{i=1}^{k}$, PORTER constructs $Z_Y \coloneqq \text{Embed}(\text{LLM}(Q_{\text{pol}}, \mathcal{D}_s, \text{task description}))$, where Q_{pol} is a prompt that instructs the LLM to summarize task-relevant policies from the given support set $\mathcal{D}_s$. Figure 3(b) shows this construction. A general summary of the support set might be unsuitable for certain tasks. We provide the LLM with the prediction task, specify the information to include, and restricts the summary to qualitative patterns supported by the examples. PORTER concatenates $Z = (Z_X, Z_Y)$ as the final representation Z used by downstream models. Note Z_X is different across patients but Z_Y is the same for patients at that site. Prompt details are provided in Appendix C.1.

The schema and distillation address complementary sources of distribution shift. Schema-guided semantic encoding aims to approximate information in clinical state and treatments (C, T) from EHRs generated under different measurement and treatment policies I_X. Policy distillation aims to approximate information about the target site's outcome policies I_Y from a labeled support set.

Table 1: Zero-shot transfer from HIRID to MIMIC, with the Qwen3.6 summarizer and the xLSTM-Mixer head. The other five transfer directions are omitted here for space. Table 7 gives the full table with all six directions. The columns on the right summarize all six directions. *Mean rank* is each method's position among the thirteen by its average rank over the thirty columns of the six directions, with that average in parentheses. *Cumulative occurrences* counts the columns, out of thirty, in which a method places in the top-1, 2, 3. The four groups are controls, the seven representations that do not use language models, the three representations that use a language model, and ours.

Method	HIRID → MIMIC					···	All six directions			
	mort AUROC↑	los MSE↓	forecast MSE↓	lab AUROC↑	drug AUROC↑		Mean rank↓	Cumulative occurrences ↑ Top-1	Top-2	Top-3
ATTACK	0.498	3.550	0.264	0.376	0.659	···	11 (10.1)	0	0	1
RANDOM	0.500	3.541	0.410	0.384	0.659	···	12 (11.1)	0	0	0
CLMBR-T (Wornow et al., 2023)	0.754	3.044	1.054	0.486	0.765	···	6 (7.1)	0	3	6
GARLIC (Wang et al., 2026)	0.584	3.098	0.321	0.419	0.681		10 (9.4)	0	0	0
KEDGN (Luo et al., 2024)	0.300	4.081	0.447	0.407	0.375	HIRID→PPICU	13 (11.1)	0	0	0
MIRA (Li et al., 2026)	0.646	3.035	0.773	0.404	0.778	MIMIC→HIRID	9 (8.0)	0	1	1
TabICL (Qu et al., 2025)	0.759	3.725	**0.157**	0.430	0.690	MIMIC→PPICU	7 (7.2)	2	3	4
MEWS (Subbe et al., 2001)	0.783	2.874	0.512	0.424	0.722	PPICU→HIRID	8 (7.3)	0	0	0
SOFA (Vincent et al., 1996)	0.801	2.873	0.346	0.465	0.849	PPICU→MIMIC	5 (5.0)	4	7	11
Markdown (Hegselmann et al., 2026)	0.767	2.777	0.269	0.468	0.838		4 (4.3)	5	9	13
Record2Vec (Ji et al., 2026)	0.819	2.963	0.206	0.452	0.861	···	3 (4.0)	1	9	15
Template (Demirel et al., 2026)	0.751	2.835	0.233	**0.494**	0.897	···	2 (3.6)	7	11	17
PORTER	**0.825**	**2.773**	0.225	0.447	**0.915**	···	**1 (2.9)**	**11**	**17**	**22**

Combined, $Z = (Z_X, Z_Y)$ approximates the outcome-relevant information in (C, T, I_Y). Figure 3(c) shows this goal. We evaluate whether the resulting representation improves prediction at unseen sites.

4 Datasets & Experiments

Datasets, tasks and metrics. We use three ICU datasets: MIMIC-IV contains 57,212 stays with 60 variables, HiRID contains 32,216 stays with 64 variables, PPICU contains 39,000 stays with 75 variables (Johnson et al., 2023; Faltys et al., 2021; Rojas et al., 2025). The variables cover laboratory tests, vital signs, and clinical interventions. Each input X is a 48-hour ICU stay window. For outcome Y, we have three classification tasks: in-hospital mortality (`mort`), which laboratory tests are ordered in the following 24 hours (`lab`), and whether a vasopressor or antibiotic is administered in the following 24 hours (`drug`); and two regression tasks: ICU length of stay (`los`) and future charted values at observed positions (`forecast`). `mort`, `lab` and `drug` are scored by the area under the receiver operating characteristic curve (AUROC), `los` by mean squared error (MSE) on the source-standardized scale, and `forecast` by MSE over observed positions. More details for datasets, and task definitions are in Appendix D.

In contrast to ICU datasets, which contain dense records over a short continuous window, longitudinal EHRs (LEHR) contain sparser records over different visit types and a much longer time-frame, usually years. We use two LEHR datasets, EHRSHOT (Wornow et al., 2023) and INSPECT (Huang et al., 2023), to stress-test PORTER on zero-shot transfer and study the effect of input context length in Section 5.4 and Appendix G.

Representation methods. We evaluate 11 methods and two controls[1]. For methods that *do not* use LLM as summarizer, we evaluate: 1. clinical sequence encoders: CLMBR-T, GARLIC, and KEDGN; 2. general time-series and tabular models: MIRA and TabICL; 3. hand-crafted assessment scores: MEWS and SOFA; 4. direct embedding of rendered record *without summarization*: Markdown and Template, which renders the record with a fixed template and replaces variable names with their standard LOINC and ATC names. For control we use RANDOM embeddings and ATTACK, a representation that is identical for every patient and therefore ignores the record. For method that use LLM to summarize, we evaluate Record2Vec and three further summarization prompts are compared in Appendix F.1. The main results are shown with Qwen3.6 (Yang et al., 2025) as the summarizer. We also experimented with other open-source LLMs; see Appendix E.1 for full results.

[1]References are included in Table 1.

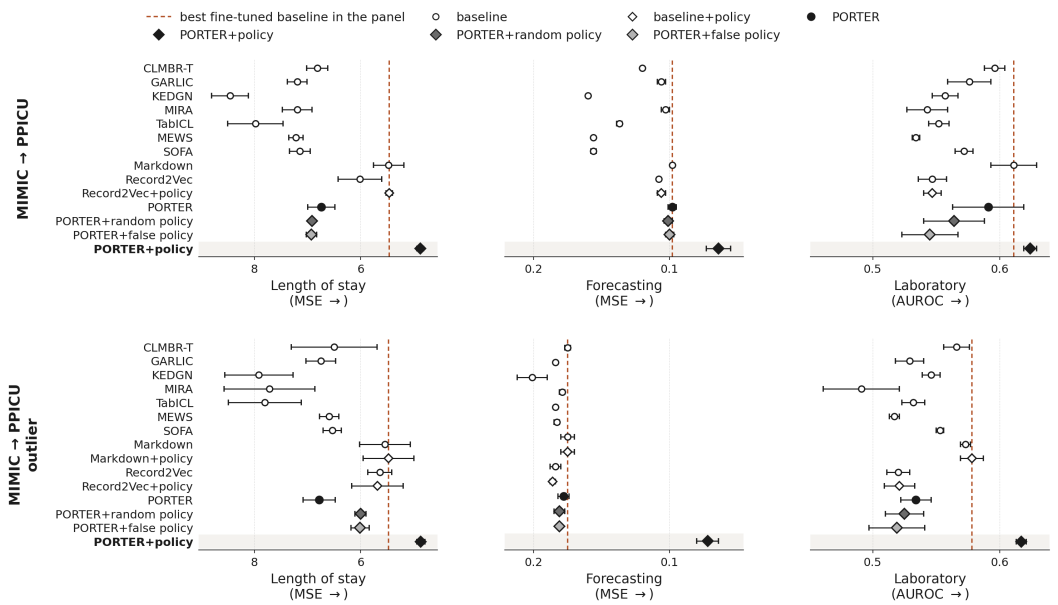

Figure 4: We evaluate the three tasks on which our methods were most challenged in the zero-shot setting. **Top:** robustness across five independently sampled sets of 16 labeled target-site records, reducing sensitivity to any particular few-shot sample. **Bottom:** a clinically motivated outlier set designed to stress-test performance on challenging cases that may arise in deployment. Circles indicate the base method and diamonds indicate its policy-augmented variant. Open markers denote baselines, whereas filled markers denote our methods, **PORTER** and **PORTER + policy**. The dashed line marks the strongest fine-tuned baseline in each panel. With only 16 labeled target-site examples, **PORTER + policy** surpasses the strongest baseline across all shown panels and remains robust under the clinical-outlier stress test. We introduce two additional attacks that replace the summarized policy with a random or false policy, validating the effect of using a correctly summarized policy.

Downstream models. We evaluate following downstream model architectures: 1. linear; 2. two-layer MLP; 3. two-layer LSTM (Hochreiter & Schmidhuber, 1997); 4. xLSTM-Mixer (Kraus et al., 2026).

Evaluation protocol. For **zero-shot transfer,** we evaluate all six ordered pairs of sites using five-fold cross-validation on the source training pool, with each fold evaluated on the held-out target test split. Because task metrics differ in scale and direction, we rank the sixteen representations within each transfer-task-head setting and summarize performance by mean rank and cumulative Top-1, Top-2, and Top-3 counts. For **few-shot adaptation,** to assess robustness, we independently sample five sets of 16 labeled target-site records, distill corresponding policy, and report performance across them. In addition, motivated by clinical deployment, we construct a set of clinically outlier cases as a stress test to evaluate robustness under outlier sampling. Lastly, for **prompt ablation,** we compare the PORTER prompt with other prompting methods and six ablation variants over the same grid and under all four summarizers. Details for LEHR evaluation protocols are shown in Appendix G.

5 Results

We evaluate PORTER along four complementary dimensions. First, we evaluate whether our approach improves zero-shot transfer over existing methods across the tasks defined above. We then study few-shot adaptation using a small amount of labeled target-site data. The purpose is to examine both the effect of target-site policy distillation and the importance of schema-guided semantic encoding as a site-independent representation. We also perform ablations on our prompt to isolate the effect and contribution of each part of the prompt. Finally we study the transferability of PORTER on longitudinal EHR. All experiments shown here use Qwen3.6 as summarizer. Detailed results, including different transfer directions and different summarizers, are in Appendix F.

5.1 ZERO-SHOT CROSS-SITE TRANSFER

PORTER with only Z_X performs well on tasks requiring general clinical state information. In Table 1, it has the best overall mean rank and places first in more columns than any other method. It is the best method on `mort` in all but one of the six transfer directions and ranks among the top three on `drug` in all six, as Table 7 shows. Summarization retains overall clinical state information while removing noise, which strengthens the signal that is useful for these tasks. Across the six directions, PORTER is less effective on tasks requiring site-dependent information. Compared to `mort`, tasks on `los` and `lab` are more dependent on site due to discharge policy and available tests. Because our schema excludes site-dependent information, this is expected. SOFA, containing only 6 features, is particularly strong on `drug` in most directions. This *reflects strong task alignment rather than representation capacity:* the cardiovascular component of SOFA directly incorporates vasopressor dose. Conversely, SOFA contains little of the charted record that `forecast` has to reproduce, which is consistent with its poor performance on that task.

Pretrained text embeddings provide a strong baseline for transferable representations. Table 1 shows that even without LLM summarization, embedding a textual rendering of the record is generally competitive with specialized EHR encoders. Consistent with prior findings (Hegselmann et al., 2026; Ji et al., 2026), pretrained embedding models map semantically related descriptions nearby, rendering direct embedding an effective alternative when summarization is cost-prohibitive.

5.2 FEW-SHOT FINE-TUNING TO THE TARGET SITE

Target-site policy distillation closes the gaps observed with schema-guided semantic encoding in the zero-shot setting. While **PORTER** remains relatively weak on `los`, `forecast`, and `lab` prediction, **PORTER + policy** consistently surpasses the strongest fine-tuned baselines on all three tasks, which were also the most challenging under zero-shot transfer. As shown in Figure 4, this pattern is consistent across five independently sampled sets of 16 labeled target-site records. We further validate the benefit of a summarized policy by showing that random or incorrect policies provide no gains and can even corrupt learned representations, as observed in lab prediction tasks. These gains further support our hypothesis that the remaining transfer gap is driven by site-specific practice, rather than missing temporal or numerical information. At the same time, **PORTER + policy** retains strong performance on tasks where it was already competitive in the zero-shot setting. Appendix F.2 reports complete results across additional tasks, prediction heads, and transfer directions.

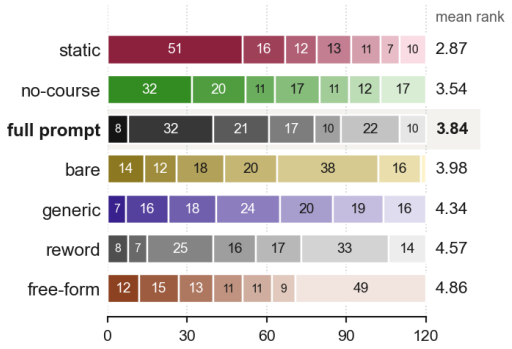

Figure 5: The PORTER prompt and its six ablation variants, ranked against one another in each of the 120 runs of the Qwen3.6 arm. Each bar counts the runs in which a variant took each rank, from rank 1 (darkest) to rank 7 (lightest); counts below seven are not printed. The right column reports mean rank.

PORTER + policy remains robust under a clinical-outlier stress test. We further evaluate all methods on a challenging set of clinically atypical target-site examples designed to stress-test few-shot adaptation under difficult deployment conditions. **PORTER + policy** maintains strong performance across all three tasks and continues to outperform the strongest fine-tuned baseline in every panel (Figure 4, bottom). These results suggest that the benefit of target-site policy distillation extends to clinically challenging cases where robust adaptation is particularly important.

Target-site policy distillation benefits from a site-independent representation on which to adapt. Applying the same target-site policy augmentation to Record2Vec and the strongest non-summarization baselines does not yield comparable improvements in mean performance (Figure 4). This contrast suggests that policy adaptation is most effective when the underlying representation has already suppressed site-specific variation. In site-dependent representations, existing source-site signals may interfere with the newly introduced target-site policy, limiting the benefit of fine-tuning.

Together, these results highlight the role of schema-guided semantic encoding as a site-independent foundation on which target-site policy can be effectively distilled.

5.3 PROMPT ABLATION

The PORTER prompt (**full** in the ablation tables) consists of five components: **A1 task framing** specifies the context and the summarization objective; **A2 overall course** asks for a concise summary of the patient's trajectory over the window; **B1 clinical-section scaffold** divides the summary into 13 sections; **B2 content guidance** specifies the relevant clinical information to summarize within each section; and **B3 temporal guidance** asks for trends.

The prompt ablations modify these components individually. **generic** modifies **B1** by replacing the 13 clinically meaningful section names with generic labels. **bare** removes the section-specific content guidance **B2** and temporal guidance **B3**, testing whether the summarizer needs explicit instructions about which variables and trends belong to each clinical section. **free-form** removes the structured section components **B1–B3**, testing the value of the schema as a whole against unstructured summarization. **no-course** removes **A2**, testing whether the separate overall-course summary adds information beyond the clinical sections. **static** modifies **A1**, **A2**, and **B3** to summarize the patient's state at the end of the window rather than over the window, testing the value of explicit temporal trajectory information. **reword** paraphrases **A1** and **A2**, providing a control for sensitivity to prompt wording. For the exact prompt and changes, see Appendix C. Figure 5 compares seven prompt variants using Gemma-4 summaries across all transfer directions, downstream tasks, and prediction heads. Detailed results are in Appendix F.3.

The clinical schema matters more than detailed extraction instructions. Restricting the comparison to prompt ablations, removing the schema entirely produces the weakest aggregate result: **free-form** has the worst mean rank and is placed last more often than any other variant. In contrast, **bare** retains the clinically named section scaffold while removing section-specific content and temporal guidance, yet remains close to **full** overall. Replacing the clinical section names with generic labels causes **generic** to fall behind **bare**. These show that the schema contributes substantial value even without extensive section-level instructions.

Relative to other representation methods, the schema-based prompts are highly competitive in aggregate. The best prompt variants achieve among the strongest mean ranks across transfer directions, tasks, and prediction heads, although weaker prompt variants overlap with strong non-prompt comparators. This advantage is clearest on `mort` and `drug`, while existing methods remain stronger on `los`. Thus, the schema appears to be a robust representation design choice, even though its best realization remains task-dependent.

Tasks prefer different, non-additive prompt biases. No prompt is uniformly best across downstream tasks. Among the ablations, **static** performs particularly well on `mort`, `los`, `lab`, and `drug`, but is less competitive on `forecast`, where prompts that retain trajectory information perform better. Moreover, **static** and **no-course** rank ahead of **full** overall despite simplifying or removing prompt components. This implies adding every component does not necessarily produce the strongest representation.

5.4 ZERO-SHOT CROSS-SITE TRANSFER ON LONGITUDINAL EHR

We next test whether these findings extend beyond the ICU to longitudinal EHR, where a record spans years of routine care. We transfer between EHRSHOT and INSPECT in both directions on five tasks, and Figure 6 shows the results.

PORTER remains effective on longitudinal EHR. PORTER achieves the best overall ranking in both transfer directions. Its advantage is the most obvious on 30-day and all records. It performs especially well on one-year hypertension and cancer onset, and next creatinine value prediction. Consistent with our findings on ICU datasets, *direct embedding* of records is a strong low-cost method, achieving a similar levels of performance on the ICU. hyperlipidemia and hypertension, although slight worse on next creatinine value.

More context generally leads to better performance on tasks that mostly depend on current patient state. On all tasks except time to admission, both direct embedding and PORTER performed

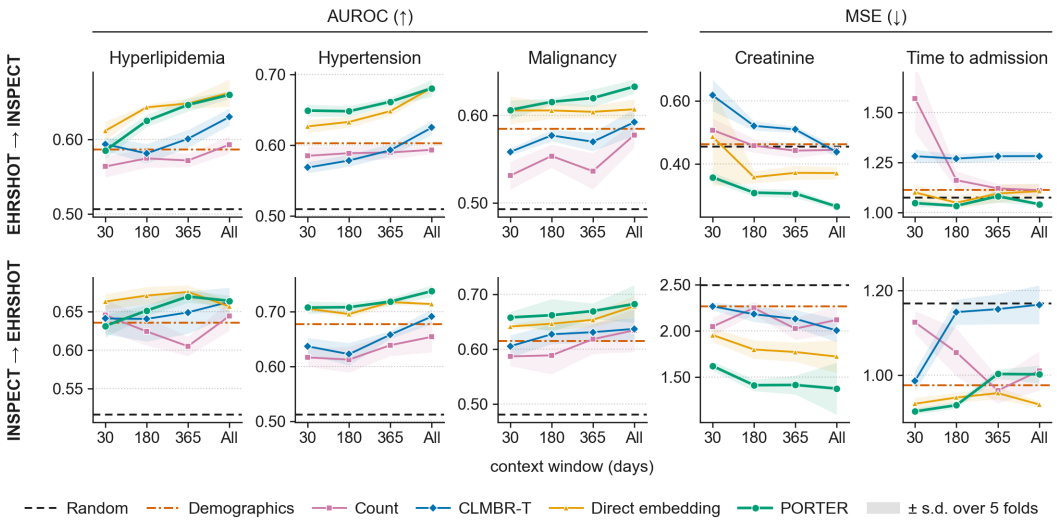

Figure 6: Zero-shot transfer between EHRSHOT and INSPECT with the MLP prediction head. Rows are the two transfer directions and columns the five tasks; each panel plots performance against the context window. Lines are the mean and bands are ± 1 standard deviation across those folds. RANDOM and Demographics do not use the context window and are drawn as horizontal lines. Detailed results are in Appendix G.

better when using all observed values for each code within the context window, compared to using only the latest. Except on time to admission, all methods perform better when using the entire history of the patient compared to using only the most recent 30 or 180 days.

6 Discussion

We proposed PORTER, a summarize-then-embed pipeline for EHR representation construction, based on an approximation of the underlying data generating process. We found that the approximation guided us towards building representations that are more transferable. PORTER achieves strong cross-site performance compared to existing methods that were not designed with the DGP in mind. This suggests that understanding the DGP of the data is important for creating transferable methods. Our comparisons further showed that relatively small open-source LLMs can produce useful summaries from raw EHRs. Our results also corroborate prior work showing that general-purpose LLMs can serve as effective EHR encoders (Hegselmann et al., 2026; Ji et al., 2026): in our cross-site experiments, directly embedding serialized EHRs produced representations that transferred better than those produced by specialized clinical encoders.

Appropriate prompting can increase the usefulness of LLM summarizers even further. Our ablations show that clinically meaningful prompt structure affects representation quality. Removing the clinical schema produces the weakest prompt variant, while retaining the clinically named section scaffold remains competitive even without detailed content and temporal guidance. This suggests that, in specialized domains such as medicine, eliciting relevant domain structure can matter more than providing extensive general writing instructions.

Limitations. First, using an LLM to summarize EHR adds significant computational cost and may introduce additional privacy and compliance requirements when applied to private patient data, reducing the immediate applicability of our method in real world settings. Second, the resulting embeddings are less interpretable than the original EHRs, making it difficult to inspect the downstream models. Third, we have not done any qualitative analysis on the summaries. Hence we cannot be certain whether they contain hallucinations and site-dependent parts, or retain meaningful predictive information.

AI USE STATEMENT

In this work, we used generative AI tools for assist in writing of proofs, assist in writing code for the experiments. We have not used generative AI tools to generate synthetic data sets, help develop theoretical models or conceptual frameworks, formulate mathematical claims, provide critical ingredients for proving mathematical claims, propose or refine hypotheses, design or provide feedback on research methodology or experiments assist with translation, clean and reformat dataset, support qualitative and thematic data analysis, interpret results. Additionally, we used generative AI tools for generating plots, summarizing and analyzing existing literature, editing writing to improve readability. We have reviewed all AI-assisted work. We manually verified proposed steps in proofs, and reviewed LLM-generated code for correctness. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

ETHICS STATEMENT

All authors completed the CITI training required for access to credentialed PhysioNet datasets and are trained in the ethical and privacy considerations involved in handling human clinical data. We followed applicable data-use requirements and privacy regulations, including HIPAA, GDPR, and local policies where relevant. All experiments involving patient data were conducted in controlled, access-restricted computing environments. Patient data were not transmitted outside these environments, and no protected health information was provided to external services. The datasets were used solely for research purposes, and no outputs from the models were used to inform clinical care.

REPRODUCIBILITY STATEMENT

All five datasets used, HiRID (Faltys et al., 2021), MIMIC-IV (Johnson et al., 2023), PPICU (Finlayson et al., 2021), EHRSHOT (Wornow et al., 2023), and INSPECT (Huang et al., 2023), are publicly accessible under data agreement. Dataset preprocessing and label construction are documented in detail in Appendix D.1.3. Exact training hyperparameters and early stopping criteria are provided in Appendix E.3. We control randomness using a fixed set of random seeds. We will release the code needed to reproduce our results after de-anonymization.

REFERENCES

Denis Agniel, Isaac S. Kohane, and Griffin M. Weber. Biases in electronic health record data due to processes within the healthcare system: retrospective observational study. *BMJ*, 361:k1479, 2018. doi: 10.1136/bmj.k1479. URL https://doi.org/10.1136/bmj.k1479.

Martin Arjovsky, Léon Bottou, Ishaan Gulrajani, and David Lopez-Paz. Invariant risk minimization. *arXiv preprint arXiv:1907.02893*, 2019.

Rahul K Arora, Jason Wei, Rebecca Soskin Hicks, Preston Bowman, Joaquin Quiñonero-Candela, Foivos Tsimpourlas, Michael Sharman, Meghan Shah, Andrea Vallone, Alex Beutel, et al. Healthbench: Evaluating large language models towards improved human health. *arXiv preprint arXiv:2505.08775*, 2025.

William Boag, Mercy Oladipo, and Peter Szolovits. Ehr safari: data is contextual. In *Machine Learning for Healthcare Conference*, pp. 391–408. PMLR, 2022.

Olivier Bodenreider. The unified medical language system (umls): Integrating biomedical terminology. *Nucleic acids research*, 32:D267–70, 02 2004. doi: 10.1093/nar/gkh061.

Adam M. Chekroud, Matt Hawrilenko, Hieronimus Loho, Julia Bondar, Ralitza Gueorguieva, Alkomiet Hasan, Joseph Kambeitz, Philip R. Corlett, Nikolaos Koutsouleris, Harlan M. Krumholz, John H. Krystal, and Martin Paulus. Illusory generalizability of clinical prediction models. *Science*, 383(6679):164–167, 2024. doi: 10.1126/science.adg8538. URL https://doi.org/10.1126/science.adg8538.

Genna R Cohen, Charles P Friedman, Andrew M Ryan, Caroline R Richardson, and Julia Adler-Milstein. Variation in physicians' electronic health record documentation and potential patient harm from that variation. *Journal of general internal medicine*, 34(11):2355–2367, 2019.

Alicia Curth, Patrick Thoral, Wilco van den Wildenberg, Peter Bijlstra, Daan de Bruin, Paul Elbers, and Mattia Fornasa. Transferring clinical prediction models across hospitals and electronic health record systems. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pp. 605–621. Springer, 2019.

Ilker Demirel, Larry Shi, Zeshan Hussain, and David Sontag. Llms can construct powerful representations and streamline sample-efficient supervised learning. *arXiv preprint arXiv:2603.11679*, 2026.

Kevin Donnelly. Snomed-ct: The advanced terminology and coding system for ehealth. *Studies in health technology and informatics*, 121:279–290, 02 2006.

Martin Faltys, Marc Zimmermann, Xinrui Lyu, Matthias Hüser, Stephanie Hyland, Gunnar Rätsch, and Tobias Merz. HiRID, a high time-resolution ICU dataset. *PhysioNet*, February 2021. doi: 10.13026/nkwc-js72. URL https://doi.org/10.13026/nkwc-js72. Version 1.1.1.

Samuel G Finlayson, Adarsh Subbaswamy, Karandeep Singh, John Bowers, Annabel Kupke, Jonathan Zittrain, Isaac S Kohane, and Suchi Saria. The clinician and dataset shift in artificial intelligence. *New England Journal of Medicine*, 385(3):283–286, 2021.

Sunyang Fu, Lester Y Leung, Anne-Olivia Raulli, David F Kallmes, Kristin A Kinsman, Kristoff B Nelson, Michael S Clark, Patrick H Luetmer, Paul R Kingsbury, David M Kent, et al. Assessment of the impact of ehr heterogeneity for clinical research through a case study of silent brain infarction. *BMC medical informatics and decision making*, 20(1):60, 2020.

Kin Wah Fung and Junchuan Xu. Synergism between the mapping projects from snomed ct to icd-10 and icd-10-cm. *AMIA ... Annual Symposium proceedings. AMIA Symposium*, 2012:218–27, 2012. URL https://api.semanticscholar.org/CorpusID:17537962.

Joseph Futoma, Morgan Simons, Trishan Panch, Finale Doshi-Velez, and Leo Anthony Celi. The myth of generalisability in clinical research and machine learning in health care. *The Lancet Digital Health*, 2(9):e489–e492, 2020. doi: 10.1016/S2589-7500(20)30186-2. URL https://doi.org/10.1016/S2589-7500(20)30186-2.

Joseph Futoma, Morgan Simons, Finale Doshi-Velez, and Rishikesan Kamaleswaran. Generalization in clinical prediction models: The blessing and curse of measurement indicator variables. *Critical Care Explorations*, 3(7):e0453, 2021. doi: 10.1097/CCE.0000000000000453. URL https://doi.org/10.1097/CCE.0000000000000453.

Lin Lawrence Guo, Stephen R Pfohl, Jason Fries, Alistair EW Johnson, Jose Posada, Catherine Aftandilian, Nigam Shah, and Lillian Sung. Evaluation of domain generalization and adaptation on improving model robustness to temporal dataset shift in clinical medicine. *Scientific reports*, 12 (1):2726, 2022.

Lin Lawrence Guo, Jason Fries, Ethan Steinberg, Scott Lanyon Fleming, Keith Morse, Catherine Aftandilian, Jose Posada, Nigam Shah, and Lillian Sung. A multi-center study on the adaptability of a shared foundation model for electronic health records. *NPJ digital medicine*, 7(1):171, 2024.

Lin Lawrence Guo, Adam Paul Yan, Emily Vettese, and Lillian Sung. Porter: Language-grounded event representations for portable structured ehr foundation models. *arXiv preprint arXiv:2606.24102*, 2026.

Hrayr Harutyunyan, Hrant Khachatrian, David C Kale, Greg Ver Steeg, and Aram Galstyan. Multitask learning and benchmarking with clinical time series data. *Scientific data*, 6(1):96, 2019.

Stefan Hegselmann, Alejandro Buendia, Hunter Lang, Monica Agrawal, Xiaoyi Jiang, and David Sontag. Tabllm: Few-shot classification of tabular data with large language models. In *International conference on artificial intelligence and statistics*, pp. 5549–5581. PMLR, 2023.

Stefan Hegselmann, Georg von Arnim, Tillmann Rheude, Noel Kronenberg, David Sontag, Gerhard Hindricks, Roland Eils, and Benjamin Wild. Large language models are powerful electronic health record encoders. *npj Digital Medicine*, 9(1):530, 2026. doi: 10.1038/s41746-026-02915-9. URL https://doi.org/10.1038/s41746-026-02915-9.

Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8): 1735–1780, 1997.

George Hripcsak, Jon D Duke, Nigam H Shah, Christian G Reich, Vojtech Huser, Martijn J Schuemie, Marc A Suchard, Rae Woong Park, Ian Chi Kei Wong, Peter R Rijnbeek, et al. Observational health data sciences and informatics (ohdsi): opportunities for observational researchers. *Studies in health technology and informatics*, 216:574, 2015.

Shih-Cheng Huang, Zepeng Huo, Ethan Steinberg, Chia-Chun Chiang, Matthew P Lungren, Curtis P Langlotz, Serena Yeung, Nigam H Shah, and Jason A Fries. Inspect: a multimodal dataset for pulmonary embolism diagnosis and prognosis. *arXiv preprint arXiv:2311.10798*, 2023.

Kyunghoon Hur, Jiyoung Lee, Jungwoo Oh, Wesley Price, Younghak Kim, and Edward Choi. Unifying heterogeneous electronic health records systems via text-based code embedding. In *Conference on Health, Inference, and Learning*, pp. 183–203. PMLR, 2022.

Kyunghoon Hur, Jungwoo Oh, Junu Kim, Jiyoun Kim, Min Jae Lee, Eunbyeol Cho, Seong-Eun Moon, Young-Hak Kim, Louis Atallah, and Edward Choi. Genhpf: General healthcare predictive framework for multi-task multi-source learning. *IEEE Journal of Biomedical and Health Informatics*, 28(1):502–513, 2023.

Kyunghoon Hur, Jungwoo Oh, Junu Kim, Jiyoun Kim, Min Jae Lee, Eunbyeol Cho, Seong-Eun Moon, Young-Hak Kim, Louis Atallah, and Edward Choi. Genhpf: General healthcare predictive framework for multi-task multi-source learning. *IEEE Journal of Biomedical and Health Informatics*, 28(1):502–513, 2024. doi: 10.1109/JBHI.2023.3327951.

Kasra Jalaldoust and Elias Bareinboim. Transportable representations for domain generalization. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 12790–12800, 2024.

Zongliang Ji, Yifei Sun, Andre Carlos Kajdacsy-Balla Amaral, Anna Goldenberg, and Rahul G Krishnan. Can we generate portable representations for clinical time series data using llms? In *The Fourteenth International Conference on Learning Representations*, 2026.

Alistair EW Johnson, Lucas Bulgarelli, Lu Shen, Alvin Gayles, Ayad Shammout, Steven Horng, Tom J Pollard, Sicheng Hao, Benjamin Moody, Brian Gow, et al. Mimic-iv, a freely accessible electronic health record dataset. *Scientific data*, 10(1):1, 2023.

Matthias Kirchler, Matteo Ferro, Veronica Lorenzini, Robin P van de Water, FinnGen Ganna Andrea 3, Christoph Lippert, and Andrea Ganna. Large language models improve transferability of electronic health record-based predictions across countries and coding systems. *npj Digital Medicine*, 2026.

Amber C Kiser, Karen Eilbeck, Jeffrey P Ferraro, David E Skarda, Matthew H Samore, and Brian Bucher. Standard vocabularies to improve machine learning model transferability with electronic health record data: retrospective cohort study using health care–associated infection. *JMIR medical informatics*, 10(8):e39057, 2022.

Zeljko Kraljevic, Thomas Searle, Anthony Shek, Lukasz Roguski, Kawsar Noor, Daniel Bean, Aurelie Mascio, Leilei Zhu, Amos A. Folarin, Angus Roberts, Rebecca Bendayan, Mark P. Richardson, Robert Stewart, Anoop D. Shah, Wai Keong Wong, Zina Ibrahim, James T. Teo, and Richard J.B. Dobson. Multi-domain clinical natural language processing with medcat: The medical concept annotation toolkit. *Artificial Intelligence in Medicine*, 117:102083, 2021. ISSN 0933-3657. doi: https://doi.org/10.1016/j.artmed.2021.102083. URL https://www.sciencedirect.com/science/article/pii/S0933365721000762.

Maurice Kraus, Felix Divo, Devendra Singh Dhami, and Kristian Kersting. xlstm-mixer: Multivariate time series forecasting by mixing via scalar memories. *Advances in Neural Information Processing Systems*, 38:6547–6581, 2026.

Thomas A Lasko, Eric V Strobl, and William W Stead. Why do probabilistic clinical models fail to transport between sites. *NPJ Digital Medicine*, 7(1):53, 2024.

Simon A Lee, Anthony Wu, and Jeffrey N Chiang. Clinical modernbert: An efficient and long context encoder for biomedical text. *arXiv preprint arXiv:2504.03964*, 2025.

Peter Leese, Adit Anand, Andrew Girvin, Amin Manna, Saaya Patel, Yun Jae Yoo, Rachel Wong, Melissa Haendel, Christopher G Chute, Tellen Bennett, et al. Clinical encounter heterogeneity and methods for resolving in networked ehr data: a study from n3c and recover programs. *Journal of the American Medical Informatics Association*, 30(6):1125–1136, 2023.

Hao Li, Bowen Deng, Chang Xu, Zhiyuan Feng, Viktor Schlegel, Yu-Hao Huang, Yizheng Sun, Jingyuan Sun, Kailai Yang, Yiyao Yu, et al. Mira: Medical time series foundation model for real-world health data. *Advances in neural information processing systems*, 38:98657–98685, 2026.

Yicheng Luo, Zhen Liu, Linghao Wang, Binquan Wu, Junhao Zheng, and Qianli Ma. Knowledge-empowered dynamic graph network for irregularly sampled medical time series. *Advances in Neural Information Processing Systems*, 37:67172–67199, 2024.

Cristina Menghini, Peter Ney, Hamza Kwisaba, Miles Turpin, Felix Binder, Jean-Christophe Testud, Aidan Boyd, Nathaniel Li, Ivan Evtimov, Klaudia Krawiecka, et al. Muse spark safety & preparedness report. *arXiv preprint arXiv:2606.12429*, 2026.

Jack Needleman, Peter Buerhaus, V. Shane Pankratz, Cynthia L. Leibson, Susanna R. Stevens, and Marcelline Harris. Nurse staffing and inpatient hospital mortality. *New England Journal of Medicine*, 364(11):1037–1045, 2011. doi: 10.1056/NEJMsa1001025. URL https://doi.org/10.1056/NEJMsa1001025.

Bret Nestor, Matthew BA McDermott, Willie Boag, Gabriela Berner, Tristan Naumann, Michael C Hughes, Anna Goldenberg, and Marzyeh Ghassemi. Feature robustness in non-stationary health records: caveats to deployable model performance in common clinical machine learning tasks. In *Machine Learning for Healthcare Conference*, pp. 381–405. PMLR, 2019.

Chao Pang, Xinzhuo Jiang, Krishna S Kalluri, Matthew Spotnitz, RuiJun Chen, Adler Perotte, and Karthik Natarajan. Cehr-bert: Incorporating temporal information from structured ehr data to improve prediction tasks. In *Machine learning for health*, pp. 239–260. PMLR, 2021.

Judea Pearl and Elias Bareinboim. Transportability of causal and statistical relations: A formal approach. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 25, pp. 247–254, 2011.

Rimma Pivovarov and Noémie Elhadad. Automated methods for the summarization of electronic health records. *Journal of the American Medical Informatics Association*, 22(5):938–947, 2015. doi: 10.1093/jamia/ocv032. URL https://doi.org/10.1093/jamia/ocv032.

Jingang Qu, David Holzmüller, Gaël Varoquaux, and Marine Le Morvan. Tabicl: A tabular foundation model for in-context learning on large data. In *International Conference on Machine Learning*, pp. 50817–50847. PMLR, 2025.

Alvin Rajkomar, Eyal Oren, Kai Chen, Andrew M. Dai, Nissan Hajaj, Michaela Hardt, Peter J. Liu, Xiaobing Liu, Jake Marcus, Mimi Sun, Patrik Sundberg, Hector Yee, Kun Zhang, Yi Zhang, Gerardo Flores, Gavin E. Duggan, Jamie Irvine, Quoc Le, Kurt Litsch, Alexander Mossin, Justin Tansuwan, De Wang, James Wexler, Jimbo Wilson, Dana Ludwig, Samuel L. Volchenboum, Katherine Chou, Michael Pearson, Srinivasan Madabushi, Nigam H. Shah, Atul J. Butte, Michael D. Howell, Claire Cui, Greg S. Corrado, and Jeffrey Dean. Scalable and accurate deep learning with electronic health records. *npj Digital Medicine*, 1(1):18, 2018. doi: 10.1038/s41746-018-0029-1. URL https://doi.org/10.1038/s41746-018-0029-1.

Pranav Rajpurkar, Emma Chen, Oishi Banerjee, and Eric J Topol. Ai in health and medicine. *Nature medicine*, 28(1):31–38, 2022.

Juan C Rojas, Patrick G Lyons, Kaveri Chhikara, Vaishvik Chaudhari, Sivasubramanium V Bhavani, Muna Nour, Kevin G Buell, Kevin D Smith, Catherine A Gao, Saki Amagai, et al. A common longitudinal intensive care unit data format (clif) for critical illness research. *Intensive care medicine*, 51(3):556–569, 2025.

Mateo Rojas-Carulla, Bernhard Schölkopf, Richard Turner, and Jonas Peters. Invariant models for causal transfer learning. *Journal of Machine Learning Research*, 19(36):1–34, 2018.

Matthew A. Rysavy, Lei Li, Edward F. Bell, Abhik Das, Susan R. Hintz, Barbara J. Stoll, Betty R. Vohr, Waldemar A. Carlo, Seetha Shankaran, Michele C. Walsh, Jon E. Tyson, C. Michael Cotten, P. Brian Smith, Jeffrey C. Murray, Tarah T. Colaizy, Jane E. Brumbaugh, and Rosemary D. Higgins. Between-hospital variation in treatment and outcomes in extremely preterm infants. *New England Journal of Medicine*, 372(19):1801–1811, 2015. doi: 10.1056/NEJMoa1410689. URL https://doi.org/10.1056/NEJMoa1410689.

Andrew Sellergren, Sahar Kazemzadeh, Tiam Jaroensri, Atilla Kiraly, Madeleine Traverse, Timo Kohlberger, Shawn Xu, Fayaz Jamil, Cían Hughes, Charles Lau, et al. Medgemma technical report. *arXiv preprint arXiv:2507.05201*, 2025.

Benjamin Shickel, Patrick James Tighe, Azra Bihorac, and Parisa Rashidi. Deep EHR: A survey of recent advances in deep learning techniques for electronic health record (EHR) analysis. *IEEE Journal of Biomedical and Health Informatics*, 22(5):1589–1604, 2018. doi: 10.1109/JBHI.2017.2767063. URL https://doi.org/10.1109/JBHI.2017.2767063.

Rose Sisk, Lijing Lin, Matthew Sperrin, Jessica K. Barrett, Brian Tom, Karla Diaz-Ordaz, Niels Peek, and Glen P. Martin. Informative presence and observation in routine health data: A review of methodology for clinical risk prediction. *Journal of the American Medical Informatics Association*, 28(1):155–166, 2021. doi: 10.1093/jamia/ocaa242. URL https://doi.org/10.1093/jamia/ocaa242.

Ethan Steinberg, Ken Jung, Jason A Fries, Conor K Corbin, Stephen R Pfohl, and Nigam H Shah. Language models are an effective representation learning technique for electronic health record data. *Journal of biomedical informatics*, 113:103637, 2021.

Adarsh Subbaswamy and Suchi Saria. From development to deployment: dataset shift, causality, and shift-stable models in health ai. *Biostatistics*, 21(2):345–352, 2020.

Christian P Subbe, Michael Kruger, Peter Rutherford, and L Gemmel. Validation of a modified early warning score in medical admissions. *Qjm*, 94(10):521–526, 2001.

Gemma Team, Sherif El Abd, Vaibhav Aggarwal, Robin Algayres, Alek Andreev, Olivier Bachem, Ian Ballantyne, Cormac Brick, Victor Cărbune, Michelle Casbon, et al. Gemma 4 technical report. *arXiv preprint arXiv:2607.02770*, 2026.

Tao Tu, Mike Schaekermann, Anil Palepu, Khaled Saab, Jan Freyberg, Ryutaro Tanno, Amy Wang, Brenna Li, Mohamed Amin, Yong Cheng, et al. Towards conversational diagnostic artificial intelligence. *Nature*, 642(8067):442–450, 2025.

Robin Van De Water, Hendrik Schmidt, Paul Elbers, Patrick Thoral, Bert Arnrich, and Patrick Rockenschaub. Yet another icu benchmark: A flexible multi-center framework for clinical ml. In *International Conference on Learning Representations*, volume 2024, pp. 44638–44693, 2024.

Dave Van Veen, Cara Van Uden, Louis Blankemeier, Jean-Benoit Delbrouck, Asad Aali, Christian Bluethgen, Anuj Pareek, Malgorzata Polacin, Eduardo Pontes Reis, Anna Seehofnerová, et al. Adapted large language models can outperform medical experts in clinical text summarization. *Nature medicine*, 30(4):1134–1142, 2024.

J-L Vincent, Rui Moreno, Jukka Takala, Sheila Willatts, Arnaldo De Mendonça, Hajo Bruining, C Kathryn Reinhart, PeterM Suter, and Lambertius G Thijs. The sofa (sepsis-related organ failure assessment) score to describe organ dysfunction/failure: On behalf of the working group on sepsis-related problems of the european society of intensive care medicine (see contributors to the project in the appendix). *Intensive care medicine*, 22(7):707–710, 1996.

Mengqian Wang, Manhua Wang, Fei Yu, Yue Yang, Jennifer Walker, and Javed Mostafa. A systematic review of automatic text summarization for biomedical literature and EHRs. *Journal of the American Medical Informatics Association*, 28(10):2287–2297, 2021. doi: 10.1093/jamia/ocab143. URL https://doi.org/10.1093/jamia/ocab143.

Ruirui Wang, Yanke Li, Manuel Günther, and Diego Paez-Granados. Garlic: Graph attention-based relational learning of multivariate time series in intensive care. In *International Conference on Learning Representations*, 2026.

Michael Wornow, Rahul Thapa, Ethan Steinberg, Jason Fries, and Nigam Shah. Ehrshot: An ehr benchmark for few-shot evaluation of foundation models. *Advances in Neural Information Processing Systems*, 36:67125–67137, 2023.

Cao Xiao, Edward Choi, and Jimeng Sun. Opportunities and challenges in developing deep learning models using electronic health records data: a systematic review. *Journal of the American Medical Informatics Association*, 25(10):1419–1428, 2018. doi: 10.1093/jamia/ocy068. URL https://doi.org/10.1093/jamia/ocy068.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025.

Haoran Zhang, Natalie Dullerud, Laleh Seyyed-Kalantari, Quaid Morris, Shalmali Joshi, and Marzyeh Ghassemi. An empirical framework for domain generalization in clinical settings. In *Proceedings of the conference on health, inference, and learning*, pp. 279–290, 2021.

Yanzhao Zhang, Mingxin Li, Dingkun Long, Xin Zhang, Huan Lin, et al. Qwen3 embedding: Advancing text embedding and reranking through foundation models. *arXiv preprint arXiv:2506.05176*, 2025. URL https://arxiv.org/abs/2506.05176.

A Related work

Site-dependent distribution-shift in clinical prediction. Clinical prediction models often degrade across sites (Futoma et al., 2020). One reason is that EHRs reflect both patient state and the local processes through which care is delivered and documented (Agniel et al., 2018; Boag et al., 2022). Differences in case mix, measurement, documentation, and treatment practices can change both the observed record and its relationship with downstream outcomes (Finlayson et al., 2021; Subbaswamy & Saria, 2020). Even under nominally standardized ICU prediction tasks, choices of dataset, cohort definition, and preprocessing can affect performance as much as or more than model class (Van De Water et al., 2024). Predictive features may consequently become unstable as healthcare processes change across sites or over time (Nestor et al., 2019; Futoma et al., 2021). A causal account of this problem distinguishes the underlying disease process from the site-dependent process through which clinical variables become observed (Lasko et al., 2024). PORTER adopts this distinction while separately allowing the mechanisms governing observation, treatment, and outcomes to vary across sites.

Transportability and invariant prediction. Causal transportability formalizes when relationships learned in one environment can be transferred to another under assumptions about which data-generating mechanisms differ (Pearl & Bareinboim, 2011). Related approaches identify predictive relationships that remain stable across environments. Rojas-Carulla et al. learn from conditional distributions assumed to be invariant under environment changes (Rojas-Carulla et al., 2018), while Invariant Risk Minimization seeks representations for which the same predictor is optimal across training environments (Arjovsky et al., 2019). In clinical prediction, however, generic domain-generalization and unsupervised-adaptation methods have not consistently improved robustness to temporal shift over empirical risk minimization (Guo et al., 2022). Transportability has also been studied directly at the representation level, where a representation may permit transport without itself being invariant across domains (Jalaldoust & Bareinboim, 2024). Complementary clinical work adapts models to target hospitals using local labeled data or continued pretraining (Curth et al., 2019; Guo et al., 2024). This distinction is central to PORTER: genuine differences in case-mix and treatments may remain represented and correlated with site. The causal model instead specifies which site-dependent variation should be normalized and which information must be retained for prediction, before any optional target-site adaptation of the downstream model.

Semantic harmonization of EHRs. Institutions differ in database structure, coding systems, encounter construction, and documentation practices (Fu et al., 2020; Leese et al., 2023; Cohen et al., 2019). Conventional harmonization approaches such as the OMOP Common Data Model standardize schemas and map local vocabularies to shared clinical concepts (Hripcsak et al., 2015). Empirically, grouping site-specific features through standard clinical vocabularies can reduce performance degradation under external validation (Kiser et al., 2022). These mappings address syntactic and vocabulary-level differences, but do not directly normalize what is measured, how frequently it is measured, or how equivalent clinical states are documented. Language-based approaches instead use textual semantics as a shared interface. Language-based approaches such as DescEmb and GenHPF use clinical descriptions of codes, fields, and values to bridge heterogeneous vocabularies and database schemas (Hur et al., 2022; 2023), while structured EHR encoders such as CEHR-BERT explicitly incorporate temporal information (Pang et al., 2021). More recent methods serialize structured EHRs for encoding by general-purpose language models or map medical concepts into shared language-model embedding spaces (Hegselmann et al., 2026; Kirchler et al., 2026). Natural-language grounding at the level of individual EHR events can similarly improve portability across vocabularies and sites (Guo et al., 2026). PORTER differs by canonicalizing an entire observation window into a fixed clinical schema, targeting variation in how clinical state is realized rather than only variation in vocabulary or event identity.

Summarize-then-embed patient representations. Clinical summarization transforms complex longitudinal records into compact descriptions of clinically salient information (Pivovarov & Elhadad, 2015; Wang et al., 2021), and adapted LLMs have demonstrated strong performance on clinician-facing summarization tasks (Van Veen et al., 2024). More broadly, self-supervised language modeling over coded EHR sequences can produce reusable patient representations for sample-efficient downstream prediction (Steinberg et al., 2021). Recent work also uses natural language itself as the

intermediate representation. Lee et al. convert separate structured EHR domains into template-based pseudo-notes and embed them with frozen language models, enabling adaptation to an unstandardized external EHR using limited target-site data (Lee et al., 2025). Demirel et al. construct structured LLM-derived representations for sample-efficient supervised learning (Demirel et al., 2026). Ji et al. use a frozen LLM to summarize irregular ICU time series and a frozen text encoder to map the summaries into fixed-dimensional representations, improving transfer across hospitals (Ji et al., 2026). PORTER builds on this summarize-then-embed paradigm but places additional structure on the intermediate text. Unlike template-based serialization, it uses an LLM to canonicalize the observation window according to a fixed clinical schema. The schema is designed to preserve case-mix information, outcome-relevant treatments and support, and predictive clinical-state information while encouraging semantically equivalent observations to be expressed consistently across sites. The representation is therefore motivated by explicit transport and predictive-sufficiency requirements rather than by summarization quality alone. Nevertheless, compressing an observation window into text can lose fine-grained numerical values and short-term temporal dynamics, particularly for tasks requiring precise forecasting.

B PROOF OF TRANSFERABILITY

Let S denote the site, C the latent clinical state, T the treatments, I_Y the site-specific outcome policies, and Y the patient outcome. Let $Z = (Z_X, Z_Y)$ denote the learned representation. Under the assumed data-generating process in Figure 2,

$$Y \perp (S, Z) \mid (C, T, I_Y). \tag{1}$$

Proposition 1 (Sufficient conditions for transferability)**.** *Assume that (C, T, I_Y) has common support across sites. If Z satisfies invariance,*

$$Z \perp S \mid (C, T, I_Y), \tag{2}$$

and sufficiency,

$$Y \perp (C, T, I_Y) \mid (Z, S), \tag{3}$$

then $Y \perp S \mid Z$.

Proof. Fix z and s with $p(z, s) > 0$, and choose any (c, t, i_Y) such that $p(c, t, i_Y, z \mid s) > 0$. By sufficiency and equation 1,

$$p(y \mid z, s) = p(y \mid c, t, i_Y, z, s) = p(y \mid c, t, i_Y). \tag{4}$$

By invariance,

$$p(z \mid c, t, i_Y, s) = p(z \mid c, t, i_Y). \tag{5}$$

Together with common support of (C, T, I_Y) across sites, this implies that for fixed z, the set of (c, t, i_Y) satisfying $p(c, t, i_Y, z \mid s) > 0$ is the same for every site at which z occurs. Therefore, by equation 4, $p(y \mid z, s)$ is the same for every such site. Denote this common value by $q(y \mid z)$.

Marginalizing over sites,

$$p(y \mid z) = \sum_s p(y \mid z, s) p(s \mid z) = \sum_s q(y \mid z) p(s \mid z) = q(y \mid z). \tag{6}$$

Hence $p(y \mid z, s) = p(y \mid z)$ for all (z, s) in the support, and therefore

$$Y \perp S \mid Z. \tag{7}$$

□

C PROMPT DESIGN FOR PORTER

We describe the clinical schema Q used by PORTER and how its parts relate to the two conditions in Section 2.

Without detailed instructions, an LLM can summarize the same record in many different ways, and the form of the summary may vary between patients and between sites. The schema asks for

the same components in the same order for every patient, so that differences between summaries reflect differences between patients rather than differences in how the record was written. We use this fixed structure to encourage invariance. The components cover the main aspects of the clinical state together with the treatments and supports the patient receives, and are intended to preserve the information required by clinical sufficiency.

The schema contains thirteen components, selected as described in Subsection 3.1. Components 1–10 describe organ systems and the general condition of the patient, and component 11 lists the active diagnoses. Together these components describe the clinical state C. Components 12 and 13 describe the treatments T. Component 12 records device and support status, and component 13 records the response to treatment.

The prompt consists of five parts, and each ablation in Appendix F.3 modifies one or more of them. Task framing (A1) asks for a concise and clinically faithful summary of the 48-hour record and restricts it to information supported by the record. The overall course (A2) asks for a one-line description of the trajectory over the 48 hours before the component details, similar to the opening of a clinical note. The section scaffold (B1) consists of the thirteen components and their names. Content guidance (B2) lists the variables to consider under each component. Temporal guidance (B3) asks whether a component is improving, worsening, or stable, and is included for eight of the thirteen components (1–6, 10, and 13). The scaffold is designed to encode the two conditions, while the other parts guide what each component emphasizes and how the summary is written. The full prompt is shown below.

	A1	**A2**	**B1**	**B2**	**B3**
Ablation	Task framing	Overall summary	Section scaffold	Content guidance	Temporal guidance
no-course	–	Delete	–	–	–
static	Modify	Modify	–	–	Delete
bare	–	–	–	Delete	Delete
free-form	–	–	Delete	Delete	Delete
generic	–	–	Rename	–	–
reword	Paraphrase	Paraphrase	–	–	–

Table 2: Prompt ablations expressed as transformations of the components of the original prompt. A dash indicates that the component is retained without modification. For **static**, A1 and A2 are replaced with 48h-window-end state framing, and B3 is removed. For **generic**, the 13 clinically meaningful section names in B1 are replaced by the generic `System 1` through `System 13`; the section heading and numbering are retained.

Here is the PORTER prompt and its correspondence to the ablation components.

A1 Task framing **A2** Overall summary **B1** Section scaffold **B2** Content guidance **B3** Temporal guidance

You are given a patient's ICU record for the past 48 hours as a chronological sequence of hourly text events describing clinically relevant observations. Write a concise, clinically faithful summary of the patient's clinical state over this 48-hour period. Include the following when supported by the record:
[Overall clinical course]
Summarize the overall 48-hour trajectory in the most concise form, stating whether the patient was improving, worsening, mixed/fluctuating, or largely stable, and the main issues driving the course.
[Clinical state]
1. Hemodynamic / cardiovascular state: blood pressure/MAP stability, heart rate/rhythm pattern, perfusion abnormalities, shock features if present, and whether instability was absent, present, improving, worsening, or persistent.
2. Respiratory state: oxygen requirement, ventilation/airway status, oxygenation pattern, respiratory distress/work of breathing if described, and whether respiratory status was improving, worsening, or unchanged.
3. Renal / fluid / elimination state: urine output, creatinine or kidney-related trends, fluid balance, edema/volume depletion if clinically relevant, dialysis dependence if part of the state, and whether renal/fluid status was stable, worsening, or improving.
4. Neurologic / mental status: consciousness, sedation, agitation, delirium, mental status, focal neurologic findings, GCS/RASS if present, and major neurologic changes across time.
5. Infectious / inflammatory / immune state: fever pattern, WBC/inflammatory markers, infection-related signals, immune-status-related findings if clinically relevant, and whether the picture was improving, worsening, persistent, or unclear.
6. Metabolic / endocrine / laboratory trajectory: major metabolic, endocrine, temperature-related, or laboratory abnormalities that materially affected the patient's state, emphasizing trends rather than isolated values.
7. Hematologic / bleeding state: clinically meaningful bleeding, hematologic abnormalities, coagulation-related issues, or blood/lymphatic/immune-system findings that affected the patient's state.
8. Gastrointestinal / nutritional state: digestive findings, feeding/nutrition status, stool/elimination abnormalities, jaundice/hepatic findings, or swallowing/eating issues if clinically active.
9. Skin / wound / tissue integrity state: wounds, erythema, skin/mucosal findings, impaired tissue integrity, stoma-related issues, or other clinically meaningful integumentary findings.
10. Functional / general clinical state: functional status, mobility/movement limitations, sleep/exertion-related state if relevant, overall patient condition/appearance, and whether general condition was improving or deteriorating.
11. Major diagnoses or syndromic impressions: summarize only the active problems clearly supported by the record, using conservative language when evidence is incomplete.
12. Device / support status: include only clinically meaningful device or support dependence that reflects the patient's current internal state, such as ventilatory support, dialysis dependence, or dependence on enabling devices.
13. Treatment / procedure response: include only therapies or procedures insofar as they changed the patient's condition, such as response to treatment, treatment-related worsening/improvement, or persistent dependence created by treatment.

C.1 TARGET-SITE POLICY DISTILLATION PROMPT

Target-site policy distillation uses a task and site conditioned prompt rather than the patient-summary prompt above. We instantiate one prompt for each of the three sites and five downstream tasks, giving 15 prompts in total. Each prompt asks the LLM to produce a compact Markdown description of the empirical site policy exhibited by 16 labelled examples. Here, "site policy" refers to observable documentation, measurement, intervention, missingness, and target-observation patterns; it does not denote an official administrative policy. The prompt prohibits causal claims, patient identifiers, sensitive attributes, and unsupported cohort-level statistics, and requires qualitative statements to be attributed to the selected examples.

The invariant prompt structure is shown below. Braces denote fields filled separately for each site and task; the block beginning with `Representative example` is repeated 16 times.

```
You are analyzing site-specific ICU documentation and {task-relevant
pattern type} for transfer learning.
Create a compact Markdown summary of the empirical site policy
represented by the 16 examples below.  Here, ''site policy'' means
{task-specific observable patterns}---not an official administrative
policy.
Dataset:  {site}
Task:  {task description}
Examples are drawn only from training and validation records.
Task definition:
{target definition, label semantics, and target names or feature
list}
Your summary must:
{task-specific content requirements}
exclude sensitive attributes and patient-level identifiers;
use only {site- and task-specific output headings}.
Leakage control:  use only the supplied example records and labels.
Do not report or infer cohort-level aggregate or distribution
statistics; describe site patterns qualitatively and attribute them
to the selected examples.
### Representative example {i}
- Source row ID: {row ID}
- {task-specific label description and value}
- SBF-short record:
{serialized 48-hour record}
```

The task-conditioned fields make the target semantics explicit. Mortality supplies the binary benchmark label; length of stay supplies the observed stay duration; laboratory and drug prediction supply the future-observation target vector together with human-readable positive targets; and forecasting supplies the observed values and mask for the $48 \times F$ input-window matrix. The forecasting prompt explicitly notes that the target does not use a shifted future horizon. For every task, each example contains its SBF-short serialization and label information, allowing the LLM to contrast observed record patterns with target values. The generated policy summary is embedded and concatenated with each patient representation for the corresponding site–task adaptation run, as described in Section 3.

D DATASETS AND TASKS

D.1 ICU

D.1.1 FEATURE INFORMATION

Features shared across datasets. The processed ICU time series share 47 features across MIMIC, HiRID, and PPICU: `ALP`, `ALT`, `Bicarbonate`, `Bilirubin`, `BloodUreaNitrogen`, `Calcium`, `Creatinine`, `Hemoglobin`, `INR`, `Lactate`, `Magnesium`, `PaCO2`, `PaO2`, `Phosphate`, `Platelets`, `Potassium`, `Sodium`, `Troponin`, `WBC`, `ph`, `AirwayPressure`, `DiastolicBloodPressure`, `FiO2`, `GCS`, `HeartRate`, `MeanBloodPressure`, `PEEP`, `RespiratoryRate`, `SAS`, `SystolicBloodPressure`, `Temperature`, `TidalVolume`,

UrineOutput, Analgesia, Antiarrhythmics, Antibiotics, Anticoagulants, Antiepileptics, Antihypertensives, Dialysis, Diuretics, ICPMonitor, MgReplacement, Paralysis, TPN, Transfusions, and Vasopressors.

MIMIC (80,749 samples). The processed ICU time series comprises 60 features, including the 47 shared features above and 13 additional features: CreatinineKinase, ICDSC, MinuteVentilation, Antipsychotics, CTScan, CaReplacement, EnteralNutrition, KReplacement, MRI, PPI, UltraSound, Ventilation, and Xray.

HiRID (36,019 samples). The processed ICU time series comprises 64 features, including the 47 shared features above and 17 additional features: AST, Chloride, CreatineKinase, Glucose, AirwayPressurePeak, FluidBalance, Saturation, Aspirin, CaReplacement, Insulin, KReplacement, LiverToxicDrug, Neuroleptics, Saline, Sedation, Steroids, and Ventilation.

PPICU (47,119 samples). The processed ICU time series comprises 75 features, including the 47 shared features above and 28 additional features: AST, Chloride, CreatinineKinase, GGT, Glucose, AirwayPressureIP, AirwayPressurePeak, FluidInput, FluidOutput, ICDSC, MinuteVentilation, PAVSupport, PC, PlateauPressure, Saturation, Antiarrhythmic, Aspirin, EKG, EVD, InhaledVasodilator, Insulin, Isoproterenol, Neuroleptics, OsmoticTherapy, PPI, PS, Sedation, and Steroids.

D.1.2 Data Preprocessing Details

We provide reproducible preprocessing pipelines for all ICU datasets considered in this work. This appendix summarizes the steps at a high level, without referencing specific files or directories.

Common setup. We organize raw and processed data under a central data directory and obtain each dataset under its respective data use agreement. Across datasets, we construct per-stay clinical time series from raw measurement and medication records, using a harmonized set of clinically meaningful features selected in consultation with experienced practicing ICU clinicians. Feature names are mapped to the shared study schema, and demographic and outcome information is merged to derive downstream targets. Unless stated otherwise, sequences are discretized at hourly resolution, and continuous features are normalized using statistics computed on the training split. The resulting time series, feature indices, labels, and dataset splits are serialized for downstream modeling.

HiRID. HiRID associates each patient with a single ICU stay, allowing records to be grouped directly by patient identifier. We retain a sparse, event-centric representation of each stay and augment the feature dictionary with event start/end times and summary statistics used for quality control and normalization. The processed representation stores time indices, value indices, and per-event counts.

MIMIC-IV. For MIMIC-IV, we first convert the raw relational tables into a columnar layout for efficient I/O and restrict the cohort to ICU stays exceeding the minimum duration threshold. Measurement identifiers are mapped to the study feature set through a separate mapping layer, allowing the mapping to be revised without repeating earlier preprocessing stages. We merge administrative and ICU stay tables to construct the cohort and exclude pediatric subjects. The processed data are organized in a subject-centric layout containing diagnoses, events, procedures, and stays, with numeric event values preferred when available and individual features stored separately for efficient access and validation.

PPICU. For PPICU, heterogeneous raw sources and multi-signal streams are first converted into a uniform tabular representation and separated into per-feature tables. We compute per-feature summary statistics for clinician review and apply clinician-verified range filters to remove implausible values. In addition to the sparse representation used for the other datasets, each stay is converted into a dense, zero-filled discretized representation. PPICU preprocessing also explicitly materializes a global time index keyed by stay identifier.

Table 3: The five tasks, in the order the results tables use. L is the number of laboratory labels and F the number of charted variables; both differ by site.

Task	Type	Label	Metric
Mortality (`mort`)	binary	benchmark mortality label	AUROC ↑
Length of stay (`los`)	scalar regression	remaining ICU stay after the input window	MSE ↓
Forecasting (`forecast`)	matrix regression	the $48 \times F$ record of the input window	MSE ↓
Laboratory (`lab`)	multi-label, L	each test measured in the following 24 hours	AUROC ↑
Drug (`drug`)	multi-label, 2	antibiotic, vasopressor	AUROC ↑

D.1.3 TASKS DEFINITION AND METRICS

Each record is one 48-hour window of an ICU stay, hours 0 to 47 of that window. Three tasks are classification and two are regression; Table 3 summarizes them and the paragraphs below give the details.

Mortality. The target is the binary mortality label that each dataset provides for a stay. All windows of a stay share this label. It is scored by AUROC.

Length of stay. The target is the remaining time in the ICU, in hours, from the end of the input window to ICU discharge. The target is standardized with the mean and standard deviation of the *source* site, and the target site uses the same constants. We report the mean squared error on this standardized scale.

Forecasting. The task is multivariate forecasting of all charted variables over the 48-hour input window. The model outputs 48 steps, and each step is the prediction for one hour of the window. The target is the corresponding $48 \times F$ matrix, and the mean squared error is taken only over the positions the record actually observes. F is 60 at MIMIC, 64 at HIRID and 75 at PPICU.

Laboratory. One binary label per laboratory test, recording whether that test is measured at least once in the 24-hour window that follows the input window, hours 48 to 71 of that window. The label records only that the test appears in the record of that window, not the value returned or whether it is abnormal.

Drug. Two binary labels over the 24-hour window that follows the input window: whether an antibiotic is recorded in it, and whether a vasopressor is recorded in it. A label is positive if the class appears at least once anywhere in that window, so the task asks whether the class is given at all rather than how much, or when, or whether it was already running before the window began. Antibiotics are the more common of the two at every site. Both labels are scored by AUROC.

Aligning label sets across sites. The sites record different laboratory tests and charted variables, and we deliberately leave them unharmonized: each site keeps its own variable set, as it would in deployment, and no name-based matching is applied. In a transfer run, the model output and target are restricted to the first $L = \min(L_{\text{source}}, L_{\text{target}})$ laboratory columns and, for forecasting, the first $F = \min(F_{\text{source}}, F_{\text{target}})$ variable columns, and compared position by position. Because the sites order their variables differently, the same position need not hold the same variable at both sites.

D.2 LONGITUDINAL EHR

D.2.1 FEATURE INFORMATION

Record content. Unlike the ICU datasets, EHRSHOT and INSPECT are not reduced to a fixed set of features. A record keeps every coded event of the patient. Each event has a code and a time, and it

may also carry a numeric value, a text value and a unit. The codes come from the standard OMOP vocabularies, such as SNOMED CT for diagnoses and LOINC for laboratory tests.

Code descriptions. The rendered record writes each code as a short description. The description comes from the code metadata of the datasets when it is available, and EHRSHOT takes precedence when both datasets describe a code. Otherwise the description is the concept name in the Athena vocabulary. A code with neither description is written as the code text after its vocabulary prefix, with underscores replaced by spaces. Sex and race codes are written as fixed words.

Demographics. Sex, race and ethnicity codes are not placed on the timeline of the rendered record. Together with the age of the patient at the prediction time, they form a header that is the same at every context window.

D.2.2 Data Preprocessing Details

Common setup. Both cohorts are in the OMOP common data model and are converted to the MEDS format. The labels and the count features each read an ontology file. The diagnosis and creatinine tasks use it to include descendant codes, and the count features use it to add ancestor codes. Each patient has one prediction time, defined in Appendix D.2.3. Every representation is computed once at that time and shared by all tasks.

Decontamination. Some patients appear in both EHRSHOT and INSPECT under the same patient ID. We remove these patients from INSPECT and leave EHRSHOT unchanged. The removal comes before any label or feature is generated, so no patient is in both the source and the target cohort. The training code also checks every run and stops if the source and target cohorts of a task share a patient.

Rendered record. The text-based representations read a rendering of the record. Each event that carries a value is written as its code description and its values, and the unit is added when a value is numeric. Events are grouped by day in time order, and each day starts with the age of the patient on that day. Values of the same code and unit on the same day share one line. Events without a value and clinical notes are left out. The demographics header comes before the events. The latest-value rendering in Appendix G keeps only the latest event of each code, in time order and without the day headings. A code is left out of it when its latest event has no value.

Context windows. Every context window ends at the prediction time and includes it. The text-based representations apply the window to the clinical events of the rendered record, and the demographics header is kept at every window. Count features and CLMBR-T apply the window to every coded event, including events without a value and demographic codes. The age column of the count features is kept at every window. Each model also has its own input limit, which applies at every window and drops the oldest events first. CLMBR-T reads at most $5{,}952$ coded events. Direct embedding reads at most $32{,}700$ tokens and the summarizer of PORTER at most $120{,}000$ tokens, and both counts include the instruction and the header. The text-based representations drop whole events and never cut an event in half.

D.2.3 Tasks definition and metrics

Each record is the history of one patient up to and including one prediction time. We construct three binary tasks predicting a new diagnosis and two regression tasks predicting a future creatinine value and time to inpatient admission. Table 4 summarizes them and the paragraphs below give the details.

Prediction time. Each patient has one prediction time, and all tasks use it. We take the days of the record between 365 days after the first clinical observation and 365 days before the last one, and we sample one of these days uniformly with a seed fixed for each patient. The prediction time is the latest time on that day within this range. A patient with no such day is left out of all tasks, so every remaining patient has at least two years between the first and the last clinical observation. To study context length, we construct representations from the 30, 180 and 365 days before the prediction time, as well as from all earlier records.

Table 4: The five LEHR tasks, in the order the results tables use. Windows are in days after the prediction time.

Task	Type	Label	Metric
New hyperlipidemia (Hyperlip.)	binary	first diagnosis in $(60, 365]$	AUROC ↑
New hypertension (Hypert.)	binary	first diagnosis in $(60, 365]$	AUROC ↑
New malignancy (Malig.)	binary	first diagnosis in $(60, 365]$	AUROC ↑
Next creatinine value (Creat.)	scalar regression	first creatinine value in $(7, 365]$	MSE ↓
Time to admission (Admit.)	scalar regression	days to the first inpatient admission in $(7, 365]$	MSE ↓

Table 5: Labels of the three diagnosis tasks.

	EHRSHOT		INSPECT	
Task	Labels	Positive rate	Labels	Positive rate
New hyperlipidemia	286	50.0%	1,062	50.0%
New hypertension	296	50.0%	1,298	50.0%
New malignancy	208	50.0%	804	50.0%

New diagnosis assignment. We predict the first recorded occurrence of hyperlipidemia (`SNOMED/55822004`), hypertension (`SNOMED/59621000`), or malignancy (`SNOMED/363346000`), including descendant codes in the ontology. A positive label requires the first occurrence to fall more than 60 and at most 365 days after prediction; a patient whose first occurrence is at or before the end of the 60-day blackout period is ineligible for that task. Otherwise, the patient is negative if no first occurrence falls in the prediction window. We retain at most 1,000 patients per class for each task by random sampling with a fixed seed. The two classes have the same number of patients. Table 5 gives the number of labels of each task.

Next creatinine value. We regress the numeric value of the earliest creatinine measurement coded as LOINC 2160-0 or a descendant code more than 7 and at most 365 days after prediction. Patients without a numeric measurement in this window receive no label; measurements during the first 7 days are ignored. We retain at most 2,000 labeled patients by random sampling with a fixed seed.

Time to admission. We regress the number of days from prediction to the earliest inpatient event coded as `Visit/IP` in `visit_occurrence` or `Visit/ERIP` in `visit_detail` more than 7 and at most 365 days later. Patients without an admission in this window receive no label; admissions during the first 7 days are ignored. We retain at most 2,000 labeled patients by random sampling with a fixed seed.

E Model architectures and hyperparameters

E.1 Summarizers and the embedder

Every summarize-then-embed representation in this paper is produced by the same two-stage pipeline. A language model reads the serialized record and writes a summary in natural language, and a separate embedding model turns that summary into a vector. Neither model is fine-tuned at any point; both are used as released, so within one arm the prompt is the only thing that varies. Generation is greedy, at temperature 0, which makes a summary a deterministic function of its record and its prompt.

E.1.1 ICU

We repeat the whole grid under four summarizers, chosen to span the axes along which the choice of model could matter. **MedGemma-27B** (Sellergren et al., 2025) is the medically adapted one,

Google's Gemma 3 27B further trained on medical text and images; we use its text-only instruction-tuned checkpoint, and it is the only summarizer here that has seen a medical corpus by design. **Gemma-4-26B-A4B** (Team et al., 2026) is the sparse one, a mixture-of-experts model that holds 26B parameters but activates roughly 4B of them per token, so it carries the capacity of a large model at the inference cost of a small one. **Qwen3.6-27B** (Yang et al., 2025) is a dense 27B model from a different pretraining lineage, with every parameter active on every forward pass. **Muse-Glimmer-30B** (Menghini et al., 2026) is a dense 30B model distilled from a larger teacher and tuned for long multi-step instruction following rather than for any medical task. The four therefore differ in scale (26–30B), in architecture (one sparse, three dense), in provenance (three organizations) and in whether medical text was part of the recipe. All four are the instruction-tuned checkpoints.

The embedder is held fixed across all four arms. **Qwen3-Embedding-8B** (Zhang et al., 2025) is the 8B member of a text embedding and reranking family built on Qwen3. We take the final hidden states, mean-pool them over the non-padding positions and L_2-normalize the result, with inputs truncated at 32 768 tokens. Because the embedder never changes, any difference between two rows of one arm is a difference in the text it was given.

E.1.2 LONGITUDINAL EHR

The LEHR experiments use one summarizer, Gemma-4-26B-A4B, with the same greedy decoding at temperature 0. Each record is given to the model in a single user turn that holds the prompt, the demographics header and the rendered record. The output is limited to 4,096 tokens. The input is limited to 120,000 tokens, counting the prompt and the header, and the oldest events are dropped first. A record with no event with a value in its context window is not sent to the model. It receives a fixed text that repeats the header and states that no clinical events were available.

The embedder is Qwen3-Embedding (Zhang et al., 2025), loaded through sentence-transformers. The LEHR code does not pool or normalize the output itself and uses the pooling and normalization in the configuration of the checkpoint. Each input starts with a one-sentence instruction that names the input type as a record, a summary or the demographics. Inputs are limited to 32,700 tokens, counting the instruction and the header, again with the oldest events dropped first. A summary is embedded without the demographics header, which the summarizer has already read.

E.2 DOWNSTREAM PREDICTION HEADS

We evaluate the following downstream prediction heads:

Linear: For classification, regression, and multilabel prediction, we mean-pool the decoded sequence over time and apply a single linear layer. For forecasting, the same linear layer is applied independently at each time step.

MLP: For classification, regression, and multilabel prediction, we flatten the decoded sequence and apply a two-layer MLP with a 256-dimensional hidden layer and ReLU activation. For forecasting, the same two-layer MLP is applied independently at each time step.

LSTM (Hochreiter & Schmidhuber, 1997): We use a two-layer unidirectional LSTM with hidden dimension 256 and no recurrent dropout. For classification, regression, and multilabel prediction, the final hidden state is passed to a linear output layer. For forecasting, the output at every time step is projected to the feature dimension.

xLSTM-Mixer (Kraus et al., 2026): We use an xLSTM-Mixer with embedding dimension 256, two xLSTM blocks, one attention head, and no dropout. For classification, regression, and multilabel prediction, we flatten the output sequence and apply a linear output layer. For forecasting, the sequence produced by the xLSTM-Mixer is used directly.

MLP on LEHR: The MLP head on LEHR takes one vector per patient. It is a two-layer MLP with a 256-dimensional hidden layer, ReLU activation and a single output. Diagnosis tasks use a binary cross-entropy loss weighted by the ratio of negative to positive training records, and regression tasks use the mean squared error on standardized targets.

E.3 TRAINING HYPERPARAMETERS

The downstream prediction heads are trained using the hyperparameters in Table 6. On LEHR, the folds split the source cohort by patient, and early stopping monitors the loss on the held-out source fold.

Table 6: Training hyperparameters of the neural downstream models on the ICU and LEHR datasets.

Hyperparameter	ICU	LEHR
Optimizer	Adam	Adam
Learning rate	10^{-4}	10^{-3}
Weight decay	10^{-4}	10^{-4}
Batch size	256	256
Maximum epochs	100	100
Early-stopping patience	10	10
Number of folds	5	5
Random seed	42	0

F DETAILED RESULTS

F.1 ZERO-SHOT TRANSFER

Table 1 shows one of the six transfer directions, at one head and under one summarizer. This section gives the rest. The grid behind it runs every representation through six directions, five tasks and four heads, and each of those runs is five-fold cross-validated on the source site and evaluated once on the held-out target cohort, so an arm is 120 runs of a representation. Table 7 gives Table 1 in full, with all six directions. Table 20 repeats that setting with the three further prompts described below, so its ranks and marks are taken among sixteen representations.

Other summarization prompts. Besides PORTER and Record2Vec, we compared three summarization prompts of our own: QA, Treatment and Organ-severity trajectory. They read the same input text as PORTER and use the same summarizers and embedding model. QA asks the model to assess short-term prognosis along five fixed points: the main current threat to survival, the expected course over the next 24 to 48 hours without escalation of care, reversibility, the single management change that would help most, and the course of the illness. Each point has to name specific variables, values and hours. Treatment asks how the patient is responding to treatment and how much physiological reserve remains, judged against the level of support rather than against absolute severity, with support anchored on vasopressors, ventilator settings, fluids and sedation. It answers four points in order: the response to treatment, the remaining reserve, whether support is being reduced, held or escalated, and the organ or problem that drives the need for support. Organ-severity trajectory asks for a fixed JSON object over nine organ systems (cardiovascular, respiratory, urinary, digestive, endocrine, immune, musculoskeletal, nervous and reproductive), with an organ-specific severity label, a trajectory label (improving, stable, worsening, fluctuating or not assessable) and a short summary for each. Only the two labels are kept. Each label is written as a short string and embedded on its own, which gives eighteen vectors per window.

None of the three reaches PORTER. Tables 8 to 23 rank the sixteen representations at each of the four heads under each of the four summarizers. In each of these tables PORTER has a better mean rank than each of the three prompts, and Treatment is the best of the three in fifteen of them. Under every summarizer, PORTER has the best mean rank of all sixteen representations at the linear and xLSTM-Mixer heads and is second only to SOFA at the MLP and LSTM heads.

Tables 8 to 11 give the metric values at the MLP head, one table per summarizer. Each table carries its six directions over three blocks of two, with the summary over all six directions on the right as in Table 1.

The remaining tables repeat the same grid at the LSTM, linear and xLSTM-Mixer heads, in the same order of summarizers.

Table 7: Zero-shot transfer, all six directions, with the Qwen3.6 summarizer and the xLSTM-Mixer head. This is Table 1 in full, over the same thirteen methods. The directions are carried over three blocks of two. Cells are the mean over the five cross-validation folds with one standard deviation. The columns to the right of the first block are the summary columns of Table 1. Groups, metrics and marking as in Table 1.

Method	HIRID → MIMIC					HIRID → PPICU				
	mort AUROC↑	los MSE↓	forecast MSE↓	lab AUROC↑	drug AUROC↑	mort AUROC↑	los MSE↓	forecast MSE↓	lab AUROC↑	drug AUROC↑
ATTACK	0.498 ±0.009	3.550 ±0.020	0.264 ±0.099	0.376 ±0.004	0.659 ±0.002	0.500 ±0.001	10.111 ±0.098	0.269 ±0.112	0.484 ±0.004	0.634 ±0.000
RANDOM	0.500 ±0.010	3.541 ±0.077	0.410 ±0.176	0.384 ±0.014	0.659 ±0.004	0.500 ±0.012	9.824 ±0.232	0.451 ±0.193	0.487 ±0.012	0.632 ±0.003
CLMBR-T (Wornow et al., 2023)	0.754 ±0.008	3.044 ±0.108	1.054 ±0.954	0.486 ±0.019	0.765 ±0.009	0.661 ±0.006	7.398 ±0.203	0.972 ±0.895	0.585 ±0.008	0.656 ±0.007
GARLIC (Wang et al., 2026)	0.584 ±0.013	3.098 ±0.085	0.321 ±0.227	0.419 ±0.024	0.681 ±0.007	0.641 ±0.008	8.192 ±0.178	0.314 ±0.214	0.543 ±0.030	0.696 ±0.014
KEDGN (Luo et al., 2024)	0.300 ±0.035	4.081 ±0.627	0.447 ±0.257	0.407 ±0.045	0.375 ±0.077	0.660 ±0.043	14.418 ±2.387	0.459 ±0.290	0.513 ±0.027	0.416 ±0.013
MIRA (Li et al., 2026)	0.646 ±0.023	3.035 ±0.045	0.773 ±0.642	0.404 ±0.028	0.778 ±0.059	0.649 ±0.019	8.096 ±0.248	0.849 ±0.670	0.509 ±0.024	0.691 ±0.022
TabICL (Qu et al., 2025)	0.759 ±0.015	3.725 ±0.148	**0.157** ±0.018	0.430 ±0.015	0.690 ±0.009	0.746 ±0.022	10.281 ±0.430	**0.160** ±0.015	0.537 ±0.025	0.711 ±0.009
MEWS (Subbe et al., 2001)	0.783 ±0.003	2.874 ±0.014	0.512 ±0.338	0.424 ±0.013	0.722 ±0.004	0.682 ±0.005	7.313 ±0.123	0.506 ±0.256	0.491 ±0.005	0.711 ±0.001
SOFA (Vincent et al., 1996)	0.801 ±0.008	2.873 ±0.042	0.346 ±0.349	0.465 ±0.016	0.849 ±0.006	0.729 ±0.005	7.180 ±0.065	0.382 ±0.422	0.537 ±0.007	**0.828** ±0.002
Markdown (Hegselmann et al., 2026)	0.767 ±0.010	2.777 ±0.118	0.269 ±0.113	0.468 ±0.013	0.838 ±0.025	0.682 ±0.004	**6.709** ±0.451	0.289 ±0.155	0.578 ±0.009	0.577 ±0.022
Record2Vec (Ji et al., 2026)	0.819 ±0.007	2.963 ±0.044	0.206 ±0.049	0.452 ±0.006	0.861 ±0.003	0.755 ±0.011	8.714 ±0.179	0.199 ±0.032	0.551 ±0.005	0.767 ±0.003
Template (Demirel et al., 2026)	0.751 ±0.011	2.835 ±0.049	0.233 ±0.104	**0.494** ±0.013	0.897 ±0.007	0.736 ±0.002	7.176 ±0.276	0.266 ±0.149	**0.596** ±0.004	0.650 ±0.012
PORTER	**0.825** ±0.004	**2.773** ±0.012	0.225 ±0.116	0.447 ±0.011	**0.915** ±0.002	**0.765** ±0.003	8.613 ±0.056	0.215 ±0.101	0.576 ±0.012	0.779 ±0.008

Method	MIMIC → HIRID					MIMIC → PPICU				
	mort AUROC↑	los MSE↓	forecast MSE↓	lab AUROC↑	drug AUROC↑	mort AUROC↑	los MSE↓	forecast MSE↓	lab AUROC↑	drug AUROC↑
ATTACK	0.498 ±0.008	0.747 ±0.004	0.160 ±0.048	0.387 ±0.012	0.697 ±0.002	0.500 ±0.000	9.859 ±0.113	0.192 ±0.067	0.488 ±0.015	0.634 ±0.000
RANDOM	0.503 ±0.031	0.791 ±0.019	0.470 ±0.287	0.411 ±0.003	0.692 ±0.006	0.503 ±0.009	9.646 ±0.135	0.443 ±0.259	0.473 ±0.004	0.631 ±0.004
CLMBR-T (Wornow et al., 2023)	0.777 ±0.017	0.564 ±0.055	0.484 ±0.204	0.524 ±0.039	0.799 ±0.004	0.695 ±0.010	7.853 ±0.404	0.426 ±0.183	0.571 ±0.020	0.656 ±0.017
GARLIC (Wang et al., 2026)	0.586 ±0.019	0.610 ±0.050	0.490 ±0.720	0.466 ±0.004	0.727 ±0.017	0.598 ±0.011	9.302 ±0.139	0.464 ±0.603	0.482 ±0.021	0.683 ±0.006
KEDGN (Luo et al., 2024)	0.613 ±0.018	1.082 ±0.195	0.647 ±0.397	0.455 ±0.020	0.572 ±0.064	0.511 ±0.039	7.919 ±0.215	0.619 ±0.391	0.512 ±0.014	0.661 ±0.024
MIRA (Li et al., 2026)	0.749 ±0.033	0.604 ±0.108	0.341 ±0.216	0.418 ±0.038	0.765 ±0.043	0.619 ±0.018	9.049 ±0.455	0.332 ±0.162	0.500 ±0.028	0.674 ±0.005
TabICL (Qu et al., 2025)	0.834 ±0.027	0.630 ±0.038	0.154 ±0.047	0.506 ±0.016	0.763 ±0.011	0.732 ±0.020	9.464 ±0.258	0.204 ±0.061	0.530 ±0.005	0.726 ±0.006
MEWS (Subbe et al., 2001)	0.831 ±0.009	0.552 ±0.022	0.427 ±0.311	0.462 ±0.007	0.748 ±0.006	0.713 ±0.006	7.885 ±0.113	0.473 ±0.346	0.516 ±0.011	0.704 ±0.002
SOFA (Vincent et al., 1996)	0.852 ±0.009	0.539 ±0.015	1.202 ±0.478	0.486 ±0.020	0.875 ±0.005	0.770 ±0.007	7.860 ±0.083	1.114 ±0.381	0.542 ±0.016	**0.844** ±0.004
Markdown (Hegselmann et al., 2026)	0.871 ±0.015	**0.399** ±0.015	0.135 ±0.019	0.543 ±0.010	0.838 ±0.008	0.671 ±0.012	6.557 ±0.366	0.167 ±0.017	0.558 ±0.006	0.631 ±0.018
Record2Vec (Ji et al., 2026)	0.891 ±0.002	0.539 ±0.025	0.129 ±0.009	0.510 ±0.008	0.853 ±0.003	0.776 ±0.003	8.272 ±0.284	0.170 ±0.011	0.534 ±0.008	0.761 ±0.003
Template (Demirel et al., 2026)	0.843 ±0.012	0.481 ±0.049	0.135 ±0.009	**0.558** ±0.015	0.849 ±0.008	0.713 ±0.012	**6.274** ±0.228	0.170 ±0.018	**0.573** ±0.009	0.713 ±0.014
PORTER	**0.901** ±0.002	0.525 ±0.040	**0.126** ±0.010	0.514 ±0.010	**0.916** ±0.001	**0.792** ±0.003	8.093 ±0.269	**0.160** ±0.020	0.547 ±0.005	0.791 ±0.005

Method	PPICU → HIRID					PPICU → MIMIC				
	mort AUROC↑	los MSE↓	forecast MSE↓	lab AUROC↑	drug AUROC↑	mort AUROC↑	los MSE↓	forecast MSE↓	lab AUROC↑	drug AUROC↑
ATTACK	0.502 ±0.006	0.324 ±0.049	0.121 ±0.010	0.513 ±0.010	0.697 ±0.002	0.505 ±0.007	0.643 ±0.038	0.135 ±0.016	0.421 ±0.010	0.661 ±0.003
RANDOM	0.517 ±0.006	0.416 ±0.053	0.428 ±0.145	0.508 ±0.004	0.699 ±0.003	0.508 ±0.010	0.718 ±0.042	0.454 ±0.180	0.434 ±0.006	0.661 ±0.003
CLMBR-T (Wornow et al., 2023)	0.781 ±0.026	0.116 ±0.003	0.285 ±0.280	0.591 ±0.061	0.750 ±0.011	0.730 ±0.031	0.850 ±0.089	0.231 ±0.186	0.491 ±0.034	0.690 ±0.005
GARLIC (Wang et al., 2026)	0.316 ±0.012	0.270 ±0.028	0.698 ±1.056	0.531 ±0.046	0.654 ±0.024	0.520 ±0.014	0.493 ±0.024	0.512 ±0.686	0.467 ±0.011	0.719 ±0.008
KEDGN (Luo et al., 2024)	0.379 ±0.067	0.388 ±0.107	0.646 ±0.622	0.520 ±0.010	0.463 ±0.077	0.666 ±0.106	0.948 ±0.136	0.592 ±0.614	0.471 ±0.027	0.589 ±0.078
MIRA (Li et al., 2026)	0.794 ±0.025	0.110 ±0.008	0.147 ±0.035	0.521 ±0.035	0.792 ±0.008	0.714 ±0.035	0.484 ±0.010	0.179 ±0.035	0.453 ±0.013	0.807 ±0.003
TabICL (Qu et al., 2025)	0.650 ±0.031	0.873 ±0.215	0.161 ±0.038	0.533 ±0.013	0.697 ±0.058	0.648 ±0.028	0.797 ±0.097	0.154 ±0.018	0.512 ±0.015	0.659 ±0.037
MEWS (Subbe et al., 2001)	0.813 ±0.010	0.146 ±0.013	0.317 ±0.162	0.548 ±0.014	0.752 ±0.008	0.804 ±0.003	0.497 ±0.011	0.287 ±0.156	0.487 ±0.003	0.749 ±0.004
SOFA (Vincent et al., 1996)	0.817 ±0.010	0.178 ±0.013	0.259 ±0.174	0.613 ±0.010	**0.857** ±0.003	0.824 ±0.004	0.538 ±0.036	0.273 ±0.149	**0.526** ±0.005	0.823 ±0.004
Markdown (Hegselmann et al., 2026)	**0.899** ±0.003	0.148 ±0.036	**0.106** ±0.006	0.611 ±0.002	0.751 ±0.023	0.819 ±0.009	0.462 ±0.019	**0.126** ±0.008	0.496 ±0.003	0.812 ±0.008
Record2Vec (Ji et al., 2026)	0.887 ±0.002	0.358 ±0.054	0.124 ±0.025	0.602 ±0.005	0.810 ±0.007	0.850 ±0.001	**0.438** ±0.017	0.134 ±0.021	0.504 ±0.005	0.814 ±0.005
Template (Demirel et al., 2026)	0.890 ±0.004	**0.101** ±0.017	0.122 ±0.036	**0.629** ±0.010	0.657 ±0.021	0.813 ±0.006	0.441 ±0.011	0.128 ±0.009	0.510 ±0.008	0.807 ±0.005
PORTER	0.894 ±0.005	0.479 ±0.064	0.120 ±0.014	0.612 ±0.008	0.817 ±0.006	**0.863** ±0.001	0.438 ±0.011	0.130 ±0.013	0.508 ±0.009	**0.828** ±0.006

All six directions			
Mean rank↓	Cumulative occurrences ↑		
	Top-1	Top-2	Top-3
11 (10.1)	0	0	1
12 (11.1)	0	0	0
6 (7.1)	0	3	6
10 (9.4)	0	0	0
13 (11.1)	0	0	0
9 (8.0)	0	1	1
7 (7.2)	2	3	4
8 (7.3)	0	0	0
5 (5.0)	4	7	11
4 (4.3)	5	9	13
3 (4.0)	1	9	15
2 (3.6)	7	11	17
1 (2.9)	**11**	**17**	**22**

F.2 FEW-SHOT FINE-TUNING

Figure 4 plots all ten panels of the few-shot experiment. The panels cover two transfer directions into PPICU, with five tasks in each direction. Table 24 gives the values behind the panels. As in Section 4, we independently sample five support sets of 16 labeled target-site records. For each support set, every evaluated method is fine-tuned on those 16 records under the same schedule, and the site context of the "+ policy" methods is distilled from the same 16 records. Each cell of that table uses an MLP head and reports the mean over the five support sets, with one standard deviation across them. Tables 25 and 26 repeat the grid with the linear and the LSTM head, so every method is reported with all three heads. *PORTER + policy* has the best mean rank over the ten columns with each of the three heads. Its mean rank is 1.4 with the MLP head, 1.5 with the linear head and 1.8 with the LSTM head. It is also first when each method is taken at its own best head. The choice of head therefore does not drive the result. No in-distribution reference was run for this experiment, so the panels have no ceiling and compare the twelve methods only with each other.

The dashed line in each panel of Figure 4 is the best of the ten fine-tuned baselines in that panel. Four different baselines hold it across the ten panels. SOFA and Markdown hold it in three panels each, and CLMBR-T and *Record2Vec + policy* in two each. Beating the line means beating the strongest baseline for that task. This baseline is not the same method in every panel. *PORTER + policy* beats the line in eight panels, and *PORTER* beats it in three. The two panels where *PORTER + policy* does not beat the line are `drug` in both directions. SOFA holds the line in both. These are also the only two of the ten columns where adding the site context makes the representation worse than *PORTER* alone.

To further evaluate the robustness of our policy-distillation method, we conduct an ablation study using clinically atypical cases. We distill policies from these cases and fine-tune models across all methods for comparison. Table 27 summarizes the impact of these outliers on fine-tuning performance,

Table 8: Zero-shot transfer, all six directions, the MedGemma arm at the MLP head. The directions are carried over three blocks of two. Cells are the mean over the five cross-validation folds with one standard deviation. The columns to the right of the first block summarize all six directions. *Mean rank* is each method's position among the sixteen by its average rank over the thirty columns, with that average in parentheses, and *Cumulative occurrences* counts the columns, out of thirty, in which a method places in the top-1, 2, 3. Groups, metrics and marking as in Table 1.

Method	HIRID → MIMIC					HIRID → PPICU				
	mort AUROC↑	los MSE↓	forecast MSE↓	lab AUROC↑	drug AUROC↑	mort AUROC↑	los MSE↓	forecast MSE↓	lab AUROC↑	drug AUROC↑
ATTACK	0.500 ±0.000	3.504 ±0.017	0.127 ±0.000	0.373 ±0.011	0.661 ±0.001	0.500 ±0.001	9.869 ±0.097	0.163 ±0.002	0.487 ±0.005	0.634 ±0.000
RANDOM	0.509 ±0.007	3.603 ±0.111	0.135 ±0.001	0.377 ±0.003	0.664 ±0.003	0.498 ±0.009	10.003 ±0.301	0.163 ±0.001	0.495 ±0.008	0.632 ±0.002
CLMBR-T (Wornow et al., 2023)	0.756 ±0.004	2.971 ±0.091	0.131 ±0.002	0.481 ±0.014	0.773 ±0.007	0.662 ±0.008	7.760 ±0.305	0.158 ±0.002	0.573 ±0.007	0.667 ±0.008
GARLIC (Wang et al., 2026)	0.547 ±0.021	3.278 ±0.081	0.124 ±0.001	0.412 ±0.006	0.680 ±0.014	0.613 ±0.017	9.101 ±0.134	0.163 ±0.002	0.562 ±0.007	0.668 ±0.006
KEDGN (Luo et al., 2024)	0.299 ±0.022	3.615 ±0.154	**0.108** ±0.002	0.442 ±0.019	0.396 ±0.071	0.616 ±0.062	10.692 ±0.271	**0.126** ±0.005	0.486 ±0.022	0.430 ±0.032
MIRA (Li et al., 2026)	0.658 ±0.011	3.012 ±0.050	0.127 ±0.002	0.433 ±0.009	0.800 ±0.004	0.667 ±0.005	8.001 ±0.150	0.160 ±0.004	0.528 ±0.015	0.695 ±0.005
TabICL (Qu et al., 2025)	0.751 ±0.020	3.742 ±0.072	0.142 ±0.002	0.431 ±0.010	0.686 ±0.007	0.745 ±0.017	10.753 ±0.392	0.190 ±0.003	0.542 ±0.010	0.708 ±0.013
MEWS (Subbe et al., 2001)	0.775 ±0.008	2.865 ±0.049	0.130 ±0.000	0.424 ±0.024	0.721 ±0.002	0.675 ±0.011	7.462 ±0.215	0.156 ±0.001	0.506 ±0.015	0.709 ±0.002
SOFA (Vincent et al., 1996)	**0.800** ±0.006	2.937 ±0.103	0.133 ±0.001	0.453 ±0.029	0.855 ±0.004	0.733 ±0.006	7.234 ±0.286	0.155 ±0.002	0.538 ±0.014	**0.825** ±0.006
Markdown (Hegselmann et al., 2026)	0.766 ±0.014	2.776 ±0.085	0.134 ±0.001	0.477 ±0.013	0.850 ±0.018	0.687 ±0.011	**6.813** ±0.207	0.157 ±0.002	0.583 ±0.008	0.592 ±0.015
Record2Vec (Ji et al., 2026)	0.751 ±0.005	3.017 ±0.054	0.134 ±0.001	0.447 ±0.005	0.834 ±0.003	0.725 ±0.002	8.553 ±0.236	0.161 ±0.001	0.547 ±0.007	0.722 ±0.003
Template (Demirel et al., 2026)	0.749 ±0.008	**2.744** ±0.218	0.136 ±0.001	**0.497** ±0.014	**0.906** ±0.002	0.737 ±0.003	7.676 ±0.124	0.156 ±0.002	**0.593** ±0.006	0.676 ±0.014
QA	0.736 ±0.005	3.227 ±0.051	0.132 ±0.001	0.439 ±0.003	0.771 ±0.009	0.672 ±0.007	8.927 ±0.258	0.162 ±0.002	0.521 ±0.010	0.696 ±0.002
Treatment	0.743 ±0.006	3.050 ±0.039	0.133 ±0.002	0.442 ±0.004	0.829 ±0.002	0.706 ±0.004	7.746 ±0.384	0.161 ±0.003	0.527 ±0.007	0.735 ±0.006
Organ-severity trajectory	0.772 ±0.010	3.332 ±0.015	0.133 ±0.001	0.441 ±0.011	0.858 ±0.005	0.722 ±0.012	9.798 ±0.149	0.163 ±0.002	0.568 ±0.003	0.744 ±0.004
PORTER	0.791 ±0.008	2.863 ±0.049	0.134 ±0.000	0.442 ±0.005	0.893 ±0.002	**0.756** ±0.005	9.110 ±0.186	0.159 ±0.001	0.554 ±0.004	0.769 ±0.003

Method	MIMIC → HIRID					MIMIC → PPICU				
	mort AUROC↑	los MSE↓	forecast MSE↓	lab AUROC↑	drug AUROC↑	mort AUROC↑	los MSE↓	forecast MSE↓	lab AUROC↑	drug AUROC↑
ATTACK	0.498 ±0.004	0.749 ±0.003	0.125 ±0.001	0.394 ±0.013	0.697 ±0.001	0.500 ±0.001	9.775 ±0.089	0.188 ±0.001	0.496 ±0.023	0.633 ±0.000
RANDOM	0.481 ±0.012	0.871 ±0.088	0.122 ±0.001	0.398 ±0.004	0.696 ±0.004	0.502 ±0.008	9.863 ±0.148	**0.168** ±0.004	0.477 ±0.003	0.634 ±0.001
CLMBR-T (Wornow et al., 2023)	0.778 ±0.012	0.566 ±0.035	0.125 ±0.003	0.510 ±0.012	0.784 ±0.014	0.687 ±0.009	8.117 ±0.253	0.178 ±0.003	0.561 ±0.009	0.634 ±0.017
GARLIC (Wang et al., 2026)	0.591 ±0.015	0.608 ±0.024	**0.120** ±0.002	0.489 ±0.017	0.667 ±0.014	0.593 ±0.014	9.682 ±0.162	0.183 ±0.003	0.510 ±0.011	0.681 ±0.004
KEDGN (Luo et al., 2024)	0.651 ±0.034	1.739 ±0.599	0.137 ±0.002	0.481 ±0.011	0.598 ±0.057	0.544 ±0.062	9.748 ±0.726	0.185 ±0.003	0.525 ±0.011	0.651 ±0.018
MIRA (Li et al., 2026)	0.743 ±0.007	0.507 ±0.051	0.126 ±0.002	0.476 ±0.022	0.786 ±0.008	0.590 ±0.004	8.537 ±0.178	0.184 ±0.003	0.519 ±0.007	0.679 ±0.005
TabICL (Qu et al., 2025)	0.816 ±0.015	0.643 ±0.049	0.121 ±0.002	0.505 ±0.011	0.749 ±0.017	0.716 ±0.016	10.445 ±0.266	0.176 ±0.006	0.523 ±0.005	0.698 ±0.037
MEWS (Subbe et al., 2001)	0.835 ±0.012	0.545 ±0.011	0.135 ±0.001	0.462 ±0.010	0.732 ±0.005	0.717 ±0.007	8.318 ±0.118	0.189 ±0.001	0.520 ±0.003	0.704 ±0.003
SOFA (Vincent et al., 1996)	0.837 ±0.013	0.578 ±0.011	0.131 ±0.001	0.490 ±0.005	0.873 ±0.005	0.758 ±0.006	8.361 ±0.125	0.189 ±0.001	0.547 ±0.002	**0.841** ±0.001
Markdown (Hegselmann et al., 2026)	**0.870** ±0.009	**0.375** ±0.005	0.138 ±0.002	0.536 ±0.013	0.845 ±0.005	0.674 ±0.013	6.804 ±0.138	0.198 ±0.002	0.557 ±0.005	0.651 ±0.019
Record2Vec (Ji et al., 2026)	0.836 ±0.012	0.517 ±0.018	0.131 ±0.001	0.494 ±0.007	0.837 ±0.005	0.737 ±0.008	8.744 ±0.310	0.190 ±0.003	0.524 ±0.004	0.733 ±0.004
Template (Demirel et al., 2026)	0.826 ±0.031	0.440 ±0.029	0.138 ±0.001	**0.554** ±0.011	0.857 ±0.008	0.709 ±0.003	**6.425** ±0.191	0.195 ±0.003	**0.569** ±0.004	0.727 ±0.016
QA	0.802 ±0.005	0.600 ±0.018	0.130 ±0.002	0.453 ±0.010	0.782 ±0.004	0.679 ±0.007	9.038 ±0.174	0.190 ±0.003	0.503 ±0.005	0.672 ±0.005
Treatment	0.817 ±0.005	0.572 ±0.020	0.130 ±0.002	0.481 ±0.016	0.850 ±0.000	0.730 ±0.003	8.795 ±0.414	0.190 ±0.003	0.522 ±0.008	0.734 ±0.003
Organ-severity trajectory	0.820 ±0.005	0.639 ±0.011	0.134 ±0.001	0.449 ±0.022	0.875 ±0.004	0.727 ±0.007	8.462 ±0.234	0.193 ±0.003	0.526 ±0.027	0.749 ±0.003
PORTER	0.866 ±0.002	0.488 ±0.009	0.134 ±0.001	0.520 ±0.006	**0.891** ±0.003	**0.758** ±0.009	8.572 ±0.192	0.190 ±0.001	0.534 ±0.004	0.765 ±0.001

Method	PPICU → HIRID					PPICU → MIMIC				
	mort AUROC↑	los MSE↓	forecast MSE↓	lab AUROC↑	drug AUROC↑	mort AUROC↑	los MSE↓	forecast MSE↓	lab AUROC↑	drug AUROC↑
ATTACK	0.501 ±0.000	0.368 ±0.018	0.105 ±0.001	0.514 ±0.003	0.698 ±0.001	0.500 ±0.000	0.675 ±0.014	0.114 ±0.001	0.425 ±0.010	0.661 ±0.002
RANDOM	0.511 ±0.016	0.402 ±0.010	0.111 ±0.003	0.505 ±0.004	0.699 ±0.005	0.510 ±0.009	0.706 ±0.010	0.119 ±0.001	0.430 ±0.006	0.658 ±0.002
CLMBR-T (Wornow et al., 2023)	0.696 ±0.052	**0.100** ±0.011	0.107 ±0.001	**0.625** ±0.006	0.738 ±0.007	0.660 ±0.016	0.812 ±0.200	0.113 ±0.002	0.507 ±0.005	0.677 ±0.023
GARLIC (Wang et al., 2026)	0.309 ±0.020	0.311 ±0.088	0.101 ±0.001	0.594 ±0.012	0.448 ±0.031	0.508 ±0.016	0.490 ±0.029	0.112 ±0.001	0.492 ±0.010	0.653 ±0.017
KEDGN (Luo et al., 2024)	0.382 ±0.071	0.228 ±0.123	**0.090** ±0.001	0.547 ±0.027	0.422 ±0.073	0.626 ±0.050	0.678 ±0.061	**0.101** ±0.001	0.470 ±0.020	0.578 ±0.092
MIRA (Li et al., 2026)	0.782 ±0.014	0.103 ±0.011	0.107 ±0.000	0.582 ±0.037	0.792 ±0.005	0.738 ±0.006	0.478 ±0.003	0.116 ±0.001	0.464 ±0.025	0.815 ±0.003
TabICL (Qu et al., 2025)	0.644 ±0.043	1.924 ±0.840	0.098 ±0.001	0.524 ±0.011	0.678 ±0.062	0.620 ±0.019	0.881 ±0.133	0.123 ±0.001	0.515 ±0.011	0.649 ±0.056
MEWS (Subbe et al., 2001)	0.815 ±0.017	0.109 ±0.008	0.109 ±0.001	0.549 ±0.007	0.752 ±0.011	0.802 ±0.003	0.460 ±0.011	0.120 ±0.001	0.486 ±0.003	0.750 ±0.006
SOFA (Vincent et al., 1996)	0.813 ±0.006	0.130 ±0.007	0.106 ±0.001	0.619 ±0.002	**0.859** ±0.005	0.823 ±0.004	0.461 ±0.020	0.120 ±0.001	**0.521** ±0.002	**0.838** ±0.002
Markdown (Hegselmann et al., 2026)	**0.903** ±0.001	0.127 ±0.029	0.104 ±0.001	0.602 ±0.008	0.744 ±0.011	0.819 ±0.006	0.445 ±0.011	0.117 ±0.001	0.489 ±0.001	0.820 ±0.008
Record2Vec (Ji et al., 2026)	0.834 ±0.006	0.333 ±0.047	0.108 ±0.001	0.594 ±0.005	0.748 ±0.004	0.804 ±0.003	0.478 ±0.027	0.118 ±0.001	0.502 ±0.005	0.761 ±0.005
Template (Demirel et al., 2026)	0.888 ±0.006	0.106 ±0.018	0.105 ±0.001	0.621 ±0.005	0.670 ±0.043	0.820 ±0.007	**0.427** ±0.012	0.122 ±0.002	0.510 ±0.004	0.809 ±0.010
QA	0.778 ±0.012	0.269 ±0.034	0.106 ±0.001	0.572 ±0.006	0.754 ±0.007	0.757 ±0.006	0.479 ±0.015	0.117 ±0.001	0.490 ±0.006	0.744 ±0.009
Treatment	0.824 ±0.009	0.211 ±0.031	0.108 ±0.000	0.582 ±0.006	0.817 ±0.003	0.788 ±0.003	0.431 ±0.004	0.119 ±0.001	0.498 ±0.005	0.806 ±0.004
Organ-severity trajectory	0.838 ±0.004	0.325 ±0.035	0.108 ±0.001	0.589 ±0.004	0.816 ±0.005	**0.832** ±0.003	0.517 ±0.012	0.119 ±0.001	0.500 ±0.008	0.800 ±0.005
PORTER	0.863 ±0.004	0.332 ±0.105	0.108 ±0.001	0.606 ±0.004	0.800 ±0.006	0.824 ±0.005	0.438 ±0.023	0.119 ±0.001	0.509 ±0.004	0.807 ±0.005

All six directions			
Mean rank↓	Cumulative occurrences↑		
	Top-1	Top-2	Top-3
15 (12.7)	0	0	1
16 (13.5)	1	1	2
5 (7.1)	2	4	8
12 (10.2)	1	3	4
14 (11.6)	4	4	4
10 (8.3)	0	1	2
13 (10.5)	0	5	5
9 (8.2)	0	0	2
1 (4.8)	6	9	13
4 (6.0)	4	9	11
6 (7.4)	0	0	1
3 (5.4)	**8**	**11**	**15**
11 (9.7)	0	0	0
7 (7.5)	0	2	2
8 (7.7)	1	2	6
2 (5.4)	3	9	14

showing that our method mitigates the adverse effects and maintains nearly unchanged performance across all tasks.

Record2Vec is given site context in the same form, distilled from the same five support sets. Over the ten columns, the site context moves Record2Vec from a mean rank of 6.1 to 5.3. It moves our representation from 3.5 to 1.4. How much the site context helps therefore depends on the representation it is added to.

F.3 Prompt ablations

Figure 5 ranks the seven prompts within the Qwen3.6 arm, and Table 28 repeats the ranking for all four summarizers. Two results hold in every arm. First, *free-form*, which deletes all thirteen components, has the worst mean rank of the seven in all four arms. It is ranked last in between a quarter and a half of each arm's runs, and no other variant has the worst mean rank in any arm. Second, *static* and *no-course* are the two best variants in all four arms, and both rank ahead of the *full* prompt in every arm. Both keep the clinical section scaffold. *Static* describes the patient's state at the end of the window and removes temporal guidance, and *no-course* removes the overall-course summary. Under Muse, *no-course* ties with *reword* for second place.

The middle of the ordering is less stable. *Reword* is a control that paraphrases the *full* prompt without changing its content, so any gap between the two comes from wording alone. Under Qwen3.6, *reword* ranks below every variant except *free-form*. In the other three arms it ranks higher. Under

Table 9: Zero-shot transfer, all six directions, the Gemma-4 arm at the MLP head. The directions are carried over three blocks of two. Cells are the mean over the five cross-validation folds with one standard deviation. The columns to the right of the first block summarize all six directions. *Mean rank* is each method's position among the sixteen by its average rank over the thirty columns, with that average in parentheses, and *Cumulative occurrences* counts the columns, out of thirty, in which a method places in the top-1, 2, 3. Groups, metrics and marking as in Table 1.

Method	HIRID → MIMIC					HIRID → PPICU				
	mort AUROC↑	los MSE↓	forecast MSE↓	lab AUROC↑	drug AUROC↑	mort AUROC↑	los MSE↓	forecast MSE↓	lab AUROC↑	drug AUROC↑
ATTACK	0.500 ±0.000	3.504 ±0.017	0.127 ±0.000	0.373 ±0.011	0.661 ±0.001	0.500 ±0.001	9.869 ±0.097	0.163 ±0.002	0.487 ±0.005	0.634 ±0.000
RANDOM	0.509 ±0.007	3.603 ±0.111	0.135 ±0.001	0.377 ±0.003	0.664 ±0.003	0.498 ±0.009	10.003 ±0.301	0.163 ±0.001	0.495 ±0.008	0.632 ±0.002
CLMBR-T (Wornow et al., 2023)	0.756 ±0.004	2.971 ±0.091	0.131 ±0.002	0.481 ±0.014	0.773 ±0.007	0.662 ±0.008	7.760 ±0.305	0.158 ±0.002	0.573 ±0.007	0.667 ±0.008
GARLIC (Wang et al., 2026)	0.547 ±0.021	3.278 ±0.081	0.124 ±0.001	0.412 ±0.006	0.680 ±0.014	0.613 ±0.017	9.101 ±0.134	0.163 ±0.002	0.562 ±0.007	0.668 ±0.006
KEDGN (Luo et al., 2024)	0.299 ±0.022	3.615 ±0.154	**0.108** ±0.002	0.442 ±0.019	0.396 ±0.071	0.616 ±0.062	10.692 ±0.271	**0.126** ±0.005	0.486 ±0.022	0.430 ±0.032
MIRA (Li et al., 2026)	0.658 ±0.011	3.012 ±0.050	0.127 ±0.002	0.433 ±0.009	0.800 ±0.004	0.667 ±0.005	8.001 ±0.150	0.160 ±0.004	0.528 ±0.015	0.695 ±0.005
TabICL (Qu et al., 2025)	0.751 ±0.020	3.742 ±0.072	0.142 ±0.002	0.431 ±0.010	0.686 ±0.007	0.745 ±0.017	10.753 ±0.392	0.190 ±0.003	0.542 ±0.010	0.708 ±0.013
MEWS (Subbe et al., 2001)	0.775 ±0.008	2.865 ±0.049	0.130 ±0.000	0.424 ±0.024	0.721 ±0.002	0.675 ±0.011	7.462 ±0.215	0.156 ±0.001	0.506 ±0.015	0.709 ±0.002
SOFA (Vincent et al., 1996)	0.800 ±0.006	2.937 ±0.103	0.133 ±0.001	0.453 ±0.029	0.855 ±0.004	0.733 ±0.006	7.234 ±0.286	0.155 ±0.002	0.538 ±0.014	**0.825** ±0.006
Markdown (Hegselmann et al., 2026)	0.766 ±0.014	2.776 ±0.085	0.134 ±0.001	0.477 ±0.013	0.850 ±0.018	0.687 ±0.011	**6.813** ±0.207	0.157 ±0.002	0.583 ±0.008	0.592 ±0.015
Record2Vec (Ji et al., 2026)	0.811 ±0.006	2.986 ±0.024	0.135 ±0.002	0.447 ±0.009	0.844 ±0.006	0.744 ±0.003	8.278 ±0.128	0.160 ±0.001	0.554 ±0.009	0.750 ±0.001
Template (Demirel et al., 2026)	0.749 ±0.008	**2.744** ±0.218	0.136 ±0.001	**0.497** ±0.014	0.906 ±0.002	0.737 ±0.003	7.676 ±0.124	0.156 ±0.002	**0.593** ±0.006	0.676 ±0.014
QA	0.794 ±0.008	3.163 ±0.040	0.134 ±0.001	0.445 ±0.004	0.811 ±0.004	0.736 ±0.005	8.855 ±0.128	0.160 ±0.002	0.540 ±0.006	0.723 ±0.003
Treatment	0.774 ±0.010	3.137 ±0.060	0.136 ±0.001	0.435 ±0.006	0.836 ±0.005	0.730 ±0.010	8.561 ±0.259	0.162 ±0.001	0.526 ±0.003	0.744 ±0.005
Organ-severity trajectory	0.794 ±0.003	3.178 ±0.088	0.132 ±0.001	0.431 ±0.006	0.865 ±0.003	0.760 ±0.005	9.148 ±0.324	0.163 ±0.001	0.556 ±0.006	0.735 ±0.002
PORTER	**0.823** ±0.005	2.920 ±0.043	0.138 ±0.001	0.439 ±0.006	**0.909** ±0.003	**0.770** ±0.005	8.687 ±0.091	0.162 ±0.002	0.567 ±0.003	0.772 ±0.006

Method	MIMIC → HIRID					MIMIC → PPICU				
	mort AUROC↑	los MSE↓	forecast MSE↓	lab AUROC↑	drug AUROC↑	mort AUROC↑	los MSE↓	forecast MSE↓	lab AUROC↑	drug AUROC↑
ATTACK	0.498 ±0.004	0.749 ±0.003	0.125 ±0.001	0.394 ±0.013	0.697 ±0.001	0.500 ±0.001	9.775 ±0.089	0.188 ±0.001	0.496 ±0.023	0.633 ±0.000
RANDOM	0.481 ±0.012	0.871 ±0.088	0.122 ±0.001	0.398 ±0.004	0.696 ±0.004	0.502 ±0.008	9.863 ±0.148	**0.168** ±0.004	0.477 ±0.003	0.634 ±0.001
CLMBR-T (Wornow et al., 2023)	0.778 ±0.012	0.566 ±0.035	0.125 ±0.003	0.510 ±0.012	0.784 ±0.014	0.687 ±0.009	8.117 ±0.253	0.178 ±0.003	0.561 ±0.009	0.634 ±0.017
GARLIC (Wang et al., 2026)	0.591 ±0.015	0.608 ±0.024	**0.120** ±0.002	0.489 ±0.017	0.667 ±0.014	0.593 ±0.014	9.682 ±0.162	0.183 ±0.003	0.510 ±0.011	0.681 ±0.004
KEDGN (Luo et al., 2024)	0.651 ±0.034	1.739 ±0.599	0.137 ±0.002	0.481 ±0.011	0.598 ±0.057	0.544 ±0.062	9.748 ±0.726	0.185 ±0.003	0.525 ±0.011	0.651 ±0.018
MIRA (Li et al., 2026)	0.743 ±0.007	0.507 ±0.051	0.126 ±0.002	0.476 ±0.022	0.786 ±0.008	0.590 ±0.004	8.537 ±0.178	0.184 ±0.003	0.519 ±0.007	0.679 ±0.005
TabICL (Qu et al., 2025)	0.816 ±0.015	0.643 ±0.049	0.121 ±0.002	0.505 ±0.011	0.749 ±0.017	0.716 ±0.016	10.445 ±0.266	0.176 ±0.006	0.523 ±0.005	0.698 ±0.037
MEWS (Subbe et al., 2001)	0.835 ±0.012	0.545 ±0.011	0.135 ±0.001	0.462 ±0.010	0.732 ±0.005	0.717 ±0.007	8.318 ±0.118	0.189 ±0.001	0.520 ±0.003	0.704 ±0.003
SOFA (Vincent et al., 1996)	0.837 ±0.013	0.578 ±0.011	0.131 ±0.001	0.490 ±0.005	0.873 ±0.005	0.758 ±0.006	8.361 ±0.125	0.189 ±0.001	0.547 ±0.002	**0.841** ±0.001
Markdown (Hegselmann et al., 2026)	0.870 ±0.009	**0.375** ±0.005	0.138 ±0.002	0.536 ±0.013	0.845 ±0.005	0.674 ±0.013	6.804 ±0.138	0.198 ±0.002	0.557 ±0.005	0.651 ±0.019
Record2Vec (Ji et al., 2026)	0.887 ±0.006	0.508 ±0.021	0.131 ±0.002	0.500 ±0.007	0.833 ±0.002	0.768 ±0.003	8.439 ±0.226	0.189 ±0.004	0.545 ±0.005	0.741 ±0.002
Template (Demirel et al., 2026)	0.826 ±0.031	0.440 ±0.029	0.138 ±0.001	**0.554** ±0.011	0.857 ±0.008	0.709 ±0.003	**6.425** ±0.191	0.195 ±0.003	**0.569** ±0.004	0.727 ±0.016
QA	0.841 ±0.018	0.645 ±0.034	0.131 ±0.002	0.447 ±0.006	0.783 ±0.010	0.757 ±0.003	8.915 ±0.144	0.191 ±0.002	0.526 ±0.004	0.715 ±0.003
Treatment	0.838 ±0.009	0.592 ±0.037	0.130 ±0.001	0.456 ±0.006	0.830 ±0.002	0.765 ±0.004	8.940 ±0.141	0.188 ±0.003	0.527 ±0.005	0.740 ±0.002
Organ-severity trajectory	0.838 ±0.008	0.729 ±0.074	0.138 ±0.003	0.462 ±0.032	0.854 ±0.005	0.783 ±0.005	9.067 ±0.594	0.194 ±0.002	0.535 ±0.025	0.743 ±0.003
PORTER	**0.900** ±0.003	0.542 ±0.060	0.133 ±0.002	0.522 ±0.006	**0.915** ±0.002	**0.794** ±0.004	8.504 ±0.299	0.190 ±0.003	0.547 ±0.004	0.783 ±0.003

Method	PPICU → HIRID					PPICU → MIMIC				
	mort AUROC↑	los MSE↓	forecast MSE↓	lab AUROC↑	drug AUROC↑	mort AUROC↑	los MSE↓	forecast MSE↓	lab AUROC↑	drug AUROC↑
ATTACK	0.501 ±0.000	0.368 ±0.018	0.105 ±0.001	0.514 ±0.003	0.698 ±0.001	0.500 ±0.000	0.675 ±0.014	0.114 ±0.001	0.425 ±0.010	0.661 ±0.002
RANDOM	0.511 ±0.016	0.402 ±0.010	0.111 ±0.003	0.505 ±0.004	0.699 ±0.005	0.510 ±0.009	0.706 ±0.010	0.119 ±0.001	0.430 ±0.006	0.658 ±0.002
CLMBR-T (Wornow et al., 2023)	0.696 ±0.052	**0.100** ±0.011	0.107 ±0.001	**0.625** ±0.006	0.738 ±0.007	0.660 ±0.016	0.812 ±0.200	0.113 ±0.002	0.507 ±0.005	0.677 ±0.023
GARLIC (Wang et al., 2026)	0.309 ±0.020	0.311 ±0.088	0.101 ±0.001	0.594 ±0.012	0.448 ±0.031	0.508 ±0.016	0.490 ±0.029	0.112 ±0.001	0.492 ±0.010	0.653 ±0.017
KEDGN (Luo et al., 2024)	0.382 ±0.071	0.228 ±0.123	**0.090** ±0.001	0.547 ±0.027	0.422 ±0.073	0.626 ±0.050	0.678 ±0.061	**0.101** ±0.001	0.470 ±0.020	0.578 ±0.092
MIRA (Li et al., 2026)	0.782 ±0.014	0.103 ±0.011	0.107 ±0.000	0.582 ±0.037	0.792 ±0.005	0.738 ±0.006	0.478 ±0.003	0.116 ±0.001	0.464 ±0.025	0.815 ±0.003
TabICL (Qu et al., 2025)	0.644 ±0.043	1.924 ±0.840	0.098 ±0.001	0.524 ±0.011	0.678 ±0.062	0.620 ±0.019	0.881 ±0.133	0.123 ±0.001	0.515 ±0.011	0.649 ±0.056
MEWS (Subbe et al., 2001)	0.815 ±0.017	0.109 ±0.008	0.109 ±0.001	0.549 ±0.007	0.752 ±0.011	0.802 ±0.003	0.460 ±0.011	0.120 ±0.001	0.486 ±0.003	0.750 ±0.006
SOFA (Vincent et al., 1996)	0.813 ±0.006	0.130 ±0.007	0.106 ±0.001	0.619 ±0.002	**0.859** ±0.005	0.823 ±0.004	0.461 ±0.020	0.120 ±0.001	**0.521** ±0.002	**0.838** ±0.002
Markdown (Hegselmann et al., 2026)	**0.903** ±0.001	0.127 ±0.029	0.104 ±0.001	0.602 ±0.008	0.744 ±0.011	0.819 ±0.006	0.445 ±0.011	0.117 ±0.001	0.489 ±0.001	0.820 ±0.008
Record2Vec (Ji et al., 2026)	0.862 ±0.005	0.469 ±0.076	0.108 ±0.000	0.602 ±0.001	0.805 ±0.004	0.824 ±0.008	0.441 ±0.012	0.120 ±0.001	0.506 ±0.002	0.759 ±0.006
Template (Demirel et al., 2026)	0.888 ±0.006	0.106 ±0.018	0.105 ±0.001	0.621 ±0.005	0.670 ±0.043	0.820 ±0.007	0.427 ±0.012	0.122 ±0.002	0.510 ±0.004	0.809 ±0.010
QA	0.795 ±0.029	0.330 ±0.034	0.108 ±0.001	0.562 ±0.005	0.786 ±0.006	0.817 ±0.005	0.476 ±0.014	0.118 ±0.001	0.493 ±0.003	0.766 ±0.014
Treatment	0.813 ±0.013	0.398 ±0.058	0.107 ±0.001	0.575 ±0.007	0.781 ±0.011	0.827 ±0.001	0.477 ±0.022	0.118 ±0.001	0.498 ±0.002	0.796 ±0.003
Organ-severity trajectory	0.817 ±0.013	0.637 ±0.106	0.111 ±0.001	0.583 ±0.002	0.770 ±0.007	0.832 ±0.003	0.499 ±0.012	0.121 ±0.001	0.499 ±0.003	0.807 ±0.005
PORTER	0.845 ±0.019	0.471 ±0.076	0.108 ±0.001	0.609 ±0.003	0.796 ±0.008	**0.860** ±0.001	**0.426** ±0.014	0.120 ±0.001	0.507 ±0.003	0.823 ±0.008

Mean rank↓	All six directions: Cumulative occurrences ↑ Top-1	Top-2	Top-3
15 (12.6)	0	0	1
16 (13.2)	1	1	2
6 (7.2)	2	4	8
12 (10.2)	1	3	4
14 (11.5)	4	4	4
8 (8.4)	0	1	2
13 (10.7)	0	4	5
11 (8.6)	0	0	2
1 (5.0)	5	8	10
4 (6.2)	3	7	11
5 (6.2)	0	3	7
3 (5.8)	6	**11**	**15**
10 (8.5)	0	0	0
7 (8.0)	0	0	1
9 (8.5)	0	3	5
2 (5.3)	**8**	**11**	13

MedGemma it ranks above *generic*, and under Gemma-4 and Muse it ranks above *generic*, *bare* and the *full* prompt. Rewording alone can therefore move a prompt across the middle of the ordering. We read only the two ends of the ordering as effects of prompt structure, with *free-form* at the bottom and *static* and *no-course* at the top.

G LONGITUDINAL EHR

TASKS

We construct three binary tasks predicting a new diagnosis and two regression tasks predicting time to inpatient admission and a future creatinine value. We first sample a prediction time for each patient and use the same prediction time for all tasks.

New diagnosis assignment. We predict the first recorded occurrence of hyperlipidemia (`SNOMED/55822004`), malignancy (`SNOMED/363346000`), or hypertension (`SNOMED/59621000`), including descendant codes in the ontology. A positive label requires the first occurrence to fall more than 60 and at most 365 days after prediction; a patient whose first occurrence is at or before the end of the 60-day blackout period is ineligible for that task. Otherwise, the patient is negative if no first occurrence falls in the prediction window. We retain at most 1,000 patients per class for each task.

Table 10: Zero-shot transfer, all six directions, the Qwen3.6 arm at the MLP head. The directions are carried over three blocks of two. Cells are the mean over the five cross-validation folds with one standard deviation. The columns to the right of the first block summarize all six directions. *Mean rank* is each method's position among the sixteen by its average rank over the thirty columns, with that average in parentheses, and *Cumulative occurrences* counts the columns, out of thirty, in which a method places in the top-1, 2, 3. Groups, metrics and marking as in Table 1.

Method	HIRID → MIMIC					HIRID → PPICU				
	mort AUROC↑	los MSE↓	forecast MSE↓	lab AUROC↑	drug AUROC↑	mort AUROC↑	los MSE↓	forecast MSE↓	lab AUROC↑	drug AUROC↑
ATTACK	0.500 ±0.000	3.504 ±0.017	0.127 ±0.000	0.373 ±0.011	0.661 ±0.001	0.500 ±0.001	9.869 ±0.097	0.163 ±0.002	0.487 ±0.005	0.634 ±0.000
RANDOM	0.509 ±0.007	3.603 ±0.111	0.135 ±0.001	0.377 ±0.003	0.664 ±0.003	0.498 ±0.009	10.003 ±0.301	0.163 ±0.001	0.495 ±0.008	0.632 ±0.002
CLMBR-T (Wornow et al., 2023)	0.756 ±0.004	2.971 ±0.091	0.131 ±0.002	0.481 ±0.014	0.773 ±0.007	0.662 ±0.008	7.760 ±0.305	0.158 ±0.002	0.573 ±0.007	0.667 ±0.008
GARLIC (Wang et al., 2026)	0.547 ±0.021	3.278 ±0.081	0.124 ±0.001	0.412 ±0.006	0.680 ±0.014	0.613 ±0.017	9.101 ±0.134	0.163 ±0.002	0.562 ±0.007	0.668 ±0.006
KEDGN (Luo et al., 2024)	0.299 ±0.022	3.615 ±0.154	**0.108** ±0.002	0.442 ±0.019	0.396 ±0.071	0.616 ±0.062	10.692 ±0.271	**0.126** ±0.005	0.486 ±0.022	0.430 ±0.032
MIRA (Li et al., 2026)	0.658 ±0.011	3.012 ±0.050	0.127 ±0.002	0.433 ±0.009	0.800 ±0.004	0.667 ±0.005	8.001 ±0.150	0.160 ±0.004	0.528 ±0.015	0.695 ±0.005
TabICL (Qu et al., 2025)	0.751 ±0.020	3.742 ±0.072	0.142 ±0.002	0.431 ±0.010	0.686 ±0.007	0.745 ±0.017	10.753 ±0.392	0.190 ±0.003	0.542 ±0.010	0.708 ±0.013
MEWS (Subbe et al., 2001)	0.775 ±0.008	2.865 ±0.049	0.130 ±0.000	0.424 ±0.024	0.721 ±0.002	0.675 ±0.011	7.462 ±0.215	0.156 ±0.001	0.506 ±0.015	0.709 ±0.002
SOFA (Vincent et al., 1996)	0.800 ±0.006	2.937 ±0.103	0.133 ±0.001	0.453 ±0.029	0.855 ±0.004	0.733 ±0.006	7.234 ±0.286	0.155 ±0.002	0.538 ±0.014	**0.825** ±0.006
Markdown (Hegselmann et al., 2026)	0.766 ±0.014	2.776 ±0.085	0.134 ±0.001	0.477 ±0.013	0.850 ±0.018	0.687 ±0.011	**6.813** ±0.207	0.157 ±0.002	0.583 ±0.008	0.592 ±0.015
Record2Vec (Ji et al., 2026)	0.822 ±0.004	2.985 ±0.053	0.136 ±0.002	0.444 ±0.006	0.866 ±0.004	0.759 ±0.003	8.801 ±0.242	0.159 ±0.002	0.553 ±0.008	0.768 ±0.003
Template (Demirel et al., 2026)	0.749 ±0.008	**2.744** ±0.218	0.136 ±0.001	**0.497** ±0.014	0.906 ±0.002	0.737 ±0.003	7.676 ±0.124	0.156 ±0.002	**0.593** ±0.006	0.676 ±0.014
QA	0.797 ±0.003	3.252 ±0.026	0.134 ±0.001	0.442 ±0.003	0.822 ±0.005	0.734 ±0.004	8.783 ±0.138	0.161 ±0.002	0.535 ±0.007	0.741 ±0.002
Treatment	0.804 ±0.005	2.993 ±0.039	0.137 ±0.001	0.441 ±0.010	0.856 ±0.002	0.743 ±0.007	8.801 ±0.154	0.161 ±0.003	0.548 ±0.012	0.772 ±0.008
Organ-severity trajectory	0.814 ±0.006	3.287 ±0.042	0.136 ±0.001	0.447 ±0.007	0.873 ±0.006	0.719 ±0.004	10.705 ±0.153	0.164 ±0.001	0.571 ±0.010	0.791 ±0.005
PORTER	**0.828** ±0.003	2.803 ±0.031	0.141 ±0.001	0.439 ±0.007	**0.916** ±0.002	**0.766** ±0.006	8.520 ±0.126	0.158 ±0.002	0.559 ±0.007	0.784 ±0.005

Method	MIMIC → HIRID					MIMIC → PPICU				
	mort AUROC↑	los MSE↓	forecast MSE↓	lab AUROC↑	drug AUROC↑	mort AUROC↑	los MSE↓	forecast MSE↓	lab AUROC↑	drug AUROC↑
ATTACK	0.498 ±0.004	0.749 ±0.003	0.125 ±0.001	0.394 ±0.013	0.697 ±0.001	0.500 ±0.001	9.775 ±0.089	0.188 ±0.001	0.496 ±0.023	0.633 ±0.000
RANDOM	0.481 ±0.012	0.871 ±0.088	0.122 ±0.001	0.398 ±0.004	0.696 ±0.004	0.502 ±0.008	9.863 ±0.148	**0.168** ±0.004	0.477 ±0.003	0.634 ±0.001
CLMBR-T (Wornow et al., 2023)	0.778 ±0.012	0.566 ±0.035	0.125 ±0.003	0.510 ±0.012	0.784 ±0.014	0.687 ±0.009	8.117 ±0.253	0.178 ±0.003	0.561 ±0.009	0.634 ±0.017
GARLIC (Wang et al., 2026)	0.591 ±0.015	0.608 ±0.024	**0.120** ±0.002	0.489 ±0.017	0.667 ±0.014	0.593 ±0.014	9.682 ±0.162	0.183 ±0.003	0.510 ±0.011	0.681 ±0.004
KEDGN (Luo et al., 2024)	0.651 ±0.034	1.739 ±0.599	0.137 ±0.002	0.481 ±0.011	0.598 ±0.057	0.544 ±0.062	9.748 ±0.726	0.185 ±0.003	0.525 ±0.011	0.651 ±0.018
MIRA (Li et al., 2026)	0.743 ±0.007	0.507 ±0.051	0.126 ±0.002	0.476 ±0.022	0.786 ±0.008	0.590 ±0.004	8.537 ±0.178	0.184 ±0.003	0.519 ±0.007	0.679 ±0.005
TabICL (Qu et al., 2025)	0.816 ±0.015	0.643 ±0.049	0.121 ±0.002	0.505 ±0.011	0.749 ±0.017	0.716 ±0.016	10.445 ±0.266	0.176 ±0.006	0.523 ±0.005	0.698 ±0.037
MEWS (Subbe et al., 2001)	0.835 ±0.012	0.545 ±0.011	0.135 ±0.001	0.462 ±0.010	0.732 ±0.005	0.717 ±0.007	8.318 ±0.118	0.189 ±0.001	0.520 ±0.003	0.704 ±0.003
SOFA (Vincent et al., 1996)	0.837 ±0.013	0.578 ±0.011	0.131 ±0.001	0.490 ±0.005	0.873 ±0.005	0.758 ±0.006	8.361 ±0.125	0.189 ±0.001	0.547 ±0.002	**0.841** ±0.001
Markdown (Hegselmann et al., 2026)	0.870 ±0.009	**0.375** ±0.005	0.138 ±0.002	0.536 ±0.013	0.845 ±0.005	0.674 ±0.013	6.804 ±0.138	0.198 ±0.002	0.557 ±0.005	0.651 ±0.019
Record2Vec (Ji et al., 2026)	0.892 ±0.004	0.523 ±0.020	0.137 ±0.002	0.508 ±0.008	0.852 ±0.002	0.776 ±0.003	8.425 ±0.225	0.191 ±0.003	0.530 ±0.009	0.760 ±0.002
Template (Demirel et al., 2026)	0.826 ±0.031	0.440 ±0.029	0.138 ±0.001	**0.554** ±0.011	0.857 ±0.008	0.709 ±0.003	**6.425** ±0.191	0.195 ±0.003	**0.569** ±0.004	0.727 ±0.016
QA	0.879 ±0.003	0.608 ±0.016	0.132 ±0.002	0.474 ±0.005	0.800 ±0.007	0.752 ±0.004	9.059 ±0.135	0.190 ±0.002	0.526 ±0.003	0.726 ±0.003
Treatment	0.879 ±0.003	0.616 ±0.015	0.133 ±0.001	0.475 ±0.009	0.855 ±0.005	0.768 ±0.003	8.929 ±0.132	0.190 ±0.002	0.536 ±0.008	0.794 ±0.004
Organ-severity trajectory	0.873 ±0.003	0.669 ±0.043	0.140 ±0.002	0.492 ±0.047	0.883 ±0.003	0.759 ±0.003	9.170 ±0.305	0.194 ±0.003	0.530 ±0.046	0.786 ±0.002
PORTER	**0.902** ±0.003	0.485 ±0.023	0.136 ±0.002	0.507 ±0.008	**0.917** ±0.002	**0.793** ±0.003	8.451 ±0.336	0.189 ±0.003	0.538 ±0.005	0.792 ±0.001

Method	PPICU → HIRID					PPICU → MIMIC				
	mort AUROC↑	los MSE↓	forecast MSE↓	lab AUROC↑	drug AUROC↑	mort AUROC↑	los MSE↓	forecast MSE↓	lab AUROC↑	drug AUROC↑
ATTACK	0.501 ±0.000	0.368 ±0.018	0.105 ±0.001	0.514 ±0.003	0.698 ±0.001	0.500 ±0.000	0.675 ±0.014	0.114 ±0.001	0.425 ±0.010	0.661 ±0.002
RANDOM	0.511 ±0.016	0.402 ±0.010	0.111 ±0.003	0.505 ±0.004	0.699 ±0.005	0.510 ±0.009	0.706 ±0.010	0.119 ±0.001	0.430 ±0.006	0.658 ±0.002
CLMBR-T (Wornow et al., 2023)	0.696 ±0.052	**0.100** ±0.011	0.107 ±0.001	**0.625** ±0.006	0.738 ±0.007	0.660 ±0.016	0.812 ±0.200	0.113 ±0.002	0.507 ±0.005	0.677 ±0.023
GARLIC (Wang et al., 2026)	0.309 ±0.020	0.311 ±0.088	0.101 ±0.001	0.594 ±0.012	0.448 ±0.031	0.508 ±0.016	0.490 ±0.029	0.112 ±0.001	0.492 ±0.010	0.653 ±0.017
KEDGN (Luo et al., 2024)	0.382 ±0.071	0.228 ±0.123	**0.090** ±0.001	0.547 ±0.027	0.422 ±0.073	0.626 ±0.050	0.678 ±0.061	**0.101** ±0.001	0.470 ±0.020	0.578 ±0.092
MIRA (Li et al., 2026)	0.782 ±0.014	0.103 ±0.011	0.107 ±0.000	0.582 ±0.037	0.792 ±0.005	0.738 ±0.006	0.478 ±0.003	0.116 ±0.001	0.464 ±0.025	0.815 ±0.003
TabICL (Qu et al., 2025)	0.644 ±0.043	1.924 ±0.840	0.098 ±0.001	0.524 ±0.011	0.678 ±0.062	0.620 ±0.019	0.881 ±0.133	0.123 ±0.001	0.515 ±0.011	0.649 ±0.056
MEWS (Subbe et al., 2001)	0.815 ±0.017	0.109 ±0.008	0.109 ±0.001	0.549 ±0.007	0.752 ±0.011	0.802 ±0.003	0.460 ±0.011	0.120 ±0.001	0.486 ±0.003	0.750 ±0.006
SOFA (Vincent et al., 1996)	0.813 ±0.006	0.130 ±0.007	0.106 ±0.001	0.619 ±0.002	**0.859** ±0.005	0.823 ±0.004	0.461 ±0.020	0.120 ±0.001	**0.521** ±0.002	**0.838** ±0.002
Markdown (Hegselmann et al., 2026)	**0.903** ±0.001	0.127 ±0.029	0.104 ±0.001	0.602 ±0.008	0.744 ±0.011	0.819 ±0.006	0.445 ±0.011	0.117 ±0.001	0.489 ±0.001	0.820 ±0.008
Record2Vec (Ji et al., 2026)	0.884 ±0.004	0.326 ±0.065	0.108 ±0.001	0.608 ±0.004	0.806 ±0.004	0.849 ±0.003	0.437 ±0.021	0.121 ±0.001	0.513 ±0.003	0.808 ±0.007
Template (Demirel et al., 2026)	0.888 ±0.006	0.106 ±0.018	0.105 ±0.001	0.621 ±0.005	0.670 ±0.043	0.820 ±0.007	0.427 ±0.012	0.122 ±0.002	0.510 ±0.004	0.809 ±0.010
QA	0.863 ±0.006	0.250 ±0.026	0.109 ±0.001	0.585 ±0.005	0.801 ±0.002	0.815 ±0.003	0.477 ±0.014	0.117 ±0.001	0.504 ±0.006	0.791 ±0.009
Treatment	0.847 ±0.005	0.409 ±0.039	0.109 ±0.001	0.585 ±0.005	0.823 ±0.003	0.812 ±0.002	0.431 ±0.017	0.119 ±0.001	0.501 ±0.003	0.819 ±0.007
Organ-severity trajectory	0.863 ±0.005	0.634 ±0.190	0.112 ±0.001	0.595 ±0.008	0.856 ±0.006	0.843 ±0.002	0.483 ±0.025	0.121 ±0.001	0.512 ±0.008	0.827 ±0.003
PORTER	0.897 ±0.005	0.407 ±0.119	0.108 ±0.001	0.612 ±0.005	0.808 ±0.006	**0.860** ±0.003	**0.415** ±0.013	0.123 ±0.001	0.512 ±0.004	0.822 ±0.007

All six directions			
Mean rank↓	Cumulative occurrences↑		
	Top-1	Top-2	Top-3
15 (12.6)	0	0	1
16 (13.1)	1	1	2
6 (7.3)	2	4	9
12 (10.3)	1	3	4
14 (11.5)	4	4	4
11 (8.9)	0	1	1
13 (10.8)	0	4	5
10 (8.6)	0	0	1
1 (5.2)	5	7	9
5 (6.4)	3	7	9
4 (6.2)	0	5	6
3 (5.9)	6	**10**	13
9 (8.3)	0	0	0
7 (7.6)	0	1	5
8 (8.0)	0	4	7
2 (5.4)	**8**	9	**14**

Time to admission. We regress the number of days from prediction to the earliest inpatient event coded as `Visit/IP` in `visit_occurrence` or `Visit/ERIP` in `visit_detail` more than 7 and at most 365 days later. Patients without an admission in this window receive no label; admissions during the first 7 days are ignored. We deterministically retain at most 2,000 labeled patients.

Next creatinine value. We regress the numeric value of the earliest creatinine measurement coded as LOINC 2160-0 or a descendant code more than 7 and at most 365 days after prediction. Patients without a numeric measurement in this window receive no label; measurements during the first 7 days are ignored. We deterministically retain at most 2,000 labeled patients.

For all tasks, each eligible patient has one seeded, task-independent prediction time: we sample uniformly from record days whose latest timestamp is at least 365 days after the first clinical observation and at least 365 days before the last. To study context length, we construct representations from the 30, 180, and 365 days preceding prediction, as well as from all earlier records.

Decontamination. Some patients appear in both EHRSHOT and INSPECT. We remove them from INSPECT before generating labels, so that no patient is in both the source and the target cohort.

ZERO-SHOT TRANSFER

Figure 6 plots the zero-shot results for every task and context window. This section gives the values and the rankings behind it. The grid behind the figure runs each representation with the MLP head through two transfer directions, five tasks and four context windows. A run is one transfer direction

Table 11: Zero-shot transfer, all six directions, the Muse arm at the MLP head. The directions are carried over three blocks of two. Cells are the mean over the five cross-validation folds with one standard deviation. The columns to the right of the first block summarize all six directions. *Mean rank* is each method's position among the sixteen by its average rank over the thirty columns, with that average in parentheses, and *Cumulative occurrences* counts the columns, out of thirty, in which a method places in the top-1, 2, 3. Groups, metrics and marking as in Table 1.

Method	HIRID → MIMIC					HIRID → PPICU					All six directions			
	mort AUROC↑	los MSE↓	forecast MSE↓	lab AUROC↑	drug AUROC↑	mort AUROC↑	los MSE↓	forecast MSE↓	lab AUROC↑	drug AUROC↑	Mean rank↓	Cumulative occurrences ↑ Top-1	Top-2	Top-3
ATTACK	0.500 ±0.001	3.500 ±0.015	0.127 ±0.000	0.374 ±0.011	0.661 ±0.001	0.501 ±0.001	9.844 ±0.089	0.163 ±0.002	0.488 ±0.006	0.634 ±0.000	15 (12.6)	0	0	1
RANDOM	0.509 ±0.007	3.603 ±0.111	0.135 ±0.001	0.377 ±0.003	0.664 ±0.003	0.498 ±0.009	10.003 ±0.301	0.163 ±0.001	0.495 ±0.008	0.632 ±0.002	16 (13.3)	1	1	2
CLMBR-T (Wornow et al., 2023)	0.756 ±0.004	2.971 ±0.091	0.131 ±0.002	0.481 ±0.014	0.773 ±0.007	0.662 ±0.008	7.760 ±0.305	0.158 ±0.002	0.573 ±0.007	0.667 ±0.008	7 (7.4)	2	4	7
GARLIC (Wang et al., 2026)	0.547 ±0.021	3.278 ±0.081	0.124 ±0.001	0.412 ±0.006	0.680 ±0.014	0.613 ±0.017	9.101 ±0.134	0.163 ±0.002	0.562 ±0.007	0.668 ±0.006	12 (10.6)	1	3	4
KEDGN (Luo et al., 2024)	0.299 ±0.022	3.615 ±0.154	**0.108** ±0.002	0.442 ±0.019	0.396 ±0.071	0.616 ±0.062	10.692 ±0.271	**0.126** ±0.005	0.486 ±0.022	0.430 ±0.032	14 (11.6)	4	4	4
MIRA (Li et al., 2026)	0.658 ±0.011	3.012 ±0.050	0.127 ±0.002	0.433 ±0.009	0.800 ±0.004	0.667 ±0.005	8.001 ±0.150	0.160 ±0.004	0.528 ±0.015	0.695 ±0.005	11 (8.8)	0	1	2
TabICL (Qu et al., 2025)	0.751 ±0.020	3.742 ±0.072	0.142 ±0.002	0.431 ±0.010	0.686 ±0.007	0.745 ±0.017	10.753 ±0.392	0.190 ±0.003	0.542 ±0.010	0.708 ±0.013	13 (10.8)	0	4	4
MEWS (Subbe et al., 2001)	0.775 ±0.008	2.865 ±0.049	0.130 ±0.000	0.424 ±0.024	0.721 ±0.002	0.675 ±0.011	7.462 ±0.215	0.156 ±0.001	0.506 ±0.015	0.709 ±0.002	10 (8.8)	0	0	2
SOFA (Vincent et al., 1996)	0.800 ±0.006	2.937 ±0.103	0.133 ±0.001	0.453 ±0.029	0.855 ±0.004	0.733 ±0.006	7.234 ±0.286	0.155 ±0.002	0.538 ±0.014	**0.825** ±0.006	**1 (5.3)**	5	9	10
Markdown (Hegselmann et al., 2026)	0.766 ±0.014	2.776 ±0.085	0.134 ±0.001	0.477 ±0.013	0.850 ±0.018	0.687 ±0.011	**6.813** ±0.207	0.157 ±0.002	0.583 ±0.008	0.592 ±0.015	4 (6.2)	3	8	11
Record2Vec (Ji et al., 2026)	0.796 ±0.010	2.982 ±0.025	0.134 ±0.001	0.446 ±0.006	0.841 ±0.004	**0.763** ±0.005	8.361 ±0.302	0.158 ±0.002	0.568 ±0.007	0.748 ±0.004	6 (6.4)	1	1	6
Template (Demirel et al., 2026)	0.749 ±0.008	**2.744** ±0.218	0.136 ±0.001	**0.497** ±0.014	0.906 ±0.002	0.737 ±0.003	7.676 ±0.124	0.156 ±0.002	**0.593** ±0.006	0.676 ±0.014	3 (5.8)	**6**	**11**	**14**
QA	0.791 ±0.005	3.172 ±0.057	0.131 ±0.001	0.437 ±0.003	0.829 ±0.003	0.745 ±0.005	8.409 ±0.146	0.159 ±0.001	0.534 ±0.005	0.721 ±0.003	9 (8.5)	0	0	1
Treatment	0.785 ±0.008	3.126 ±0.034	0.134 ±0.001	0.444 ±0.003	0.862 ±0.005	0.731 ±0.009	7.788 ±0.167	0.160 ±0.001	0.552 ±0.007	0.774 ±0.005	5 (6.3)	0	2	4
Organ-severity trajectory	**0.809** ±0.006	3.186 ±0.053	0.134 ±0.001	0.442 ±0.006	0.864 ±0.007	0.730 ±0.005	10.193 ±0.110	0.164 ±0.001	0.572 ±0.008	0.773 ±0.009	8 (8.2)	2	5	8
PORTER	0.799 ±0.007	2.887 ±0.022	0.139 ±0.002	0.441 ±0.007	**0.913** ±0.002	0.762 ±0.005	8.276 ±0.192	0.161 ±0.003	0.572 ±0.008	0.754 ±0.002	2 (5.4)	5	7	10

Method	MIMIC → HIRID					MIMIC → PPICU				
	mort AUROC↑	los MSE↓	forecast MSE↓	lab AUROC↑	drug AUROC↑	mort AUROC↑	los MSE↓	forecast MSE↓	lab AUROC↑	drug AUROC↑
ATTACK	0.498 ±0.003	0.748 ±0.003	0.125 ±0.000	0.395 ±0.014	0.697 ±0.001	0.500 ±0.001	9.801 ±0.092	0.188 ±0.001	0.494 ±0.023	0.634 ±0.000
RANDOM	0.481 ±0.012	0.871 ±0.088	0.122 ±0.001	0.398 ±0.004	0.696 ±0.004	0.502 ±0.008	9.863 ±0.148	**0.168** ±0.004	0.477 ±0.003	0.634 ±0.001
CLMBR-T (Wornow et al., 2023)	0.778 ±0.012	0.566 ±0.035	0.125 ±0.003	0.510 ±0.012	0.784 ±0.014	0.687 ±0.009	8.117 ±0.253	0.178 ±0.003	0.561 ±0.009	0.634 ±0.017
GARLIC (Wang et al., 2026)	0.591 ±0.015	0.608 ±0.024	**0.120** ±0.002	0.489 ±0.017	0.667 ±0.014	0.593 ±0.014	9.682 ±0.162	0.183 ±0.003	0.510 ±0.011	0.681 ±0.004
KEDGN (Luo et al., 2024)	0.651 ±0.034	1.739 ±0.599	0.137 ±0.002	0.481 ±0.011	0.598 ±0.057	0.544 ±0.062	9.748 ±0.726	0.185 ±0.003	0.525 ±0.011	0.651 ±0.018
MIRA (Li et al., 2026)	0.743 ±0.007	0.507 ±0.051	0.126 ±0.002	0.476 ±0.022	0.786 ±0.008	0.590 ±0.004	8.537 ±0.178	0.184 ±0.003	0.519 ±0.007	0.679 ±0.005
TabICL (Qu et al., 2025)	0.816 ±0.015	0.643 ±0.049	0.121 ±0.002	0.505 ±0.011	0.749 ±0.017	0.716 ±0.016	10.445 ±0.266	0.176 ±0.006	0.523 ±0.005	0.698 ±0.037
MEWS (Subbe et al., 2001)	0.835 ±0.012	0.545 ±0.011	0.135 ±0.001	0.462 ±0.010	0.732 ±0.005	0.717 ±0.007	8.318 ±0.118	0.189 ±0.001	0.520 ±0.003	0.704 ±0.003
SOFA (Vincent et al., 1996)	0.837 ±0.013	0.578 ±0.011	0.131 ±0.001	0.490 ±0.005	0.873 ±0.005	0.758 ±0.006	8.361 ±0.125	0.189 ±0.001	0.547 ±0.002	**0.841** ±0.001
Markdown (Hegselmann et al., 2026)	0.870 ±0.009	**0.375** ±0.005	0.138 ±0.002	0.536 ±0.013	0.845 ±0.005	0.674 ±0.013	6.804 ±0.138	0.198 ±0.002	0.557 ±0.005	0.651 ±0.019
Record2Vec (Ji et al., 2026)	0.860 ±0.012	0.451 ±0.010	0.132 ±0.001	0.530 ±0.010	0.856 ±0.003	0.758 ±0.006	9.007 ±0.252	0.190 ±0.002	0.551 ±0.007	0.749 ±0.002
Template (Demirel et al., 2026)	0.826 ±0.031	0.440 ±0.029	0.138 ±0.001	**0.554** ±0.011	0.857 ±0.008	0.709 ±0.003	**6.425** ±0.191	0.195 ±0.003	**0.569** ±0.004	0.727 ±0.016
QA	0.839 ±0.017	0.588 ±0.020	0.131 ±0.001	0.479 ±0.009	0.791 ±0.006	0.738 ±0.005	8.832 ±0.239	0.190 ±0.001	0.523 ±0.006	0.723 ±0.004
Treatment	0.859 ±0.004	0.475 ±0.005	0.131 ±0.002	0.491 ±0.006	0.847 ±0.014	0.764 ±0.006	8.523 ±0.160	0.186 ±0.004	0.530 ±0.003	0.781 ±0.002
Organ-severity trajectory	0.856 ±0.008	0.817 ±0.181	0.141 ±0.001	0.494 ±0.041	0.866 ±0.003	0.769 ±0.001	9.169 ±0.476	0.194 ±0.002	0.546 ±0.016	0.783 ±0.004
PORTER	**0.882** ±0.007	0.456 ±0.042	0.134 ±0.002	0.528 ±0.010	**0.908** ±0.002	**0.795** ±0.009	8.045 ±0.227	0.188 ±0.004	0.552 ±0.007	0.776 ±0.003

Method	PPICU → HIRID					PPICU → MIMIC				
	mort AUROC↑	los MSE↓	forecast MSE↓	lab AUROC↑	drug AUROC↑	mort AUROC↑	los MSE↓	forecast MSE↓	lab AUROC↑	drug AUROC↑
ATTACK	0.501 ±0.001	0.366 ±0.024	0.105 ±0.001	0.513 ±0.006	0.700 ±0.001	0.501 ±0.000	0.674 ±0.018	0.114 ±0.001	0.421 ±0.009	0.661 ±0.002
RANDOM	0.511 ±0.016	0.402 ±0.010	0.111 ±0.003	0.505 ±0.004	0.699 ±0.005	0.510 ±0.009	0.706 ±0.010	0.119 ±0.001	0.430 ±0.006	0.658 ±0.002
CLMBR-T (Wornow et al., 2023)	0.696 ±0.052	**0.100** ±0.011	0.107 ±0.001	**0.625** ±0.006	0.738 ±0.007	0.660 ±0.016	0.812 ±0.200	0.113 ±0.002	0.507 ±0.005	0.677 ±0.023
GARLIC (Wang et al., 2026)	0.309 ±0.020	0.311 ±0.088	0.101 ±0.001	0.594 ±0.012	0.448 ±0.031	0.508 ±0.016	0.490 ±0.029	0.112 ±0.001	0.492 ±0.010	0.653 ±0.017
KEDGN (Luo et al., 2024)	0.382 ±0.071	0.228 ±0.123	**0.090** ±0.001	0.547 ±0.027	0.422 ±0.073	0.626 ±0.050	0.678 ±0.061	**0.101** ±0.001	0.470 ±0.020	0.578 ±0.092
MIRA (Li et al., 2026)	0.782 ±0.014	0.103 ±0.011	0.107 ±0.000	0.582 ±0.037	0.792 ±0.005	0.738 ±0.006	0.478 ±0.003	0.116 ±0.001	0.464 ±0.025	0.815 ±0.003
TabICL (Qu et al., 2025)	0.644 ±0.043	1.924 ±0.840	0.098 ±0.001	0.524 ±0.011	0.678 ±0.062	0.620 ±0.019	0.881 ±0.133	0.123 ±0.001	0.515 ±0.011	0.649 ±0.056
MEWS (Subbe et al., 2001)	0.815 ±0.017	0.109 ±0.008	0.109 ±0.001	0.549 ±0.007	0.752 ±0.011	0.802 ±0.003	0.460 ±0.011	0.120 ±0.001	0.486 ±0.003	0.750 ±0.006
SOFA (Vincent et al., 1996)	0.813 ±0.006	0.130 ±0.007	0.106 ±0.001	0.619 ±0.002	**0.859** ±0.005	0.823 ±0.004	0.461 ±0.020	0.120 ±0.001	**0.521** ±0.002	**0.838** ±0.002
Markdown (Hegselmann et al., 2026)	**0.903** ±0.001	0.127 ±0.029	0.104 ±0.001	0.602 ±0.008	0.744 ±0.011	0.819 ±0.006	0.445 ±0.011	0.117 ±0.001	0.489 ±0.001	0.820 ±0.008
Record2Vec (Ji et al., 2026)	0.878 ±0.004	0.270 ±0.039	0.109 ±0.001	0.619 ±0.008	0.781 ±0.005	0.839 ±0.002	0.444 ±0.017	0.121 ±0.001	0.509 ±0.003	0.770 ±0.005
Template (Demirel et al., 2026)	0.888 ±0.006	0.106 ±0.018	0.105 ±0.001	0.621 ±0.005	0.670 ±0.043	0.820 ±0.007	0.427 ±0.012	0.122 ±0.002	0.510 ±0.004	0.809 ±0.010
QA	0.842 ±0.007	0.242 ±0.012	0.109 ±0.001	0.589 ±0.003	0.780 ±0.006	0.817 ±0.002	0.487 ±0.019	0.118 ±0.001	0.495 ±0.004	0.764 ±0.005
Treatment	0.836 ±0.014	0.213 ±0.041	0.107 ±0.001	0.599 ±0.001	0.791 ±0.006	0.822 ±0.003	0.457 ±0.017	0.118 ±0.001	0.504 ±0.002	0.827 ±0.005
Organ-severity trajectory	0.871 ±0.007	0.662 ±0.118	0.111 ±0.001	0.596 ±0.007	0.824 ±0.004	**0.843** ±0.002	0.526 ±0.009	0.121 ±0.001	0.505 ±0.003	0.803 ±0.004
PORTER	0.888 ±0.004	0.349 ±0.093	0.107 ±0.001	0.613 ±0.004	0.785 ±0.006	0.841 ±0.002	**0.421** ±0.020	0.120 ±0.000	0.509 ±0.004	0.806 ±0.004

and one task, so a context window has ten runs of a representation. Each run is five-fold cross-validated on the source cohort and evaluated on the whole target cohort. RANDOM and Demographics do not read the context window, so they are run once and their results are reused at every window.

Table 29 takes the ranking over the ten runs of each context window and over all forty runs, and Table 30 takes it one task at a time. PORTER has the best mean rank over all forty runs and wins 29 of them outright. Direct embedding has the second-best mean rank and wins ten. RANDOM wins the remaining run, which is time to admission from EHRSHOT to INSPECT at the last 365 days. The best MSE on time to admission from EHRSHOT to INSPECT is within 0.05 of RANDOM at every context window. None of the other five representations predicts this task in this direction much better than RANDOM, which carries no information about the patient. PORTER is first in both directions on new malignancy and next creatinine value at every context window. Direct embedding is first in both directions on new hyperlipidemia at the three shorter context windows. PORTER is first in both directions on new hypertension at the same three windows. The worst cell of PORTER in Table 30 is new hyperlipidemia at the last 30 days, with a mean rank of 4.5. New hyperlipidemia and new hypertension split by direction with the full record. PORTER is first on both from INSPECT to EHRSHOT, and direct embedding is first on both from EHRSHOT to INSPECT. On new hypertension from EHRSHOT to INSPECT the two are level at 0.681 after rounding. The first place on time to admission changes with the context window. PORTER is first in both directions at the last 30 and 180 days. Its mean rank is 3.0 at the last 365 days, and PORTER and direct embedding each win one direction with the full record.

Table 12: Zero-shot transfer, all six directions, the MedGemma arm at the LSTM head. The directions are carried over three blocks of two. Cells are the mean over the five cross-validation folds with one standard deviation. The columns to the right of the first block summarize all six directions. *Mean rank* is each method's position among the sixteen by its average rank over the thirty columns, with that average in parentheses, and *Cumulative occurrences* counts the columns, out of thirty, in which a method places in the top-1, 2, 3. Groups, metrics and marking as in Table 1.

Method	HIRID → MIMIC					HIRID → PPICU					All six directions			
	mort AUROC↑	los MSE↓	forecast MSE↓	lab AUROC↑	drug AUROC↑	mort AUROC↑	los MSE↓	forecast MSE↓	lab AUROC↑	drug AUROC↑	Mean rank↓	Cumulative occurrences ↑ Top-1	Top-2	Top-3
ATTACK	0.500 ±0.000	3.483 ±0.006	0.127 ±0.000	0.373 ±0.034	0.661 ±0.003	0.500 ±0.000	9.745 ±0.039	0.161 ±0.001	0.475 ±0.027	0.634 ±0.000	15 (12.4)	0	0	1
RANDOM	0.507 ±0.010	3.588 ±0.079	0.135 ±0.002	0.394 ±0.010	0.663 ±0.003	0.493 ±0.010	10.094 ±0.213	0.162 ±0.001	0.491 ±0.018	0.634 ±0.003	16 (12.9)	0	1	2
CLMBR-T (Wornow et al., 2023)	0.753 ±0.011	2.934 ±0.043	0.132 ±0.001	**0.468** ±0.020	0.772 ±0.006	0.662 ±0.008	7.758 ±0.485	0.156 ±0.002	0.574 ±0.009	0.663 ±0.007	5 (7.2)	4	5	7
GARLIC (Wang et al., 2026)	0.550 ±0.028	3.284 ±0.043	0.127 ±0.001	0.401 ±0.006	0.690 ±0.006	0.623 ±0.017	9.025 ±0.213	0.162 ±0.001	0.531 ±0.024	0.673 ±0.012	12 (9.6)	1	2	4
KEDGN (Luo et al., 2024)	0.304 ±0.024	3.663 ±0.166	**0.117** ±0.003	0.430 ±0.024	0.329 ±0.024	0.629 ±0.031	11.271 ±0.147	**0.138** ±0.003	0.446 ±0.020	0.430 ±0.017	14 (10.6)	5	5	5
MIRA (Li et al., 2026)	0.639 ±0.014	3.003 ±0.045	0.128 ±0.001	0.369 ±0.037	0.797 ±0.003	0.663 ±0.005	7.919 ±0.316	0.164 ±0.001	0.500 ±0.014	0.693 ±0.004	10 (8.6)	0	1	2
TabICL (Qu et al., 2025)	0.751 ±0.017	3.825 ±0.030	0.140 ±0.002	0.424 ±0.009	0.695 ±0.010	0.743 ±0.024	10.934 ±0.207	0.175 ±0.007	0.541 ±0.025	0.716 ±0.011	11 (9.4)	0	4	5
MEWS (Subbe et al., 2001)	0.784 ±0.003	2.856 ±0.052	0.131 ±0.001	0.419 ±0.011	0.729 ±0.003	0.679 ±0.006	7.596 ±0.268	0.156 ±0.001	0.513 ±0.010	0.711 ±0.002	6 (7.6)	0	1	4
SOFA (Vincent et al., 1996)	**0.807** ±0.008	2.928 ±0.089	0.134 ±0.001	0.460 ±0.006	0.857 ±0.003	0.738 ±0.015	7.312 ±0.216	0.156 ±0.001	0.557 ±0.006	**0.832** ±0.004	**1 (4.6)**	**8**	**10**	**14**
Markdown (Hegselmann et al., 2026)	0.766 ±0.005	2.776 ±0.072	0.136 ±0.001	0.455 ±0.006	0.853 ±0.013	0.704 ±0.008	**6.694** ±0.156	0.155 ±0.001	0.578 ±0.005	0.613 ±0.033	4 (6.8)	4	9	9
Record2Vec (Ji et al., 2026)	0.751 ±0.006	3.009 ±0.042	0.132 ±0.001	0.415 ±0.029	0.818 ±0.009	0.725 ±0.003	8.561 ±0.193	0.155 ±0.002	0.540 ±0.016	0.724 ±0.006	7 (7.6)	0	0	3
Template (Demirel et al., 2026)	0.747 ±0.008	**2.675** ±0.162	0.135 ±0.001	0.467 ±0.013	**0.895** ±0.004	0.734 ±0.002	7.556 ±0.278	0.156 ±0.001	**0.587** ±0.007	0.658 ±0.009	3 (6.7)	4	**10**	12
QA	0.734 ±0.010	3.223 ±0.057	0.131 ±0.002	0.427 ±0.006	0.767 ±0.004	0.668 ±0.007	8.704 ±0.243	0.158 ±0.001	0.519 ±0.003	0.695 ±0.002	13 (9.7)	0	0	0
Treatment	0.737 ±0.005	3.031 ±0.032	0.132 ±0.001	0.423 ±0.034	0.825 ±0.005	0.707 ±0.006	7.424 ±0.190	0.157 ±0.001	0.534 ±0.013	0.738 ±0.003	8 (7.7)	1	1	2
Organ-severity trajectory	0.765 ±0.011	3.430 ±0.090	0.131 ±0.002	0.375 ±0.024	0.860 ±0.003	0.721 ±0.010	10.148 ±0.361	0.160 ±0.001	0.480 ±0.031	0.747 ±0.004	9 (8.3)	1	3	6
PORTER	0.791 ±0.002	2.825 ±0.048	0.133 ±0.002	0.426 ±0.007	0.890 ±0.004	**0.758** ±0.002	9.069 ±0.138	0.157 ±0.002	0.553 ±0.009	0.763 ±0.002	2 (6.2)	2	8	**14**

Method	MIMIC → HIRID					MIMIC → PPICU				
	mort AUROC↑	los MSE↓	forecast MSE↓	lab AUROC↑	drug AUROC↑	mort AUROC↑	los MSE↓	forecast MSE↓	lab AUROC↑	drug AUROC↑
ATTACK	0.497 ±0.005	0.760 ±0.003	0.125 ±0.001	0.418 ±0.023	0.697 ±0.001	0.500 ±0.001	9.587 ±0.036	0.186 ±0.001	0.475 ±0.023	0.634 ±0.000
RANDOM	0.499 ±0.015	0.841 ±0.026	0.125 ±0.001	0.392 ±0.005	0.696 ±0.002	0.499 ±0.010	10.159 ±0.108	0.180 ±0.001	0.476 ±0.004	0.632 ±0.001
CLMBR-T (Wornow et al., 2023)	0.787 ±0.012	0.527 ±0.043	0.128 ±0.001	0.449 ±0.012	0.801 ±0.003	0.682 ±0.017	8.079 ±0.314	0.186 ±0.002	0.530 ±0.007	0.654 ±0.005
GARLIC (Wang et al., 2026)	0.598 ±0.007	0.651 ±0.073	**0.121** ±0.002	0.442 ±0.018	0.702 ±0.026	0.596 ±0.013	9.710 ±0.204	0.184 ±0.001	0.490 ±0.007	0.682 ±0.005
KEDGN (Luo et al., 2024)	0.605 ±0.048	0.903 ±0.172	0.130 ±0.002	0.440 ±0.021	0.605 ±0.029	0.511 ±0.050	9.771 ±0.722	**0.179** ±0.001	0.487 ±0.024	0.675 ±0.007
MIRA (Li et al., 2026)	0.749 ±0.010	0.488 ±0.050	0.127 ±0.001	0.394 ±0.020	0.787 ±0.003	0.593 ±0.002	8.663 ±0.303	0.186 ±0.001	0.486 ±0.027	0.674 ±0.004
TabICL (Qu et al., 2025)	0.823 ±0.018	0.618 ±0.070	0.124 ±0.003	0.458 ±0.016	0.768 ±0.010	0.719 ±0.017	10.306 ±0.244	0.183 ±0.002	0.514 ±0.006	0.730 ±0.008
MEWS (Subbe et al., 2001)	0.853 ±0.004	0.526 ±0.008	0.135 ±0.001	0.453 ±0.014	0.760 ±0.005	0.719 ±0.005	8.377 ±0.246	0.190 ±0.001	0.520 ±0.010	0.712 ±0.003
SOFA (Vincent et al., 1996)	0.854 ±0.005	0.561 ±0.007	0.132 ±0.001	**0.473** ±0.007	0.875 ±0.002	**0.763** ±0.009	8.309 ±0.165	0.188 ±0.001	**0.540** ±0.012	**0.846** ±0.001
Markdown (Hegselmann et al., 2026)	**0.867** ±0.012	**0.389** ±0.008	0.139 ±0.001	0.438 ±0.040	0.822 ±0.014	0.672 ±0.007	6.665 ±0.268	0.194 ±0.002	0.486 ±0.040	0.624 ±0.016
Record2Vec (Ji et al., 2026)	0.840 ±0.006	0.507 ±0.004	0.132 ±0.002	0.419 ±0.023	0.830 ±0.005	0.737 ±0.006	8.908 ±0.195	0.189 ±0.001	0.475 ±0.022	0.734 ±0.002
Template (Demirel et al., 2026)	0.799 ±0.027	0.451 ±0.025	0.140 ±0.003	0.412 ±0.024	0.819 ±0.027	0.704 ±0.003	**6.389** ±0.097	0.195 ±0.002	0.476 ±0.024	0.711 ±0.014
QA	0.807 ±0.010	0.598 ±0.023	0.130 ±0.002	0.420 ±0.023	0.781 ±0.009	0.682 ±0.004	9.019 ±0.156	0.188 ±0.002	0.474 ±0.023	0.674 ±0.006
Treatment	0.818 ±0.006	0.574 ±0.011	0.132 ±0.001	0.420 ±0.025	0.848 ±0.003	0.732 ±0.006	8.811 ±0.225	0.188 ±0.003	0.475 ±0.025	0.732 ±0.002
Organ-severity trajectory	0.822 ±0.005	0.644 ±0.032	0.131 ±0.001	0.412 ±0.028	0.877 ±0.002	0.730 ±0.007	8.445 ±0.257	0.188 ±0.001	0.479 ±0.018	0.749 ±0.003
PORTER	0.859 ±0.005	0.503 ±0.010	0.134 ±0.001	0.416 ±0.025	**0.892** ±0.002	0.761 ±0.012	8.392 ±0.178	0.191 ±0.002	0.472 ±0.025	0.765 ±0.002

Method	PPICU → HIRID					PPICU → MIMIC				
	mort AUROC↑	los MSE↓	forecast MSE↓	lab AUROC↑	drug AUROC↑	mort AUROC↑	los MSE↓	forecast MSE↓	lab AUROC↑	drug AUROC↑
ATTACK	0.496 ±0.004	0.416 ±0.011	0.105 ±0.001	0.499 ±0.013	0.696 ±0.002	0.495 ±0.004	0.713 ±0.009	0.115 ±0.001	0.416 ±0.026	0.661 ±0.002
RANDOM	0.515 ±0.012	0.449 ±0.012	0.107 ±0.001	0.487 ±0.016	0.698 ±0.002	0.507 ±0.008	0.743 ±0.011	0.116 ±0.001	0.425 ±0.031	0.660 ±0.003
CLMBR-T (Wornow et al., 2023)	0.665 ±0.040	**0.092** ±0.008	0.108 ±0.001	**0.616** ±0.010	0.748 ±0.013	0.675 ±0.027	0.647 ±0.061	0.116 ±0.001	**0.502** ±0.008	0.679 ±0.009
GARLIC (Wang et al., 2026)	0.314 ±0.010	0.257 ±0.054	0.101 ±0.001	0.572 ±0.032	0.485 ±0.052	0.509 ±0.024	0.494 ±0.021	0.114 ±0.001	0.472 ±0.010	0.670 ±0.026
KEDGN (Luo et al., 2024)	0.317 ±0.074	0.129 ±0.033	**0.091** ±0.001	0.530 ±0.029	0.513 ±0.064	0.658 ±0.081	0.528 ±0.012	**0.103** ±0.001	0.449 ±0.016	0.580 ±0.066
MIRA (Li et al., 2026)	0.793 ±0.017	0.109 ±0.019	0.105 ±0.001	0.534 ±0.026	0.788 ±0.007	0.741 ±0.005	0.485 ±0.015	0.114 ±0.001	0.434 ±0.021	0.812 ±0.003
TabICL (Qu et al., 2025)	0.668 ±0.049	0.863 ±0.610	0.096 ±0.002	0.509 ±0.009	0.716 ±0.038	0.694 ±0.019	0.690 ±0.117	0.120 ±0.002	0.472 ±0.034	0.660 ±0.030
MEWS (Subbe et al., 2001)	0.829 ±0.009	0.104 ±0.006	0.109 ±0.001	0.505 ±0.029	0.760 ±0.010	0.804 ±0.005	0.458 ±0.005	0.121 ±0.000	0.450 ±0.034	0.753 ±0.004
SOFA (Vincent et al., 1996)	0.807 ±0.015	0.117 ±0.006	0.108 ±0.001	0.571 ±0.050	**0.869** ±0.003	0.827 ±0.003	0.459 ±0.020	0.121 ±0.000	0.475 ±0.045	**0.837** ±0.002
Markdown (Hegselmann et al., 2026)	**0.900** ±0.003	0.138 ±0.008	0.106 ±0.001	0.548 ±0.054	0.737 ±0.029	0.816 ±0.009	0.445 ±0.011	0.122 ±0.003	0.450 ±0.039	0.830 ±0.003
Record2Vec (Ji et al., 2026)	0.835 ±0.004	0.298 ±0.025	0.108 ±0.001	0.574 ±0.032	0.741 ±0.009	0.803 ±0.003	0.464 ±0.016	0.119 ±0.001	0.473 ±0.049	0.762 ±0.008
Template (Demirel et al., 2026)	0.876 ±0.020	0.106 ±0.017	0.106 ±0.000	0.599 ±0.043	0.706 ±0.047	0.818 ±0.004	0.436 ±0.008	0.124 ±0.001	0.486 ±0.031	0.820 ±0.008
QA	0.780 ±0.013	0.269 ±0.043	0.107 ±0.001	0.501 ±0.012	0.755 ±0.003	0.758 ±0.007	0.475 ±0.012	0.117 ±0.001	0.418 ±0.028	0.747 ±0.003
Treatment	0.826 ±0.010	0.237 ±0.022	0.108 ±0.001	0.518 ±0.036	0.802 ±0.006	0.786 ±0.003	**0.425** ±0.007	0.120 ±0.000	0.435 ±0.037	0.798 ±0.008
Organ-severity trajectory	0.836 ±0.007	0.299 ±0.025	0.106 ±0.001	0.515 ±0.012	0.816 ±0.007	**0.832** ±0.003	0.506 ±0.011	0.120 ±0.000	0.418 ±0.013	0.792 ±0.015
PORTER	0.852 ±0.011	0.315 ±0.081	0.109 ±0.001	0.560 ±0.048	0.808 ±0.009	0.820 ±0.004	0.444 ±0.022	0.119 ±0.001	0.479 ±0.033	0.816 ±0.005

Results by context window. Table 32 shows the MLP results for each context window. PORTER has the best mean rank at all four context windows. PORTER is in the top three for at least eight of the ten columns at every context window. History beyond one year helps PORTER more than direct embedding of the rendered record. From the last 365 days to the full record, PORTER improves on nine of the ten columns and direct embedding improves on seven. This is consistent with the summary making use of a long record that direct embedding has to truncate. Count features and CLMBR-T rank below PORTER and direct embedding at every context window with the MLP head.

Table 13: Zero-shot transfer, all six directions, the MedGemma arm at the linear head. The directions are carried over three blocks of two. Cells are the mean over the five cross-validation folds with one standard deviation. The columns to the right of the first block summarize all six directions. *Mean rank* is each method's position among the sixteen by its average rank over the thirty columns, with that average in parentheses, and *Cumulative occurrences* counts the columns, out of thirty, in which a method places in the top-1, 2, 3. Groups, metrics and marking as in Table 1.

Method	HIRID → MIMIC					HIRID → PPICU				
	mort AUROC↑	los MSE↓	forecast MSE↓	lab AUROC↑	drug AUROC↑	mort AUROC↑	los MSE↓	forecast MSE↓	lab AUROC↑	drug AUROC↑
ATTACK	0.502 ±0.003	3.527 ±0.021	0.127 ±0.000	0.390 ±0.032	0.662 ±0.003	0.500 ±0.001	9.993 ±0.104	0.162 ±0.001	0.499 ±0.023	0.634 ±0.000
RANDOM	0.517 ±0.008	3.516 ±0.023	0.140 ±0.001	0.384 ±0.007	0.663 ±0.002	0.504 ±0.015	9.816 ±0.040	0.166 ±0.002	0.493 ±0.014	0.632 ±0.002
CLMBR-T (Wornow et al., 2023)	0.761 ±0.007	2.952 ±0.022	0.130 ±0.002	0.477 ±0.018	0.760 ±0.010	0.670 ±0.010	7.777 ±0.250	0.159 ±0.003	0.565 ±0.019	0.658 ±0.009
GARLIC (Wang et al., 2026)	0.550 ±0.041	3.158 ±0.080	0.126 ±0.002	0.399 ±0.028	0.673 ±0.009	0.621 ±0.023	8.509 ±0.175	0.163 ±0.002	0.549 ±0.011	0.650 ±0.012
KEDGN (Luo et al., 2024)	0.281 ±0.015	4.373 ±0.692	**0.110** ±0.005	0.397 ±0.011	0.371 ±0.048	0.639 ±0.027	16.610 ±1.327	**0.138** ±0.008	0.516 ±0.024	0.462 ±0.029
MIRA (Li et al., 2026)	0.658 ±0.010	2.974 ±0.030	0.128 ±0.001	0.426 ±0.031	0.806 ±0.004	0.669 ±0.005	7.781 ±0.150	0.161 ±0.002	0.499 ±0.040	0.700 ±0.003
TabICL (Qu et al., 2025)	0.768 ±0.006	3.566 ±0.108	0.146 ±0.003	0.419 ±0.009	0.679 ±0.004	**0.758** ±0.007	9.316 ±0.356	0.191 ±0.006	0.549 ±0.011	0.700 ±0.008
MEWS (Subbe et al., 2001)	0.783 ±0.005	2.853 ±0.016	0.129 ±0.001	0.424 ±0.020	0.729 ±0.002	0.679 ±0.004	7.240 ±0.034	0.155 ±0.002	0.498 ±0.013	0.709 ±0.001
SOFA (Vincent et al., 1996)	**0.804** ±0.003	2.865 ±0.011	0.133 ±0.001	0.443 ±0.025	0.855 ±0.004	0.733 ±0.007	7.045 ±0.082	0.155 ±0.002	0.533 ±0.009	**0.824** ±0.003
Markdown (Hegselmann et al., 2026)	0.769 ±0.009	**2.758** ±0.057	0.136 ±0.003	0.462 ±0.012	0.854 ±0.012	0.680 ±0.006	**6.632** ±0.181	0.159 ±0.004	0.574 ±0.007	0.571 ±0.019
Record2Vec (Ji et al., 2026)	0.760 ±0.004	2.979 ±0.023	0.132 ±0.001	0.446 ±0.006	0.830 ±0.006	0.727 ±0.002	8.233 ±0.192	0.157 ±0.001	0.546 ±0.006	0.722 ±0.006
Template (Demirel et al., 2026)	0.757 ±0.005	2.766 ±0.079	0.136 ±0.002	**0.493** ±0.021	**0.904** ±0.004	0.736 ±0.003	7.487 ±0.195	0.160 ±0.005	**0.593** ±0.005	0.673 ±0.012
QA	0.732 ±0.006	3.178 ±0.039	0.131 ±0.000	0.418 ±0.031	0.769 ±0.004	0.671 ±0.008	8.552 ±0.213	0.159 ±0.003	0.525 ±0.008	0.696 ±0.003
Treatment	0.743 ±0.008	3.033 ±0.043	0.132 ±0.001	0.420 ±0.034	0.827 ±0.003	0.704 ±0.008	7.378 ±0.070	0.157 ±0.002	0.530 ±0.006	0.734 ±0.008
Organ-severity trajectory	0.768 ±0.007	3.474 ±0.074	0.133 ±0.001	0.437 ±0.008	0.859 ±0.003	0.722 ±0.008	10.354 ±0.239	0.163 ±0.001	0.569 ±0.011	0.741 ±0.004
PORTER	0.794 ±0.003	2.799 ±0.027	0.134 ±0.001	0.423 ±0.035	0.893 ±0.001	0.756 ±0.003	8.740 ±0.314	0.157 ±0.003	0.554 ±0.007	0.770 ±0.003

Method	MIMIC → HIRID					MIMIC → PPICU				
	mort AUROC↑	los MSE↓	forecast MSE↓	lab AUROC↑	drug AUROC↑	mort AUROC↑	los MSE↓	forecast MSE↓	lab AUROC↑	drug AUROC↑
ATTACK	0.500 ±0.001	0.749 ±0.003	0.125 ±0.000	0.408 ±0.026	0.697 ±0.001	0.500 ±0.000	9.863 ±0.189	0.186 ±0.001	0.470 ±0.038	0.634 ±0.000
RANDOM	0.505 ±0.014	0.807 ±0.004	0.128 ±0.001	0.409 ±0.008	0.695 ±0.002	0.494 ±0.008	9.588 ±0.067	**0.179** ±0.004	0.473 ±0.004	0.629 ±0.001
CLMBR-T (Wornow et al., 2023)	0.787 ±0.009	0.524 ±0.032	0.131 ±0.003	0.496 ±0.013	0.792 ±0.004	0.669 ±0.019	7.692 ±0.510	0.183 ±0.002	0.546 ±0.015	0.635 ±0.010
GARLIC (Wang et al., 2026)	0.595 ±0.015	0.591 ±0.020	**0.120** ±0.002	0.457 ±0.021	0.688 ±0.017	0.595 ±0.010	9.381 ±0.199	0.182 ±0.002	0.501 ±0.019	0.685 ±0.004
KEDGN (Luo et al., 2024)	0.626 ±0.022	1.064 ±0.175	0.143 ±0.003	0.452 ±0.019	0.597 ±0.057	0.517 ±0.042	8.034 ±0.231	0.190 ±0.006	0.516 ±0.009	0.655 ±0.030
MIRA (Li et al., 2026)	0.750 ±0.011	0.517 ±0.086	0.126 ±0.002	0.439 ±0.031	0.788 ±0.004	0.599 ±0.006	8.704 ±0.295	0.186 ±0.004	0.474 ±0.032	0.676 ±0.007
TabICL (Qu et al., 2025)	0.822 ±0.017	0.646 ±0.033	0.135 ±0.003	0.501 ±0.012	0.759 ±0.011	0.692 ±0.034	8.943 ±0.548	0.206 ±0.004	0.519 ±0.002	0.716 ±0.012
MEWS (Subbe et al., 2001)	0.840 ±0.007	0.509 ±0.017	0.136 ±0.001	0.457 ±0.017	0.749 ±0.004	0.723 ±0.003	7.846 ±0.099	0.188 ±0.001	0.506 ±0.022	0.708 ±0.002
SOFA (Vincent et al., 1996)	0.848 ±0.009	0.518 ±0.010	0.131 ±0.000	0.485 ±0.002	0.880 ±0.001	**0.762** ±0.004	7.785 ±0.094	0.187 ±0.002	0.545 ±0.005	**0.843** ±0.002
Markdown (Hegselmann et al., 2026)	**0.874** ±0.007	**0.376** ±0.005	0.137 ±0.003	0.538 ±0.014	0.839 ±0.012	0.680 ±0.012	6.752 ±0.290	0.195 ±0.006	0.551 ±0.013	0.638 ±0.024
Record2Vec (Ji et al., 2026)	0.838 ±0.010	0.513 ±0.016	0.130 ±0.001	0.494 ±0.004	0.834 ±0.003	0.740 ±0.005	8.592 ±0.125	0.184 ±0.002	0.527 ±0.005	0.729 ±0.003
Template (Demirel et al., 2026)	0.833 ±0.013	0.481 ±0.013	0.138 ±0.005	**0.559** ±0.015	0.856 ±0.006	0.717 ±0.009	**6.208** ±0.112	0.199 ±0.008	**0.567** ±0.013	0.727 ±0.019
QA	0.813 ±0.007	0.567 ±0.015	0.130 ±0.001	0.451 ±0.006	0.781 ±0.006	0.681 ±0.004	8.732 ±0.182	0.188 ±0.002	0.501 ±0.006	0.671 ±0.004
Treatment	0.814 ±0.006	0.562 ±0.011	0.129 ±0.001	0.466 ±0.008	0.848 ±0.002	0.730 ±0.005	8.289 ±0.202	0.187 ±0.002	0.515 ±0.006	0.730 ±0.005
Organ-severity trajectory	0.816 ±0.004	0.637 ±0.013	0.132 ±0.001	0.443 ±0.035	0.876 ±0.004	0.723 ±0.013	8.939 ±0.308	0.193 ±0.001	0.515 ±0.033	0.752 ±0.006
PORTER	0.864 ±0.004	0.505 ±0.030	0.131 ±0.001	0.514 ±0.009	**0.892** ±0.002	0.758 ±0.011	8.223 ±0.137	0.186 ±0.004	0.535 ±0.005	0.764 ±0.003

Method	PPICU → HIRID					PPICU → MIMIC				
	mort AUROC↑	los MSE↓	forecast MSE↓	lab AUROC↑	drug AUROC↑	mort AUROC↑	los MSE↓	forecast MSE↓	lab AUROC↑	drug AUROC↑
ATTACK	0.499 ±0.003	0.354 ±0.018	0.105 ±0.000	0.501 ±0.013	0.698 ±0.001	0.498 ±0.003	0.665 ±0.013	0.115 ±0.000	0.427 ±0.025	0.661 ±0.002
RANDOM	0.512 ±0.015	0.483 ±0.037	0.108 ±0.001	0.509 ±0.018	0.696 ±0.004	0.506 ±0.005	0.782 ±0.031	0.120 ±0.001	0.439 ±0.018	0.654 ±0.003
CLMBR-T (Wornow et al., 2023)	0.690 ±0.064	0.116 ±0.020	0.107 ±0.001	0.629 ±0.006	0.746 ±0.008	0.664 ±0.044	0.712 ±0.028	0.112 ±0.002	0.510 ±0.005	0.692 ±0.008
GARLIC (Wang et al., 2026)	0.317 ±0.004	0.263 ±0.059	0.100 ±0.001	0.572 ±0.027	0.469 ±0.015	0.514 ±0.023	0.465 ±0.004	0.111 ±0.001	0.463 ±0.018	0.660 ±0.008
KEDGN (Luo et al., 2024)	0.347 ±0.040	0.457 ±0.092	**0.085** ±0.002	0.540 ±0.021	0.449 ±0.049	0.697 ±0.059	1.476 ±0.539	**0.097** ±0.001	0.456 ±0.010	0.543 ±0.036
MIRA (Li et al., 2026)	0.795 ±0.002	**0.112** ±0.009	0.107 ±0.001	0.561 ±0.032	0.791 ±0.006	0.739 ±0.001	0.467 ±0.006	0.117 ±0.001	0.451 ±0.024	0.816 ±0.002
TabICL (Qu et al., 2025)	0.644 ±0.019	1.140 ±0.365	0.098 ±0.003	0.529 ±0.008	0.590 ±0.057	0.632 ±0.015	1.029 ±0.103	0.128 ±0.004	**0.511** ±0.008	0.615 ±0.067
MEWS (Subbe et al., 2001)	0.815 ±0.012	0.125 ±0.006	0.111 ±0.001	0.519 ±0.017	0.760 ±0.005	0.806 ±0.002	0.474 ±0.005	0.121 ±0.000	0.453 ±0.017	0.752 ±0.003
SOFA (Vincent et al., 1996)	0.822 ±0.006	0.156 ±0.004	0.107 ±0.001	0.570 ±0.054	**0.861** ±0.002	0.828 ±0.002	0.510 ±0.005	0.121 ±0.000	0.509 ±0.010	**0.833** ±0.004
Markdown (Hegselmann et al., 2026)	**0.902** ±0.002	0.146 ±0.016	0.103 ±0.001	0.611 ±0.006	0.738 ±0.023	0.824 ±0.008	0.449 ±0.009	0.117 ±0.001	0.494 ±0.005	0.814 ±0.014
Record2Vec (Ji et al., 2026)	0.831 ±0.004	0.386 ±0.036	0.108 ±0.000	0.598 ±0.005	0.743 ±0.008	0.804 ±0.002	0.501 ±0.016	0.118 ±0.001	0.502 ±0.003	0.763 ±0.007
Template (Demirel et al., 2026)	0.884 ±0.004	0.113 ±0.011	0.104 ±0.001	**0.630** ±0.003	0.669 ±0.028	0.821 ±0.005	**0.439** ±0.005	0.120 ±0.002	0.505 ±0.006	0.793 ±0.007
QA	0.780 ±0.014	0.248 ±0.025	0.106 ±0.000	0.555 ±0.038	0.758 ±0.006	0.753 ±0.005	0.475 ±0.013	0.116 ±0.001	0.475 ±0.024	0.752 ±0.005
Treatment	0.819 ±0.014	0.241 ±0.034	0.107 ±0.001	0.550 ±0.040	0.810 ±0.006	0.788 ±0.004	0.445 ±0.016	0.119 ±0.001	0.469 ±0.031	0.802 ±0.006
Organ-severity trajectory	0.832 ±0.006	0.311 ±0.073	0.108 ±0.000	0.574 ±0.024	0.825 ±0.002	**0.830** ±0.005	0.506 ±0.028	0.119 ±0.001	0.485 ±0.016	0.795 ±0.002
PORTER	0.862 ±0.007	0.346 ±0.022	0.107 ±0.001	0.596 ±0.037	0.795 ±0.003	0.825 ±0.006	0.460 ±0.014	0.118 ±0.001	0.495 ±0.020	0.800 ±0.005

All six directions			
Mean rank↓	Cumulative occurrences ↑		
	Top-1	Top-2	Top-3
15 (12.6)	0	1	2
16 (13.7)	1	1	1
7 (7.4)	0	3	8
12 (10.1)	1	4	5
14 (12.0)	4	4	4
10 (8.5)	1	2	3
13 (10.6)	2	3	3
9 (8.3)	0	0	3
2 (5.1)	6	10	12
4 (5.7)	5	9	**13**
5 (6.9)	0	0	1
3 (5.5)	**8**	**12**	**13**
11 (9.4)	0	0	0
6 (7.2)	0	1	2
8 (8.0)	1	2	7
1 (5.1)	1	8	**13**

Table 14: Zero-shot transfer, all six directions, the MedGemma arm at the xLSTM-Mixer head. The directions are carried over three blocks of two. Cells are the mean over the five cross-validation folds with one standard deviation. The columns to the right of the first block summarize all six directions. *Mean rank* is each method's position among the sixteen by its average rank over the thirty columns, with that average in parentheses, and *Cumulative occurrences* counts the columns, out of thirty, in which a method places in the top-1, 2, 3. Groups, metrics and marking as in Table 1.

Method	HIRID → MIMIC					HIRID → PPICU				
	mort AUROC↑	los MSE↓	forecast MSE↓	lab AUROC↑	drug AUROC↑	mort AUROC↑	los MSE↓	forecast MSE↓	lab AUROC↑	drug AUROC↑
ATTACK	0.498 ±0.009	3.550 ±0.020	0.264 ±0.099	0.376 ±0.004	0.659 ±0.002	0.500 ±0.001	10.111 ±0.098	0.269 ±0.112	0.484 ±0.004	0.634 ±0.000
RANDOM	0.500 ±0.010	3.541 ±0.077	0.410 ±0.176	0.384 ±0.014	0.659 ±0.004	0.500 ±0.012	9.824 ±0.232	0.451 ±0.193	0.487 ±0.012	0.632 ±0.003
CLMBR-T (Wornow et al., 2023)	0.754 ±0.008	3.044 ±0.108	1.054 ±0.954	0.486 ±0.019	0.765 ±0.009	0.661 ±0.006	7.398 ±0.203	0.972 ±0.895	0.585 ±0.008	0.656 ±0.007
GARLIC (Wang et al., 2026)	0.584 ±0.013	3.098 ±0.085	0.321 ±0.227	0.419 ±0.024	0.681 ±0.007	0.641 ±0.008	8.192 ±0.178	0.314 ±0.214	0.543 ±0.030	0.696 ±0.014
KEDGN (Luo et al., 2024)	0.300 ±0.035	4.081 ±0.627	0.447 ±0.257	0.407 ±0.045	0.375 ±0.077	0.660 ±0.043	14.418 ±2.387	0.459 ±0.290	0.513 ±0.027	0.416 ±0.013
MIRA (Li et al., 2026)	0.646 ±0.023	3.035 ±0.045	0.773 ±0.642	0.404 ±0.028	0.778 ±0.059	0.649 ±0.019	8.096 ±0.248	0.849 ±0.670	0.509 ±0.024	0.691 ±0.022
TabICL (Qu et al., 2025)	0.759 ±0.015	3.725 ±0.148	0.157 ±0.018	0.430 ±0.015	0.690 ±0.009	0.746 ±0.022	10.281 ±0.430	0.160 ±0.015	0.537 ±0.025	0.711 ±0.009
MEWS (Subbe et al., 2001)	0.783 ±0.003	2.874 ±0.014	0.512 ±0.338	0.424 ±0.013	0.722 ±0.004	0.682 ±0.005	7.313 ±0.123	0.506 ±0.256	0.491 ±0.005	0.711 ±0.001
SOFA (Vincent et al., 1996)	**0.801** ±0.008	2.873 ±0.042	0.346 ±0.349	0.465 ±0.016	0.849 ±0.006	0.729 ±0.005	7.180 ±0.065	0.382 ±0.422	0.537 ±0.007	**0.828** ±0.002
Markdown (Hegselmann et al., 2026)	0.767 ±0.010	2.777 ±0.118	0.269 ±0.113	0.468 ±0.013	0.838 ±0.025	0.682 ±0.004	**6.709** ±0.451	0.289 ±0.155	0.578 ±0.009	0.577 ±0.022
Record2Vec (Ji et al., 2026)	0.758 ±0.006	2.948 ±0.029	0.190 ±0.067	0.457 ±0.006	0.837 ±0.004	0.727 ±0.003	8.147 ±0.076	0.197 ±0.064	0.549 ±0.013	0.722 ±0.006
Template (Demirel et al., 2026)	0.751 ±0.011	2.835 ±0.049	0.233 ±0.104	**0.494** ±0.013	**0.897** ±0.007	0.736 ±0.002	7.176 ±0.276	0.266 ±0.149	**0.596** ±0.004	0.650 ±0.012
QA	0.735 ±0.004	3.165 ±0.044	0.201 ±0.087	0.435 ±0.003	0.769 ±0.008	0.670 ±0.006	8.617 ±0.143	0.235 ±0.134	0.518 ±0.004	0.695 ±0.006
Treatment	0.744 ±0.007	3.013 ±0.035	**0.153** ±0.022	0.447 ±0.003	0.830 ±0.003	0.704 ±0.006	7.378 ±0.334	**0.156** ±0.003	0.521 ±0.006	0.734 ±0.006
Organ-severity trajectory	0.777 ±0.007	3.329 ±0.087	4.743 ±2.446	0.437 ±0.028	0.858 ±0.005	0.728 ±0.005	9.651 ±0.419	3.835 ±2.079	0.563 ±0.030	0.742 ±0.004
PORTER	0.794 ±0.003	**2.768** ±0.024	0.210 ±0.131	0.456 ±0.007	0.892 ±0.002	**0.758** ±0.001	8.949 ±0.086	0.229 ±0.169	0.559 ±0.010	0.769 ±0.001

Method	MIMIC → HIRID					MIMIC → PPICU				
	mort AUROC↑	los MSE↓	forecast MSE↓	lab AUROC↑	drug AUROC↑	mort AUROC↑	los MSE↓	forecast MSE↓	lab AUROC↑	drug AUROC↑
ATTACK	0.498 ±0.008	0.747 ±0.004	0.160 ±0.048	0.387 ±0.012	0.697 ±0.002	0.500 ±0.000	9.859 ±0.113	0.192 ±0.067	0.488 ±0.015	0.634 ±0.000
RANDOM	0.503 ±0.031	0.791 ±0.019	0.470 ±0.287	0.411 ±0.003	0.692 ±0.006	0.503 ±0.009	9.646 ±0.135	0.443 ±0.259	0.473 ±0.004	0.631 ±0.004
CLMBR-T (Wornow et al., 2023)	0.777 ±0.017	0.564 ±0.055	0.484 ±0.204	0.524 ±0.039	0.799 ±0.004	0.695 ±0.010	7.853 ±0.404	0.426 ±0.183	0.571 ±0.020	0.656 ±0.017
GARLIC (Wang et al., 2026)	0.586 ±0.019	0.610 ±0.050	0.490 ±0.720	0.466 ±0.004	0.727 ±0.017	0.598 ±0.011	9.302 ±0.139	0.464 ±0.603	0.482 ±0.021	0.683 ±0.006
KEDGN (Luo et al., 2024)	0.613 ±0.018	1.082 ±0.195	0.647 ±0.397	0.455 ±0.020	0.572 ±0.064	0.511 ±0.039	7.919 ±0.215	0.619 ±0.391	0.512 ±0.014	0.661 ±0.024
MIRA (Li et al., 2026)	0.749 ±0.033	0.604 ±0.108	0.341 ±0.216	0.418 ±0.038	0.765 ±0.043	0.619 ±0.018	9.049 ±0.455	0.332 ±0.162	0.500 ±0.028	0.674 ±0.005
TabICL (Qu et al., 2025)	0.834 ±0.027	0.630 ±0.038	0.154 ±0.047	0.506 ±0.016	0.763 ±0.011	0.732 ±0.020	9.464 ±0.258	0.204 ±0.061	0.530 ±0.005	0.726 ±0.006
MEWS (Subbe et al., 2001)	0.831 ±0.009	0.552 ±0.022	0.427 ±0.311	0.462 ±0.007	0.748 ±0.006	0.713 ±0.006	7.885 ±0.113	0.473 ±0.346	0.516 ±0.011	0.704 ±0.002
SOFA (Vincent et al., 1996)	0.852 ±0.009	0.539 ±0.015	1.202 ±0.478	0.486 ±0.020	0.875 ±0.005	**0.770** ±0.007	7.860 ±0.083	1.114 ±0.381	0.542 ±0.016	**0.844** ±0.004
Markdown (Hegselmann et al., 2026)	**0.871** ±0.015	**0.399** ±0.015	0.135 ±0.019	0.543 ±0.010	0.838 ±0.008	0.671 ±0.012	6.557 ±0.366	0.167 ±0.017	0.558 ±0.006	0.631 ±0.018
Record2Vec (Ji et al., 2026)	0.833 ±0.013	0.519 ±0.019	0.141 ±0.015	0.495 ±0.010	0.833 ±0.005	0.740 ±0.004	8.765 ±0.275	**0.162** ±0.025	0.530 ±0.004	0.732 ±0.004
Template (Demirel et al., 2026)	0.843 ±0.012	0.481 ±0.049	0.135 ±0.009	**0.558** ±0.015	0.849 ±0.008	0.713 ±0.012	**6.274** ±0.228	0.170 ±0.018	**0.573** ±0.009	0.713 ±0.014
QA	0.815 ±0.008	0.575 ±0.009	**0.126** ±0.004	0.448 ±0.007	0.783 ±0.007	0.683 ±0.007	8.720 ±0.123	0.168 ±0.011	0.500 ±0.005	0.674 ±0.005
Treatment	0.822 ±0.006	0.569 ±0.007	0.143 ±0.030	0.471 ±0.007	0.847 ±0.003	0.732 ±0.006	8.446 ±0.242	0.175 ±0.016	0.517 ±0.003	0.732 ±0.005
Organ-severity trajectory	0.814 ±0.010	0.654 ±0.036	0.176 ±0.077	0.444 ±0.032	0.875 ±0.005	0.733 ±0.003	8.619 ±0.385	0.202 ±0.054	0.519 ±0.037	0.745 ±0.007
PORTER	0.863 ±0.003	0.505 ±0.018	0.138 ±0.008	0.520 ±0.019	**0.892** ±0.003	0.760 ±0.009	8.607 ±0.096	0.169 ±0.020	0.545 ±0.010	0.762 ±0.003

Method	PPICU → HIRID					PPICU → MIMIC				
	mort AUROC↑	los MSE↓	forecast MSE↓	lab AUROC↑	drug AUROC↑	mort AUROC↑	los MSE↓	forecast MSE↓	lab AUROC↑	drug AUROC↑
ATTACK	0.502 ±0.006	0.324 ±0.049	0.121 ±0.010	0.513 ±0.010	0.697 ±0.002	0.505 ±0.007	0.643 ±0.038	0.135 ±0.016	0.421 ±0.010	0.661 ±0.003
RANDOM	0.517 ±0.006	0.416 ±0.053	0.428 ±0.145	0.508 ±0.004	0.699 ±0.003	0.508 ±0.010	0.718 ±0.042	0.454 ±0.180	0.434 ±0.006	0.661 ±0.003
CLMBR-T (Wornow et al., 2023)	0.781 ±0.026	0.116 ±0.003	0.285 ±0.280	0.591 ±0.061	0.750 ±0.011	0.730 ±0.031	0.850 ±0.089	0.231 ±0.186	0.491 ±0.034	0.690 ±0.005
GARLIC (Wang et al., 2026)	0.316 ±0.012	0.270 ±0.028	0.698 ±1.056	0.531 ±0.046	0.654 ±0.024	0.520 ±0.014	0.493 ±0.024	0.512 ±0.686	0.467 ±0.011	0.719 ±0.008
KEDGN (Luo et al., 2024)	0.379 ±0.067	0.388 ±0.107	0.646 ±0.622	0.520 ±0.010	0.463 ±0.077	0.666 ±0.106	0.948 ±0.136	0.592 ±0.614	0.471 ±0.027	0.589 ±0.078
MIRA (Li et al., 2026)	0.794 ±0.025	0.110 ±0.008	0.147 ±0.035	0.521 ±0.035	0.792 ±0.008	0.714 ±0.035	0.484 ±0.010	0.179 ±0.035	0.453 ±0.013	0.807 ±0.003
TabICL (Qu et al., 2025)	0.650 ±0.031	0.873 ±0.215	0.161 ±0.038	0.533 ±0.013	0.697 ±0.058	0.648 ±0.028	0.797 ±0.097	0.154 ±0.018	0.512 ±0.015	0.659 ±0.037
MEWS (Subbe et al., 2001)	0.813 ±0.010	0.146 ±0.013	0.317 ±0.162	0.548 ±0.014	0.752 ±0.008	0.804 ±0.003	0.497 ±0.011	0.287 ±0.156	0.487 ±0.003	0.749 ±0.004
SOFA (Vincent et al., 1996)	0.817 ±0.010	0.178 ±0.013	0.259 ±0.174	0.613 ±0.010	**0.857** ±0.003	0.824 ±0.004	0.538 ±0.036	0.273 ±0.149	**0.526** ±0.005	**0.823** ±0.004
Markdown (Hegselmann et al., 2026)	**0.899** ±0.003	0.148 ±0.036	**0.106** ±0.006	0.611 ±0.002	0.751 ±0.023	0.819 ±0.009	0.462 ±0.019	0.126 ±0.008	0.496 ±0.003	0.812 ±0.008
Record2Vec (Ji et al., 2026)	0.838 ±0.007	0.421 ±0.047	0.112 ±0.002	0.596 ±0.005	0.746 ±0.009	0.803 ±0.004	0.502 ±0.020	0.121 ±0.003	0.503 ±0.005	0.766 ±0.007
Template (Demirel et al., 2026)	0.890 ±0.004	**0.101** ±0.017	0.122 ±0.036	**0.629** ±0.010	0.657 ±0.021	0.813 ±0.006	0.441 ±0.011	0.128 ±0.009	0.510 ±0.008	0.807 ±0.005
QA	0.779 ±0.012	0.266 ±0.008	0.112 ±0.004	0.577 ±0.007	0.763 ±0.005	0.757 ±0.004	0.480 ±0.007	0.121 ±0.004	0.490 ±0.004	0.757 ±0.006
Treatment	0.823 ±0.008	0.243 ±0.020	0.113 ±0.003	0.575 ±0.008	0.810 ±0.003	0.789 ±0.002	**0.439** ±0.008	0.122 ±0.003	0.496 ±0.006	0.803 ±0.003
Organ-severity trajectory	0.842 ±0.003	0.254 ±0.051	0.450 ±0.311	0.565 ±0.027	0.822 ±0.005	**0.831** ±0.003	0.495 ±0.010	0.522 ±0.413	0.479 ±0.029	0.796 ±0.009
PORTER	0.861 ±0.008	0.364 ±0.038	0.112 ±0.002	0.611 ±0.002	0.803 ±0.007	0.824 ±0.006	0.456 ±0.014	**0.120** ±0.001	0.506 ±0.007	0.807 ±0.004

All six directions			
Mean rank↓	Cumulative occurrences↑		
	Top-1	Top-2	Top-3
14 (13.0)	0	0	0
16 (13.9)	0	0	0
9 (8.8)	0	3	6
13 (11.9)	0	0	0
15 (13.8)	0	0	0
12 (10.2)	0	1	2
11 (9.0)	0	4	4
10 (8.9)	0	0	1
4 (5.4)	7	9	12
3 (4.9)	5	10	14
5 (5.9)	1	3	6
2 (4.2)	**8**	**13**	**16**
8 (8.2)	1	1	4
6 (6.1)	3	3	4
7 (8.0)	1	3	6
1 (3.8)	4	10	15

Table 15: Zero-shot transfer, all six directions, the Gemma-4 arm at the LSTM head. The directions are carried over three blocks of two. Cells are the mean over the five cross-validation folds with one standard deviation. The columns to the right of the first block summarize all six directions. *Mean rank* is each method's position among the sixteen by its average rank over the thirty columns, with that average in parentheses, and *Cumulative occurrences* counts the columns, out of thirty, in which a method places in the top-1, 2, 3. Groups, metrics and marking as in Table 1.

Method	HIRID → MIMIC					HIRID → PPICU				
	mort AUROC↑	los MSE↓	forecast MSE↓	lab AUROC↑	drug AUROC↑	mort AUROC↑	los MSE↓	forecast MSE↓	lab AUROC↑	drug AUROC↑
ATTACK	0.500 ±0.000	3.483 ±0.006	0.127 ±0.000	0.373 ±0.034	0.661 ±0.003	0.500 ±0.000	9.745 ±0.039	0.161 ±0.001	0.475 ±0.027	0.634 ±0.000
RANDOM	0.507 ±0.010	3.588 ±0.079	0.135 ±0.002	0.394 ±0.010	0.663 ±0.003	0.493 ±0.010	10.094 ±0.213	0.162 ±0.001	0.491 ±0.018	0.634 ±0.003
CLMBR-T (Wornow et al., 2023)	0.753 ±0.011	2.934 ±0.043	0.132 ±0.001	**0.468** ±0.020	0.772 ±0.006	0.662 ±0.008	7.758 ±0.485	0.156 ±0.002	0.574 ±0.009	0.663 ±0.007
GARLIC (Wang et al., 2026)	0.550 ±0.028	3.284 ±0.043	0.127 ±0.001	0.401 ±0.006	0.690 ±0.006	0.623 ±0.017	9.025 ±0.213	0.162 ±0.001	0.531 ±0.024	0.673 ±0.012
KEDGN (Luo et al., 2024)	0.304 ±0.024	3.663 ±0.166	**0.117** ±0.003	0.430 ±0.024	0.329 ±0.024	0.629 ±0.031	11.271 ±0.147	**0.138** ±0.003	0.446 ±0.020	0.430 ±0.017
MIRA (Li et al., 2026)	0.639 ±0.014	3.003 ±0.045	0.128 ±0.001	0.369 ±0.037	0.797 ±0.003	0.663 ±0.005	7.919 ±0.316	0.164 ±0.001	0.500 ±0.014	0.693 ±0.004
TabICL (Qu et al., 2025)	0.751 ±0.017	3.825 ±0.030	0.140 ±0.002	0.424 ±0.009	0.695 ±0.010	0.743 ±0.024	10.934 ±0.207	0.175 ±0.007	0.541 ±0.025	0.716 ±0.011
MEWS (Subbe et al., 2001)	0.784 ±0.003	2.856 ±0.052	0.131 ±0.001	0.419 ±0.011	0.729 ±0.003	0.679 ±0.006	7.596 ±0.268	0.156 ±0.001	0.513 ±0.010	0.711 ±0.002
SOFA (Vincent et al., 1996)	0.807 ±0.008	2.928 ±0.089	0.134 ±0.001	0.460 ±0.006	0.857 ±0.003	0.738 ±0.015	7.312 ±0.216	0.156 ±0.001	0.557 ±0.006	**0.832** ±0.004
Markdown (Hegselmann et al., 2026)	0.766 ±0.005	2.776 ±0.072	0.136 ±0.001	0.455 ±0.006	0.853 ±0.013	0.704 ±0.008	**6.694** ±0.156	0.155 ±0.001	0.578 ±0.005	0.613 ±0.033
Record2Vec (Ji et al., 2026)	0.806 ±0.006	3.011 ±0.023	0.134 ±0.001	0.418 ±0.030	0.845 ±0.005	0.746 ±0.003	8.226 ±0.076	0.156 ±0.002	0.536 ±0.014	0.748 ±0.003
Template (Demirel et al., 2026)	0.747 ±0.008	**2.675** ±0.162	0.135 ±0.001	0.467 ±0.013	0.895 ±0.004	0.734 ±0.002	7.556 ±0.278	0.156 ±0.001	**0.587** ±0.007	0.658 ±0.009
QA	0.796 ±0.010	3.193 ±0.058	0.133 ±0.002	0.417 ±0.031	0.811 ±0.006	0.734 ±0.008	8.890 ±0.144	0.156 ±0.002	0.527 ±0.013	0.721 ±0.004
Treatment	0.776 ±0.007	3.179 ±0.064	0.134 ±0.002	0.408 ±0.026	0.836 ±0.003	0.734 ±0.005	8.680 ±0.214	0.157 ±0.001	0.526 ±0.018	0.747 ±0.003
Organ-severity trajectory	0.796 ±0.011	3.240 ±0.021	0.132 ±0.001	0.376 ±0.026	0.867 ±0.004	0.761 ±0.004	9.352 ±0.074	0.160 ±0.001	0.479 ±0.030	0.736 ±0.004
PORTER	**0.809** ±0.021	2.922 ±0.053	0.135 ±0.001	0.415 ±0.035	**0.909** ±0.004	**0.765** ±0.014	8.677 ±0.165	0.159 ±0.001	0.545 ±0.019	0.774 ±0.005

Method	MIMIC → HIRID					MIMIC → PPICU				
	mort AUROC↑	los MSE↓	forecast MSE↓	lab AUROC↑	drug AUROC↑	mort AUROC↑	los MSE↓	forecast MSE↓	lab AUROC↑	drug AUROC↑
ATTACK	0.497 ±0.005	0.760 ±0.003	0.125 ±0.001	0.418 ±0.023	0.697 ±0.001	0.500 ±0.001	9.587 ±0.036	0.186 ±0.001	0.475 ±0.023	0.634 ±0.000
RANDOM	0.499 ±0.015	0.841 ±0.026	0.125 ±0.001	0.392 ±0.005	0.696 ±0.002	0.499 ±0.010	10.159 ±0.108	0.180 ±0.001	0.476 ±0.004	0.632 ±0.001
CLMBR-T (Wornow et al., 2023)	0.787 ±0.012	0.527 ±0.043	0.128 ±0.001	0.449 ±0.012	0.801 ±0.003	0.682 ±0.017	8.079 ±0.314	0.186 ±0.002	0.530 ±0.007	0.654 ±0.005
GARLIC (Wang et al., 2026)	0.598 ±0.007	0.651 ±0.073	**0.121** ±0.002	0.442 ±0.018	0.702 ±0.026	0.596 ±0.013	9.710 ±0.204	0.184 ±0.001	0.490 ±0.007	0.682 ±0.005
KEDGN (Luo et al., 2024)	0.605 ±0.048	0.903 ±0.172	0.130 ±0.002	0.440 ±0.021	0.605 ±0.029	0.511 ±0.050	9.771 ±0.722	**0.179** ±0.001	0.487 ±0.024	0.675 ±0.007
MIRA (Li et al., 2026)	0.749 ±0.010	0.488 ±0.050	0.127 ±0.001	0.394 ±0.020	0.787 ±0.003	0.593 ±0.002	8.663 ±0.303	0.186 ±0.001	0.486 ±0.027	0.674 ±0.004
TabICL (Qu et al., 2025)	0.823 ±0.018	0.618 ±0.070	0.124 ±0.003	0.458 ±0.016	0.768 ±0.010	0.719 ±0.017	10.306 ±0.244	0.183 ±0.002	0.514 ±0.006	0.730 ±0.008
MEWS (Subbe et al., 2001)	0.853 ±0.004	0.526 ±0.008	0.135 ±0.001	0.453 ±0.014	0.760 ±0.005	0.719 ±0.005	8.377 ±0.246	0.190 ±0.001	0.520 ±0.010	0.712 ±0.003
SOFA (Vincent et al., 1996)	0.854 ±0.005	0.561 ±0.007	0.132 ±0.001	**0.473** ±0.007	0.875 ±0.002	0.763 ±0.009	8.309 ±0.165	0.188 ±0.001	**0.540** ±0.012	**0.846** ±0.001
Markdown (Hegselmann et al., 2026)	0.867 ±0.012	**0.389** ±0.008	0.139 ±0.001	0.438 ±0.040	0.822 ±0.014	0.672 ±0.007	6.665 ±0.268	0.194 ±0.002	0.486 ±0.040	0.624 ±0.016
Record2Vec (Ji et al., 2026)	0.886 ±0.006	0.508 ±0.018	0.132 ±0.001	0.422 ±0.022	0.828 ±0.004	0.769 ±0.002	8.397 ±0.163	0.188 ±0.002	0.479 ±0.021	0.739 ±0.002
Template (Demirel et al., 2026)	0.799 ±0.027	0.451 ±0.025	0.140 ±0.003	0.412 ±0.024	0.819 ±0.027	0.704 ±0.003	**6.389** ±0.097	0.195 ±0.002	0.476 ±0.024	0.711 ±0.014
QA	0.849 ±0.014	0.672 ±0.045	0.131 ±0.002	0.419 ±0.025	0.784 ±0.007	0.756 ±0.003	8.967 ±0.073	0.186 ±0.005	0.475 ±0.024	0.720 ±0.003
Treatment	0.839 ±0.009	0.582 ±0.028	0.132 ±0.001	0.420 ±0.025	0.826 ±0.005	0.766 ±0.003	9.172 ±0.271	0.187 ±0.002	0.474 ±0.024	0.745 ±0.002
Organ-severity trajectory	0.830 ±0.015	0.878 ±0.067	0.133 ±0.001	0.412 ±0.028	0.858 ±0.006	0.782 ±0.005	8.709 ±0.213	0.189 ±0.001	0.480 ±0.017	0.744 ±0.004
PORTER	**0.900** ±0.005	0.612 ±0.035	0.133 ±0.002	0.418 ±0.025	**0.914** ±0.003	**0.798** ±0.004	8.027 ±0.091	0.189 ±0.002	0.473 ±0.025	0.783 ±0.004

Method	PPICU → HIRID					PPICU → MIMIC				
	mort AUROC↑	los MSE↓	forecast MSE↓	lab AUROC↑	drug AUROC↑	mort AUROC↑	los MSE↓	forecast MSE↓	lab AUROC↑	drug AUROC↑
ATTACK	0.496 ±0.004	0.416 ±0.011	0.105 ±0.001	0.499 ±0.013	0.696 ±0.002	0.495 ±0.004	0.713 ±0.009	0.115 ±0.001	0.416 ±0.026	0.661 ±0.002
RANDOM	0.515 ±0.012	0.449 ±0.012	0.107 ±0.001	0.487 ±0.016	0.698 ±0.002	0.507 ±0.008	0.743 ±0.011	0.116 ±0.001	0.425 ±0.031	0.660 ±0.003
CLMBR-T (Wornow et al., 2023)	0.665 ±0.040	**0.092** ±0.008	0.108 ±0.001	**0.616** ±0.010	0.748 ±0.013	0.675 ±0.027	0.647 ±0.061	0.116 ±0.001	**0.502** ±0.008	0.679 ±0.009
GARLIC (Wang et al., 2026)	0.314 ±0.010	0.257 ±0.054	0.101 ±0.001	0.572 ±0.032	0.485 ±0.052	0.509 ±0.024	0.494 ±0.021	0.114 ±0.001	0.472 ±0.010	0.670 ±0.026
KEDGN (Luo et al., 2024)	0.317 ±0.074	0.129 ±0.033	**0.091** ±0.001	0.530 ±0.029	0.513 ±0.064	0.658 ±0.081	0.528 ±0.012	**0.103** ±0.001	0.449 ±0.016	0.580 ±0.066
MIRA (Li et al., 2026)	0.793 ±0.017	0.109 ±0.019	0.105 ±0.001	0.534 ±0.026	0.788 ±0.007	0.741 ±0.005	0.485 ±0.015	0.114 ±0.001	0.434 ±0.021	0.812 ±0.003
TabICL (Qu et al., 2025)	0.668 ±0.049	0.863 ±0.610	0.096 ±0.002	0.509 ±0.009	0.716 ±0.038	0.694 ±0.019	0.690 ±0.117	0.120 ±0.002	0.472 ±0.034	0.660 ±0.030
MEWS (Subbe et al., 2001)	0.829 ±0.009	0.104 ±0.006	0.109 ±0.001	0.505 ±0.029	0.760 ±0.010	0.804 ±0.005	0.458 ±0.005	0.121 ±0.000	0.450 ±0.034	0.753 ±0.004
SOFA (Vincent et al., 1996)	0.807 ±0.015	0.117 ±0.006	0.108 ±0.001	0.571 ±0.050	**0.869** ±0.003	0.827 ±0.003	0.459 ±0.020	0.121 ±0.000	0.475 ±0.045	**0.837** ±0.002
Markdown (Hegselmann et al., 2026)	**0.900** ±0.003	0.138 ±0.008	0.106 ±0.001	0.548 ±0.054	0.737 ±0.029	0.816 ±0.009	0.445 ±0.011	0.122 ±0.003	0.450 ±0.039	0.830 ±0.003
Record2Vec (Ji et al., 2026)	0.863 ±0.007	0.422 ±0.160	0.109 ±0.001	0.582 ±0.036	0.813 ±0.008	0.833 ±0.002	**0.426** ±0.020	0.121 ±0.001	0.481 ±0.040	0.777 ±0.010
Template (Demirel et al., 2026)	0.876 ±0.020	0.106 ±0.017	0.106 ±0.000	0.599 ±0.043	0.706 ±0.047	0.818 ±0.004	0.436 ±0.008	0.124 ±0.001	0.486 ±0.031	0.820 ±0.008
QA	0.804 ±0.022	0.284 ±0.028	0.109 ±0.001	0.555 ±0.024	0.787 ±0.005	0.814 ±0.003	0.459 ±0.015	0.119 ±0.001	0.484 ±0.029	0.766 ±0.009
Treatment	0.812 ±0.011	0.406 ±0.065	0.109 ±0.001	0.547 ±0.036	0.769 ±0.012	0.827 ±0.002	0.470 ±0.013	0.119 ±0.001	0.464 ±0.048	0.797 ±0.008
Organ-severity trajectory	0.798 ±0.006	0.622 ±0.104	0.108 ±0.001	0.516 ±0.012	0.768 ±0.002	0.829 ±0.002	0.496 ±0.014	0.120 ±0.001	0.418 ±0.013	0.812 ±0.008
PORTER	0.856 ±0.003	0.564 ±0.145	0.108 ±0.001	0.584 ±0.035	0.788 ±0.006	**0.858** ±0.002	0.441 ±0.018	0.122 ±0.001	0.487 ±0.030	0.826 ±0.006

Mean rank↓	All six directions: Cumulative occurrences ↑ Top-1	Top-2	Top-3
15 (12.4)	0	0	1
16 (12.9)	0	1	2
5 (7.2)	4	5	6
12 (9.7)	1	2	4
14 (10.7)	5	5	5
10 (8.8)	0	1	2
12 (9.7)	0	3	4
7 (7.9)	0	1	4
1 (4.8)	6	9	10
5 (7.2)	3	8	9
3 (6.2)	1	4	10
4 (7.1)	3	9	12
9 (8.3)	0	0	0
8 (8.1)	0	0	1
11 (9.0)	0	2	5
2 (6.1)	**7**	**10**	**15**

Table 16: Zero-shot transfer, all six directions, the Gemma-4 arm at the linear head. The directions are carried over three blocks of two. Cells are the mean over the five cross-validation folds with one standard deviation. The columns to the right of the first block summarize all six directions. *Mean rank* is each method's position among the sixteen by its average rank over the thirty columns, with that average in parentheses, and *Cumulative occurrences* counts the columns, out of thirty, in which a method places in the top-1, 2, 3. Groups, metrics and marking as in Table 1.

Method	HIRID → MIMIC					HIRID → PPICU				
	mort AUROC↑	los MSE↓	forecast MSE↓	lab AUROC↑	drug AUROC↑	mort AUROC↑	los MSE↓	forecast MSE↓	lab AUROC↑	drug AUROC↑
ATTACK	0.502 ±0.003	3.527 ±0.021	0.127 ±0.000	0.390 ±0.032	0.662 ±0.003	0.500 ±0.001	9.993 ±0.104	0.162 ±0.001	0.499 ±0.023	0.634 ±0.000
RANDOM	0.517 ±0.008	3.516 ±0.023	0.140 ±0.001	0.384 ±0.007	0.663 ±0.002	0.504 ±0.015	9.816 ±0.040	0.166 ±0.002	0.493 ±0.014	0.632 ±0.002
CLMBR-T (Wornow et al., 2023)	0.761 ±0.007	2.952 ±0.022	0.130 ±0.002	0.477 ±0.018	0.760 ±0.010	0.670 ±0.010	7.777 ±0.250	0.159 ±0.003	0.565 ±0.019	0.658 ±0.009
GARLIC (Wang et al., 2026)	0.550 ±0.041	3.158 ±0.080	0.126 ±0.002	0.399 ±0.028	0.673 ±0.009	0.621 ±0.023	8.509 ±0.175	0.163 ±0.002	0.549 ±0.011	0.650 ±0.012
KEDGN (Luo et al., 2024)	0.281 ±0.015	4.373 ±0.692	**0.110** ±0.005	0.397 ±0.011	0.371 ±0.048	0.639 ±0.027	16.610 ±1.327	**0.138** ±0.008	0.516 ±0.024	0.462 ±0.029
MIRA (Li et al., 2026)	0.658 ±0.010	2.974 ±0.030	0.128 ±0.001	0.426 ±0.031	0.806 ±0.004	0.669 ±0.005	7.781 ±0.150	0.161 ±0.002	0.499 ±0.040	0.700 ±0.003
TabICL (Qu et al., 2025)	0.768 ±0.006	3.566 ±0.108	0.146 ±0.003	0.419 ±0.009	0.679 ±0.004	0.758 ±0.007	9.316 ±0.356	0.191 ±0.006	0.549 ±0.011	0.700 ±0.008
MEWS (Subbe et al., 2001)	0.783 ±0.005	2.853 ±0.016	0.129 ±0.001	0.424 ±0.020	0.729 ±0.002	0.679 ±0.004	7.240 ±0.034	0.155 ±0.002	0.498 ±0.013	0.709 ±0.001
SOFA (Vincent et al., 1996)	0.804 ±0.003	2.865 ±0.011	0.133 ±0.001	0.443 ±0.025	0.855 ±0.004	0.733 ±0.007	7.045 ±0.082	0.155 ±0.002	0.533 ±0.009	**0.824** ±0.003
Markdown (Hegselmann et al., 2026)	0.769 ±0.009	**2.758** ±0.057	0.136 ±0.003	0.462 ±0.012	0.854 ±0.012	0.680 ±0.006	**6.632** ±0.181	0.159 ±0.004	0.574 ±0.007	0.571 ±0.019
Record2Vec (Ji et al., 2026)	0.810 ±0.008	2.989 ±0.020	0.134 ±0.001	0.442 ±0.008	0.840 ±0.005	0.745 ±0.002	7.939 ±0.204	0.159 ±0.002	0.552 ±0.005	0.749 ±0.000
Template (Demirel et al., 2026)	0.757 ±0.005	2.766 ±0.079	0.136 ±0.002	**0.493** ±0.021	0.904 ±0.004	0.736 ±0.003	7.487 ±0.195	0.160 ±0.005	**0.593** ±0.005	0.673 ±0.012
QA	0.792 ±0.007	3.148 ±0.045	0.134 ±0.001	0.441 ±0.011	0.810 ±0.004	0.736 ±0.009	8.761 ±0.177	0.159 ±0.002	0.533 ±0.007	0.723 ±0.002
Treatment	0.774 ±0.012	3.103 ±0.023	0.134 ±0.001	0.420 ±0.033	0.838 ±0.003	0.729 ±0.009	8.058 ±0.238	0.156 ±0.002	0.531 ±0.007	0.744 ±0.004
Organ-severity trajectory	0.786 ±0.013	3.197 ±0.063	0.132 ±0.001	0.434 ±0.006	0.866 ±0.003	0.753 ±0.010	9.156 ±0.246	0.163 ±0.002	0.560 ±0.019	0.740 ±0.002
PORTER	**0.822** ±0.007	2.876 ±0.031	0.138 ±0.001	0.438 ±0.008	**0.909** ±0.003	**0.768** ±0.005	8.178 ±0.308	0.162 ±0.002	0.563 ±0.008	0.772 ±0.005

Method	MIMIC → HIRID					MIMIC → PPICU				
	mort AUROC↑	los MSE↓	forecast MSE↓	lab AUROC↑	drug AUROC↑	mort AUROC↑	los MSE↓	forecast MSE↓	lab AUROC↑	drug AUROC↑
ATTACK	0.500 ±0.001	0.749 ±0.003	0.125 ±0.000	0.408 ±0.026	0.697 ±0.001	0.500 ±0.000	9.863 ±0.189	0.186 ±0.001	0.470 ±0.038	0.634 ±0.000
RANDOM	0.505 ±0.014	0.807 ±0.004	0.128 ±0.001	0.409 ±0.008	0.695 ±0.002	0.494 ±0.008	9.588 ±0.067	**0.179** ±0.004	0.473 ±0.004	0.629 ±0.001
CLMBR-T (Wornow et al., 2023)	0.787 ±0.009	0.524 ±0.032	0.131 ±0.003	0.496 ±0.013	0.792 ±0.004	0.669 ±0.019	7.692 ±0.510	0.183 ±0.002	0.546 ±0.015	0.635 ±0.010
GARLIC (Wang et al., 2026)	0.595 ±0.015	0.591 ±0.020	**0.120** ±0.002	0.457 ±0.021	0.688 ±0.017	0.595 ±0.010	9.381 ±0.199	0.182 ±0.002	0.501 ±0.019	0.685 ±0.004
KEDGN (Luo et al., 2024)	0.626 ±0.022	1.064 ±0.175	0.143 ±0.003	0.452 ±0.019	0.597 ±0.057	0.517 ±0.042	8.034 ±0.231	0.190 ±0.006	0.516 ±0.009	0.655 ±0.030
MIRA (Li et al., 2026)	0.750 ±0.011	0.517 ±0.086	0.126 ±0.002	0.439 ±0.031	0.788 ±0.004	0.599 ±0.006	8.704 ±0.295	0.186 ±0.004	0.474 ±0.032	0.676 ±0.007
TabICL (Qu et al., 2025)	0.822 ±0.017	0.646 ±0.033	0.135 ±0.003	0.501 ±0.012	0.759 ±0.011	0.692 ±0.034	8.943 ±0.548	0.206 ±0.004	0.519 ±0.002	0.716 ±0.012
MEWS (Subbe et al., 2001)	0.840 ±0.007	0.509 ±0.017	0.136 ±0.001	0.457 ±0.017	0.749 ±0.004	0.723 ±0.003	7.846 ±0.099	0.188 ±0.001	0.506 ±0.022	0.708 ±0.002
SOFA (Vincent et al., 1996)	0.848 ±0.009	0.518 ±0.010	0.131 ±0.000	0.485 ±0.002	0.880 ±0.001	0.762 ±0.004	7.785 ±0.094	0.187 ±0.002	0.545 ±0.005	**0.843** ±0.002
Markdown (Hegselmann et al., 2026)	0.874 ±0.007	**0.376** ±0.005	0.137 ±0.003	0.538 ±0.014	0.839 ±0.012	0.680 ±0.012	6.752 ±0.290	0.195 ±0.006	0.551 ±0.013	0.638 ±0.024
Record2Vec (Ji et al., 2026)	0.884 ±0.005	0.542 ±0.020	0.131 ±0.002	0.493 ±0.008	0.830 ±0.004	0.762 ±0.007	7.990 ±0.088	0.189 ±0.002	0.540 ±0.007	0.740 ±0.002
Template (Demirel et al., 2026)	0.833 ±0.013	0.481 ±0.013	0.138 ±0.005	**0.559** ±0.015	0.856 ±0.006	0.717 ±0.009	**6.208** ±0.112	0.199 ±0.008	**0.567** ±0.013	0.727 ±0.019
QA	0.828 ±0.033	0.691 ±0.045	0.129 ±0.001	0.438 ±0.015	0.777 ±0.009	0.758 ±0.004	8.594 ±0.134	0.185 ±0.003	0.520 ±0.010	0.717 ±0.006
Treatment	0.830 ±0.016	0.593 ±0.024	0.129 ±0.001	0.457 ±0.003	0.833 ±0.005	0.763 ±0.003	8.643 ±0.102	0.184 ±0.002	0.529 ±0.007	0.740 ±0.003
Organ-severity trajectory	0.828 ±0.011	0.875 ±0.056	0.137 ±0.003	0.446 ±0.035	0.858 ±0.006	0.779 ±0.008	8.727 ±0.179	0.194 ±0.002	0.510 ±0.030	0.744 ±0.002
PORTER	**0.904** ±0.003	0.584 ±0.035	0.130 ±0.002	0.524 ±0.008	**0.913** ±0.003	**0.793** ±0.011	8.134 ±0.204	0.186 ±0.004	0.552 ±0.007	0.779 ±0.004

Method	PPICU → HIRID					PPICU → MIMIC				
	mort AUROC↑	los MSE↓	forecast MSE↓	lab AUROC↑	drug AUROC↑	mort AUROC↑	los MSE↓	forecast MSE↓	lab AUROC↑	drug AUROC↑
ATTACK	0.499 ±0.003	0.354 ±0.018	0.105 ±0.000	0.501 ±0.013	0.698 ±0.001	0.498 ±0.003	0.665 ±0.013	0.115 ±0.000	0.427 ±0.025	0.661 ±0.002
RANDOM	0.512 ±0.015	0.483 ±0.037	0.108 ±0.001	0.509 ±0.018	0.696 ±0.004	0.506 ±0.005	0.782 ±0.031	0.120 ±0.001	0.439 ±0.018	0.654 ±0.003
CLMBR-T (Wornow et al., 2023)	0.690 ±0.064	0.116 ±0.020	0.107 ±0.001	0.629 ±0.006	0.746 ±0.008	0.664 ±0.044	0.712 ±0.028	0.112 ±0.002	0.510 ±0.005	0.692 ±0.008
GARLIC (Wang et al., 2026)	0.317 ±0.004	0.263 ±0.059	0.100 ±0.001	0.572 ±0.027	0.469 ±0.015	0.514 ±0.023	0.465 ±0.004	0.111 ±0.001	0.463 ±0.018	0.660 ±0.008
KEDGN (Luo et al., 2024)	0.347 ±0.040	0.457 ±0.092	**0.085** ±0.002	0.540 ±0.021	0.449 ±0.049	0.697 ±0.059	1.476 ±0.539	**0.097** ±0.001	0.456 ±0.010	0.543 ±0.036
MIRA (Li et al., 2026)	0.795 ±0.002	**0.112** ±0.009	0.107 ±0.001	0.561 ±0.032	0.791 ±0.006	0.739 ±0.001	0.467 ±0.006	0.117 ±0.001	0.451 ±0.024	0.816 ±0.002
TabICL (Qu et al., 2025)	0.644 ±0.019	1.140 ±0.365	0.098 ±0.003	0.529 ±0.008	0.590 ±0.057	0.632 ±0.015	1.029 ±0.103	0.128 ±0.004	0.511 ±0.008	0.615 ±0.067
MEWS (Subbe et al., 2001)	0.815 ±0.012	0.125 ±0.006	0.111 ±0.001	0.519 ±0.017	0.760 ±0.005	0.806 ±0.002	0.474 ±0.005	0.121 ±0.000	0.453 ±0.017	0.752 ±0.003
SOFA (Vincent et al., 1996)	0.822 ±0.006	0.156 ±0.004	0.107 ±0.001	0.570 ±0.054	**0.861** ±0.002	0.828 ±0.002	0.510 ±0.005	0.121 ±0.000	0.509 ±0.010	**0.833** ±0.004
Markdown (Hegselmann et al., 2026)	**0.902** ±0.002	0.146 ±0.016	0.103 ±0.001	0.611 ±0.006	0.738 ±0.023	0.824 ±0.008	0.449 ±0.009	0.117 ±0.001	0.494 ±0.005	0.814 ±0.014
Record2Vec (Ji et al., 2026)	0.862 ±0.009	0.475 ±0.075	0.108 ±0.001	0.607 ±0.006	0.813 ±0.004	0.826 ±0.004	0.448 ±0.016	0.120 ±0.001	**0.511** ±0.006	0.775 ±0.007
Template (Demirel et al., 2026)	0.884 ±0.004	0.113 ±0.011	0.104 ±0.001	**0.630** ±0.003	0.669 ±0.028	0.821 ±0.005	**0.439** ±0.005	0.120 ±0.002	0.505 ±0.006	0.793 ±0.007
QA	0.796 ±0.015	0.356 ±0.059	0.108 ±0.001	0.564 ±0.007	0.784 ±0.004	0.816 ±0.001	0.486 ±0.023	0.118 ±0.001	0.494 ±0.002	0.766 ±0.007
Treatment	0.800 ±0.018	0.413 ±0.074	0.108 ±0.001	0.564 ±0.024	0.782 ±0.007	0.825 ±0.003	0.490 ±0.014	0.118 ±0.001	0.492 ±0.012	0.796 ±0.006
Organ-severity trajectory	0.812 ±0.010	0.711 ±0.171	0.110 ±0.002	0.583 ±0.008	0.772 ±0.011	0.830 ±0.003	0.498 ±0.037	0.121 ±0.001	0.491 ±0.005	0.807 ±0.008
PORTER	0.842 ±0.011	0.550 ±0.061	0.107 ±0.000	0.607 ±0.005	0.789 ±0.005	**0.858** ±0.002	0.453 ±0.010	0.120 ±0.001	0.501 ±0.003	0.820 ±0.004

All six directions			
Mean rank↓	Cumulative occurrences↑		
	Top-1	Top-2	Top-3
15 (12.5)	0	1	2
16 (13.5)	1	1	1
6 (7.4)	0	2	8
12 (10.1)	1	4	5
14 (12.0)	4	4	4
10 (8.6)	1	1	4
13 (11.0)	0	3	3
10 (8.6)	0	0	4
2 (5.2)	4	7	9
4 (5.9)	4	7	**12**
5 (6.0)	1	5	7
3 (5.7)	**7**	**12**	**12**
8 (8.2)	0	0	0
7 (7.6)	0	0	1
9 (8.5)	0	2	6
1 (5.1)	**7**	11	**12**

Table 17: Zero-shot transfer, all six directions, the Gemma-4 arm at the xLSTM-Mixer head. The directions are carried over three blocks of two. Cells are the mean over the five cross-validation folds with one standard deviation. The columns to the right of the first block summarize all six directions. *Mean rank* is each method's position among the sixteen by its average rank over the thirty columns, with that average in parentheses, and *Cumulative occurrences* counts the columns, out of thirty, in which a method places in the top-1, 2, 3. Groups, metrics and marking as in Table 1.

Method	HIRID → MIMIC					HIRID → PPICU					All six directions			
	mort AUROC↑	los MSE↓	forecast MSE↓	lab AUROC↑	drug AUROC↑	mort AUROC↑	los MSE↓	forecast MSE↓	lab AUROC↑	drug AUROC↑	Mean rank↓	Cumulative occurrences ↑ Top-1	Top-2	Top-3
ATTACK	0.498 ±0.009	3.550 ±0.020	0.264 ±0.099	0.376 ±0.004	0.659 ±0.002	0.500 ±0.001	10.111 ±0.098	0.269 ±0.112	0.484 ±0.004	0.634 ±0.000	14 (12.7)	0	0	0
RANDOM	0.500 ±0.010	3.541 ±0.077	0.410 ±0.176	0.384 ±0.014	0.659 ±0.004	0.500 ±0.012	9.824 ±0.232	0.451 ±0.193	0.487 ±0.012	0.632 ±0.003	16 (13.9)	0	0	0
CLMBR-T (Wornow et al., 2023)	0.754 ±0.008	3.044 ±0.108	1.054 ±0.954	0.486 ±0.019	0.765 ±0.009	0.661 ±0.006	7.398 ±0.203	0.972 ±0.895	0.585 ±0.008	0.656 ±0.007	9 (8.9)	0	3	6
GARLIC (Wang et al., 2026)	0.584 ±0.013	3.098 ±0.085	0.321 ±0.227	0.419 ±0.024	0.681 ±0.007	0.641 ±0.008	8.192 ±0.178	0.314 ±0.214	0.543 ±0.030	0.696 ±0.014	13 (11.9)	0	0	0
KEDGN (Luo et al., 2024)	0.300 ±0.035	4.081 ±0.627	0.447 ±0.257	0.407 ±0.045	0.375 ±0.077	0.660 ±0.043	14.418 ±2.387	0.459 ±0.290	0.513 ±0.027	0.416 ±0.013	15 (13.8)	0	0	0
MIRA (Li et al., 2026)	0.646 ±0.023	3.035 ±0.045	0.773 ±0.642	0.404 ±0.028	0.778 ±0.059	0.649 ±0.019	8.096 ±0.248	0.849 ±0.670	0.509 ±0.024	0.691 ±0.022	12 (10.2)	0	1	2
TabICL (Qu et al., 2025)	0.759 ±0.015	3.725 ±0.148	**0.157** ±0.018	0.430 ±0.015	0.690 ±0.009	0.746 ±0.022	10.281 ±0.430	0.160 ±0.015	0.537 ±0.025	0.711 ±0.009	10 (9.3)	1	2	4
MEWS (Subbe et al., 2001)	0.783 ±0.003	2.874 ±0.014	0.512 ±0.338	0.424 ±0.013	0.722 ±0.004	0.682 ±0.005	7.313 ±0.123	0.506 ±0.256	0.491 ±0.005	0.711 ±0.001	11 (9.4)	0	0	0
SOFA (Vincent et al., 1996)	0.801 ±0.008	2.873 ±0.042	0.346 ±0.349	0.465 ±0.016	0.849 ±0.006	0.729 ±0.005	7.180 ±0.065	0.382 ±0.422	0.537 ±0.007	**0.828** ±0.002	5 (5.7)	5	7	11
Markdown (Hegselmann et al., 2026)	0.767 ±0.010	**2.777** ±0.118	0.269 ±0.113	0.468 ±0.013	0.838 ±0.025	0.682 ±0.004	**6.709** ±0.451	0.289 ±0.155	0.578 ±0.009	0.577 ±0.022	4 (5.3)	5	7	12
Record2Vec (Ji et al., 2026)	0.804 ±0.011	2.978 ±0.044	0.236 ±0.105	0.450 ±0.006	0.838 ±0.004	0.738 ±0.012	8.130 ±0.218	0.202 ±0.079	0.559 ±0.007	0.750 ±0.002	3 (4.7)	1	7	11
Template (Demirel et al., 2026)	0.751 ±0.011	2.835 ±0.049	0.233 ±0.104	**0.494** ±0.013	0.897 ±0.007	0.736 ±0.002	7.176 ±0.276	0.266 ±0.149	**0.596** ±0.004	0.650 ±0.012	2 (4.5)	7	13	14
QA	0.796 ±0.005	3.122 ±0.056	0.230 ±0.110	0.448 ±0.017	0.809 ±0.004	0.734 ±0.007	8.689 ±0.211	0.224 ±0.116	0.536 ±0.013	0.722 ±0.005	6 (7.1)	1	3	5
Treatment	0.773 ±0.007	3.066 ±0.018	0.238 ±0.143	0.436 ±0.009	0.835 ±0.005	0.731 ±0.007	8.126 ±0.226	0.219 ±0.138	0.528 ±0.010	0.742 ±0.005	7 (7.5)	0	0	1
Organ-severity trajectory	0.797 ±0.005	3.275 ±0.263	4.862 ±2.447	0.436 ±0.010	0.863 ±0.002	0.763 ±0.006	9.373 ±0.836	3.897 ±2.116	0.567 ±0.014	0.739 ±0.007	8 (7.6)	0	3	8
PORTER	**0.820** ±0.004	2.861 ±0.039	0.158 ±0.032	0.444 ±0.005	**0.907** ±0.002	**0.767** ±0.005	8.541 ±0.158	**0.150** ±0.023	0.566 ±0.013	0.772 ±0.004	**1 (3.6)**	**10**	**14**	**16**

Method	MIMIC → HIRID					MIMIC → PPICU				
	mort AUROC↑	los MSE↓	forecast MSE↓	lab AUROC↑	drug AUROC↑	mort AUROC↑	los MSE↓	forecast MSE↓	lab AUROC↑	drug AUROC↑
ATTACK	0.498 ±0.008	0.747 ±0.004	0.160 ±0.048	0.387 ±0.012	0.697 ±0.002	0.500 ±0.000	9.859 ±0.113	0.192 ±0.067	0.488 ±0.015	0.634 ±0.000
RANDOM	0.503 ±0.031	0.791 ±0.019	0.470 ±0.287	0.411 ±0.003	0.692 ±0.006	0.503 ±0.009	9.646 ±0.135	0.443 ±0.259	0.473 ±0.004	0.631 ±0.004
CLMBR-T (Wornow et al., 2023)	0.777 ±0.017	0.564 ±0.055	0.484 ±0.204	0.524 ±0.039	0.799 ±0.004	0.695 ±0.010	7.853 ±0.404	0.426 ±0.183	0.571 ±0.020	0.656 ±0.017
GARLIC (Wang et al., 2026)	0.586 ±0.019	0.610 ±0.050	0.490 ±0.720	0.466 ±0.004	0.727 ±0.017	0.598 ±0.011	9.302 ±0.139	0.464 ±0.603	0.482 ±0.021	0.683 ±0.006
KEDGN (Luo et al., 2024)	0.613 ±0.018	1.082 ±0.195	0.647 ±0.397	0.455 ±0.020	0.572 ±0.064	0.511 ±0.039	7.919 ±0.215	0.619 ±0.391	0.512 ±0.014	0.661 ±0.024
MIRA (Li et al., 2026)	0.749 ±0.033	0.604 ±0.108	0.341 ±0.216	0.418 ±0.038	0.765 ±0.043	0.619 ±0.018	9.049 ±0.455	0.332 ±0.162	0.500 ±0.028	0.674 ±0.005
TabICL (Qu et al., 2025)	0.834 ±0.027	0.630 ±0.038	0.154 ±0.047	0.506 ±0.016	0.763 ±0.011	0.732 ±0.020	9.464 ±0.258	0.204 ±0.061	0.530 ±0.005	0.726 ±0.006
MEWS (Subbe et al., 2001)	0.831 ±0.009	0.552 ±0.022	0.427 ±0.311	0.462 ±0.007	0.748 ±0.006	0.713 ±0.006	7.885 ±0.113	0.473 ±0.346	0.516 ±0.011	0.704 ±0.002
SOFA (Vincent et al., 1996)	0.852 ±0.009	0.539 ±0.015	1.202 ±0.478	0.486 ±0.020	0.875 ±0.005	0.770 ±0.007	7.860 ±0.083	1.114 ±0.381	0.542 ±0.016	**0.844** ±0.004
Markdown (Hegselmann et al., 2026)	0.871 ±0.015	**0.399** ±0.015	0.135 ±0.019	0.543 ±0.010	0.838 ±0.008	0.671 ±0.012	6.557 ±0.366	0.167 ±0.017	0.558 ±0.006	0.631 ±0.018
Record2Vec (Ji et al., 2026)	0.884 ±0.001	0.540 ±0.038	0.160 ±0.038	0.495 ±0.008	0.831 ±0.003	0.769 ±0.004	8.287 ±0.439	0.164 ±0.018	0.542 ±0.007	0.742 ±0.002
Template (Demirel et al., 2026)	0.843 ±0.012	0.481 ±0.049	0.135 ±0.009	**0.558** ±0.015	0.849 ±0.008	0.713 ±0.012	**6.274** ±0.228	0.170 ±0.018	**0.573** ±0.009	0.713 ±0.014
QA	0.852 ±0.021	0.619 ±0.015	0.129 ±0.009	0.447 ±0.002	0.778 ±0.007	0.757 ±0.003	8.738 ±0.150	0.165 ±0.017	0.533 ±0.006	0.712 ±0.003
Treatment	0.833 ±0.011	0.599 ±0.034	0.186 ±0.092	0.457 ±0.007	0.835 ±0.006	0.764 ±0.004	8.859 ±0.297	0.246 ±0.119	0.531 ±0.003	0.742 ±0.002
Organ-severity trajectory	0.832 ±0.013	0.798 ±0.081	0.179 ±0.053	0.464 ±0.029	0.859 ±0.005	0.780 ±0.005	8.620 ±0.476	0.231 ±0.064	0.538 ±0.021	0.744 ±0.003
PORTER	**0.899** ±0.007	0.592 ±0.091	**0.129** ±0.009	0.514 ±0.013	**0.914** ±0.003	**0.797** ±0.005	8.285 ±0.297	**0.153** ±0.019	0.549 ±0.006	0.780 ±0.003

Method	PPICU → HIRID					PPICU → MIMIC				
	mort AUROC↑	los MSE↓	forecast MSE↓	lab AUROC↑	drug AUROC↑	mort AUROC↑	los MSE↓	forecast MSE↓	lab AUROC↑	drug AUROC↑
ATTACK	0.502 ±0.006	0.324 ±0.049	0.121 ±0.010	0.513 ±0.010	0.697 ±0.002	0.505 ±0.007	0.643 ±0.038	0.135 ±0.016	0.421 ±0.010	0.661 ±0.003
RANDOM	0.517 ±0.006	0.416 ±0.053	0.428 ±0.145	0.508 ±0.004	0.699 ±0.003	0.508 ±0.010	0.718 ±0.042	0.454 ±0.180	0.434 ±0.006	0.661 ±0.003
CLMBR-T (Wornow et al., 2023)	0.781 ±0.026	0.116 ±0.003	0.285 ±0.280	0.591 ±0.061	0.750 ±0.011	0.730 ±0.031	0.850 ±0.089	0.231 ±0.186	0.491 ±0.034	0.690 ±0.005
GARLIC (Wang et al., 2026)	0.316 ±0.012	0.270 ±0.028	0.698 ±1.056	0.531 ±0.046	0.654 ±0.024	0.520 ±0.014	0.493 ±0.024	0.512 ±0.686	0.467 ±0.011	0.719 ±0.008
KEDGN (Luo et al., 2024)	0.379 ±0.067	0.388 ±0.107	0.646 ±0.622	0.520 ±0.010	0.463 ±0.077	0.666 ±0.106	0.948 ±0.136	0.592 ±0.614	0.471 ±0.027	0.589 ±0.078
MIRA (Li et al., 2026)	0.794 ±0.025	0.110 ±0.008	0.147 ±0.035	0.521 ±0.035	0.792 ±0.008	0.714 ±0.035	0.484 ±0.010	0.179 ±0.035	0.453 ±0.013	0.807 ±0.003
TabICL (Qu et al., 2025)	0.650 ±0.031	0.873 ±0.215	0.161 ±0.038	0.533 ±0.013	0.697 ±0.058	0.648 ±0.028	0.797 ±0.097	0.154 ±0.018	0.512 ±0.015	0.659 ±0.037
MEWS (Subbe et al., 2001)	0.813 ±0.010	0.146 ±0.013	0.317 ±0.162	0.548 ±0.014	0.752 ±0.008	0.804 ±0.003	0.497 ±0.011	0.287 ±0.156	0.487 ±0.003	0.749 ±0.004
SOFA (Vincent et al., 1996)	0.817 ±0.010	0.178 ±0.013	0.259 ±0.174	0.613 ±0.010	**0.857** ±0.003	0.824 ±0.004	0.538 ±0.036	0.273 ±0.149	**0.526** ±0.005	**0.823** ±0.004
Markdown (Hegselmann et al., 2026)	**0.899** ±0.003	0.148 ±0.036	**0.106** ±0.006	0.611 ±0.002	0.751 ±0.023	0.819 ±0.009	0.462 ±0.019	0.126 ±0.008	0.496 ±0.003	0.812 ±0.008
Record2Vec (Ji et al., 2026)	0.861 ±0.005	0.439 ±0.061	0.112 ±0.001	0.606 ±0.005	0.813 ±0.006	0.823 ±0.007	**0.433** ±0.011	0.123 ±0.002	0.513 ±0.006	0.766 ±0.008
Template (Demirel et al., 2026)	0.890 ±0.004	**0.101** ±0.017	0.122 ±0.036	**0.629** ±0.010	0.657 ±0.021	0.813 ±0.006	0.441 ±0.011	0.128 ±0.009	0.510 ±0.008	0.807 ±0.005
QA	0.802 ±0.005	0.382 ±0.029	0.110 ±0.001	0.557 ±0.004	0.782 ±0.005	0.813 ±0.003	0.480 ±0.016	**0.121** ±0.002	0.494 ±0.003	0.763 ±0.011
Treatment	0.804 ±0.016	0.438 ±0.047	0.125 ±0.032	0.567 ±0.004	0.786 ±0.012	0.825 ±0.002	0.488 ±0.016	0.141 ±0.039	0.495 ±0.003	0.794 ±0.007
Organ-severity trajectory	0.816 ±0.007	0.420 ±0.099	0.393 ±0.328	0.587 ±0.005	0.772 ±0.008	0.832 ±0.005	0.445 ±0.022	0.291 ±0.190	0.499 ±0.005	0.816 ±0.006
PORTER	0.846 ±0.017	0.613 ±0.065	0.121 ±0.022	0.602 ±0.006	0.790 ±0.008	**0.859** ±0.002	0.448 ±0.009	0.124 ±0.004	0.501 ±0.005	0.819 ±0.008

Table 18: Zero-shot transfer, all six directions, the Qwen3.6 arm at the LSTM head. The directions are carried over three blocks of two. Cells are the mean over the five cross-validation folds with one standard deviation. The columns to the right of the first block summarize all six directions. *Mean rank* is each method's position among the sixteen by its average rank over the thirty columns, with that average in parentheses, and *Cumulative occurrences* counts the columns, out of thirty, in which a method places in the top-1, 2, 3. Groups, metrics and marking as in Table 1.

Method	HIRID → MIMIC					HIRID → PPICU				
	mort AUROC↑	los MSE↓	forecast MSE↓	lab AUROC↑	drug AUROC↑	mort AUROC↑	los MSE↓	forecast MSE↓	lab AUROC↑	drug AUROC↑
ATTACK	0.500 ±0.000	3.483 ±0.006	0.127 ±0.000	0.373 ±0.034	0.661 ±0.003	0.500 ±0.000	9.745 ±0.039	0.161 ±0.001	0.475 ±0.027	0.634 ±0.000
RANDOM	0.507 ±0.010	3.588 ±0.079	0.135 ±0.002	0.394 ±0.010	0.663 ±0.003	0.493 ±0.010	10.094 ±0.213	0.162 ±0.001	0.491 ±0.018	0.634 ±0.003
CLMBR-T (Wornow et al., 2023)	0.753 ±0.011	2.934 ±0.043	0.132 ±0.001	**0.468** ±0.020	0.772 ±0.006	0.662 ±0.008	7.758 ±0.485	0.156 ±0.002	0.574 ±0.009	0.663 ±0.007
GARLIC (Wang et al., 2026)	0.550 ±0.028	3.284 ±0.043	0.127 ±0.001	0.401 ±0.006	0.690 ±0.006	0.623 ±0.017	9.025 ±0.213	0.162 ±0.001	0.531 ±0.024	0.673 ±0.012
KEDGN (Luo et al., 2024)	0.304 ±0.024	3.663 ±0.166	**0.117** ±0.003	0.430 ±0.024	0.329 ±0.024	0.629 ±0.031	11.271 ±0.147	**0.138** ±0.003	0.446 ±0.020	0.430 ±0.017
MIRA (Li et al., 2026)	0.639 ±0.014	3.003 ±0.045	0.128 ±0.001	0.369 ±0.037	0.797 ±0.003	0.663 ±0.005	7.919 ±0.316	0.164 ±0.001	0.500 ±0.014	0.693 ±0.004
TabICL (Qu et al., 2025)	0.751 ±0.017	3.825 ±0.030	0.140 ±0.002	0.424 ±0.009	0.695 ±0.010	0.743 ±0.024	10.934 ±0.207	0.175 ±0.007	0.541 ±0.025	0.716 ±0.011
MEWS (Subbe et al., 2001)	0.784 ±0.003	2.856 ±0.052	0.131 ±0.001	0.419 ±0.011	0.729 ±0.003	0.679 ±0.006	7.596 ±0.268	0.156 ±0.001	0.513 ±0.010	0.711 ±0.002
SOFA (Vincent et al., 1996)	0.807 ±0.008	2.928 ±0.089	0.134 ±0.001	0.460 ±0.006	0.857 ±0.003	0.738 ±0.015	7.312 ±0.216	0.156 ±0.001	0.557 ±0.006	**0.832** ±0.004
Markdown (Hegselmann et al., 2026)	0.766 ±0.005	2.776 ±0.072	0.136 ±0.001	0.455 ±0.006	0.853 ±0.013	0.704 ±0.008	**6.694** ±0.156	0.155 ±0.001	0.578 ±0.005	0.613 ±0.033
Record2Vec (Ji et al., 2026)	0.819 ±0.003	2.950 ±0.037	0.135 ±0.001	0.424 ±0.032	0.860 ±0.004	0.759 ±0.002	8.663 ±0.193	0.157 ±0.002	0.540 ±0.015	0.769 ±0.002
Template (Demirel et al., 2026)	0.747 ±0.008	**2.675** ±0.162	0.135 ±0.001	0.467 ±0.013	0.895 ±0.004	0.734 ±0.002	7.556 ±0.278	0.156 ±0.001	**0.587** ±0.007	0.658 ±0.009
QA	0.793 ±0.005	3.284 ±0.064	0.133 ±0.001	0.428 ±0.006	0.821 ±0.006	0.735 ±0.009	8.875 ±0.107	0.157 ±0.002	0.528 ±0.001	0.739 ±0.004
Treatment	0.800 ±0.004	2.962 ±0.040	0.134 ±0.001	0.398 ±0.042	0.856 ±0.004	0.742 ±0.001	8.563 ±0.366	0.158 ±0.003	0.525 ±0.041	0.778 ±0.005
Organ-severity trajectory	0.814 ±0.010	3.325 ±0.048	0.136 ±0.001	0.382 ±0.029	0.871 ±0.008	0.720 ±0.009	10.862 ±0.168	0.159 ±0.001	0.503 ±0.028	0.790 ±0.003
PORTER	**0.825** ±0.004	2.775 ±0.041	0.136 ±0.001	0.411 ±0.031	**0.915** ±0.002	**0.768** ±0.005	8.515 ±0.201	0.155 ±0.001	0.546 ±0.018	0.788 ±0.002

Method	MIMIC → HIRID					MIMIC → PPICU				
	mort AUROC↑	los MSE↓	forecast MSE↓	lab AUROC↑	drug AUROC↑	mort AUROC↑	los MSE↓	forecast MSE↓	lab AUROC↑	drug AUROC↑
ATTACK	0.497 ±0.005	0.760 ±0.003	0.125 ±0.001	0.418 ±0.023	0.697 ±0.001	0.500 ±0.001	9.587 ±0.036	0.186 ±0.001	0.475 ±0.023	0.634 ±0.000
RANDOM	0.499 ±0.015	0.841 ±0.026	0.125 ±0.001	0.392 ±0.005	0.696 ±0.002	0.499 ±0.010	10.159 ±0.108	0.180 ±0.001	0.476 ±0.004	0.632 ±0.001
CLMBR-T (Wornow et al., 2023)	0.787 ±0.012	0.527 ±0.043	0.128 ±0.001	0.449 ±0.012	0.801 ±0.003	0.682 ±0.017	8.079 ±0.314	0.186 ±0.002	0.530 ±0.007	0.654 ±0.005
GARLIC (Wang et al., 2026)	0.598 ±0.007	0.651 ±0.073	**0.121** ±0.002	0.442 ±0.018	0.702 ±0.026	0.596 ±0.013	9.710 ±0.204	0.184 ±0.001	0.490 ±0.007	0.682 ±0.005
KEDGN (Luo et al., 2024)	0.605 ±0.048	0.903 ±0.172	0.130 ±0.002	0.440 ±0.021	0.605 ±0.029	0.511 ±0.050	9.771 ±0.722	**0.179** ±0.001	0.487 ±0.024	0.675 ±0.007
MIRA (Li et al., 2026)	0.749 ±0.010	0.488 ±0.050	0.127 ±0.001	0.394 ±0.020	0.787 ±0.003	0.593 ±0.002	8.663 ±0.303	0.186 ±0.001	0.486 ±0.027	0.674 ±0.004
TabICL (Qu et al., 2025)	0.823 ±0.018	0.618 ±0.070	0.124 ±0.003	0.458 ±0.016	0.768 ±0.010	0.719 ±0.017	10.306 ±0.244	0.183 ±0.002	0.514 ±0.006	0.730 ±0.008
MEWS (Subbe et al., 2001)	0.853 ±0.004	0.526 ±0.008	0.135 ±0.001	0.453 ±0.014	0.760 ±0.005	0.719 ±0.005	8.377 ±0.246	0.190 ±0.001	0.520 ±0.010	0.712 ±0.003
SOFA (Vincent et al., 1996)	0.854 ±0.005	0.561 ±0.007	0.132 ±0.001	**0.473** ±0.007	0.875 ±0.002	0.763 ±0.009	8.309 ±0.165	0.188 ±0.001	**0.540** ±0.012	**0.846** ±0.001
Markdown (Hegselmann et al., 2026)	0.867 ±0.012	**0.389** ±0.008	0.139 ±0.001	0.438 ±0.040	0.822 ±0.014	0.672 ±0.007	6.665 ±0.268	0.194 ±0.002	0.486 ±0.040	0.624 ±0.016
Record2Vec (Ji et al., 2026)	0.894 ±0.002	0.516 ±0.009	0.137 ±0.001	0.420 ±0.024	0.846 ±0.004	0.777 ±0.003	8.892 ±0.173	0.189 ±0.002	0.474 ±0.022	0.764 ±0.003
Template (Demirel et al., 2026)	0.799 ±0.027	0.451 ±0.025	0.140 ±0.003	0.412 ±0.024	0.819 ±0.027	0.704 ±0.003	**6.389** ±0.097	0.195 ±0.002	0.476 ±0.024	0.711 ±0.014
QA	0.874 ±0.008	0.609 ±0.012	0.132 ±0.002	0.418 ±0.025	0.801 ±0.007	0.753 ±0.005	9.056 ±0.226	0.187 ±0.003	0.475 ±0.025	0.731 ±0.006
Treatment	0.875 ±0.002	0.602 ±0.014	0.132 ±0.002	0.420 ±0.025	0.856 ±0.005	0.768 ±0.002	8.998 ±0.205	0.186 ±0.003	0.474 ±0.025	0.797 ±0.003
Organ-severity trajectory	0.876 ±0.002	0.802 ±0.106	0.135 ±0.002	0.432 ±0.052	0.880 ±0.004	0.759 ±0.004	8.838 ±0.392	0.189 ±0.001	0.494 ±0.031	0.788 ±0.002
PORTER	**0.897** ±0.006	0.500 ±0.024	0.136 ±0.003	0.418 ±0.025	**0.914** ±0.004	**0.791** ±0.006	8.158 ±0.254	0.190 ±0.002	0.472 ±0.026	0.794 ±0.002

Method	PPICU → HIRID					PPICU → MIMIC				
	mort AUROC↑	los MSE↓	forecast MSE↓	lab AUROC↑	drug AUROC↑	mort AUROC↑	los MSE↓	forecast MSE↓	lab AUROC↑	drug AUROC↑
ATTACK	0.496 ±0.004	0.416 ±0.011	0.105 ±0.001	0.499 ±0.013	0.696 ±0.002	0.495 ±0.004	0.713 ±0.009	0.115 ±0.001	0.416 ±0.026	0.661 ±0.002
RANDOM	0.515 ±0.012	0.449 ±0.012	0.107 ±0.001	0.487 ±0.016	0.698 ±0.002	0.507 ±0.008	0.743 ±0.011	0.116 ±0.001	0.425 ±0.031	0.660 ±0.003
CLMBR-T (Wornow et al., 2023)	0.665 ±0.040	**0.092** ±0.008	0.108 ±0.001	**0.616** ±0.010	0.748 ±0.013	0.675 ±0.027	0.647 ±0.061	0.116 ±0.001	0.502 ±0.008	0.679 ±0.009
GARLIC (Wang et al., 2026)	0.314 ±0.010	0.257 ±0.054	0.101 ±0.001	0.572 ±0.032	0.485 ±0.052	0.509 ±0.024	0.494 ±0.021	0.114 ±0.001	0.472 ±0.010	0.670 ±0.026
KEDGN (Luo et al., 2024)	0.317 ±0.074	0.129 ±0.033	**0.091** ±0.001	0.530 ±0.029	0.513 ±0.064	0.658 ±0.081	0.528 ±0.012	**0.103** ±0.001	0.449 ±0.016	0.580 ±0.066
MIRA (Li et al., 2026)	0.793 ±0.017	0.109 ±0.019	0.105 ±0.001	0.534 ±0.026	0.788 ±0.007	0.741 ±0.005	0.485 ±0.015	0.114 ±0.001	0.434 ±0.021	0.812 ±0.003
TabICL (Qu et al., 2025)	0.668 ±0.049	0.863 ±0.610	0.096 ±0.002	0.509 ±0.009	0.716 ±0.038	0.694 ±0.019	0.690 ±0.117	0.120 ±0.002	0.472 ±0.034	0.660 ±0.030
MEWS (Subbe et al., 2001)	0.829 ±0.009	0.104 ±0.006	0.109 ±0.001	0.505 ±0.029	0.760 ±0.010	0.804 ±0.005	0.458 ±0.005	0.121 ±0.000	0.450 ±0.034	0.753 ±0.004
SOFA (Vincent et al., 1996)	0.807 ±0.015	0.117 ±0.006	0.108 ±0.001	0.571 ±0.050	**0.869** ±0.003	0.827 ±0.003	0.459 ±0.020	0.121 ±0.000	0.475 ±0.045	**0.837** ±0.002
Markdown (Hegselmann et al., 2026)	**0.900** ±0.003	0.138 ±0.008	0.106 ±0.001	0.548 ±0.054	0.737 ±0.029	0.816 ±0.009	0.445 ±0.011	0.122 ±0.003	0.450 ±0.039	0.830 ±0.003
Record2Vec (Ji et al., 2026)	0.885 ±0.003	0.345 ±0.040	0.109 ±0.001	0.594 ±0.005	0.810 ±0.006	0.849 ±0.002	0.440 ±0.014	0.122 ±0.001	**0.503** ±0.007	0.812 ±0.008
Template (Demirel et al., 2026)	0.876 ±0.020	0.106 ±0.017	0.106 ±0.000	0.599 ±0.043	0.706 ±0.047	0.818 ±0.004	0.436 ±0.008	0.124 ±0.001	0.486 ±0.031	0.820 ±0.008
QA	0.858 ±0.004	0.318 ±0.064	0.109 ±0.001	0.564 ±0.027	0.802 ±0.004	0.813 ±0.002	0.511 ±0.022	0.117 ±0.001	0.481 ±0.027	0.793 ±0.007
Treatment	0.851 ±0.004	0.397 ±0.039	0.109 ±0.001	0.569 ±0.029	0.819 ±0.005	0.811 ±0.003	0.426 ±0.016	0.120 ±0.001	0.483 ±0.027	0.820 ±0.007
Organ-severity trajectory	0.860 ±0.004	0.675 ±0.092	0.109 ±0.001	0.564 ±0.036	0.859 ±0.007	0.846 ±0.004	0.495 ±0.005	0.121 ±0.001	0.468 ±0.044	0.836 ±0.006
PORTER	0.894 ±0.003	0.430 ±0.130	0.110 ±0.001	0.584 ±0.035	0.816 ±0.003	**0.860** ±0.002	**0.425** ±0.008	0.123 ±0.002	0.488 ±0.030	0.835 ±0.006

All six directions			
Mean rank↓	Cumulative occurrences ↑		
	Top-1	Top-2	Top-3
15 (12.4)	0	0	1
16 (12.9)	0	1	2
5 (7.2)	3	5	7
12 (9.7)	1	2	4
14 (10.8)	5	5	5
11 (9.1)	0	1	2
13 (9.7)	0	3	5
9 (8.2)	0	1	3
1 (5.0)	6	7	9
7 (7.5)	3	6	7
3 (6.7)	1	6	8
4 (7.2)	3	7	10
10 (8.2)	0	0	0
6 (7.3)	0	2	4
8 (8.1)	0	4	8
2 (6.0)	**8**	**10**	**15**

Table 19: Zero-shot transfer, all six directions, the Qwen3.6 arm at the linear head. The directions are carried over three blocks of two. Cells are the mean over the five cross-validation folds with one standard deviation. The columns to the right of the first block summarize all six directions. *Mean rank* is each method's position among the sixteen by its average rank over the thirty columns, with that average in parentheses, and *Cumulative occurrences* counts the columns, out of thirty, in which a method places in the top-1, 2, 3. Groups, metrics and marking as in Table 1.

Method	HIRID → MIMIC					HIRID → PPICU				
	mort AUROC↑	los MSE↓	forecast MSE↓	lab AUROC↑	drug AUROC↑	mort AUROC↑	los MSE↓	forecast MSE↓	lab AUROC↑	drug AUROC↑
ATTACK	0.502 ±0.003	3.527 ±0.021	0.127 ±0.000	0.390 ±0.032	0.662 ±0.003	0.500 ±0.001	9.993 ±0.104	0.162 ±0.001	0.499 ±0.023	0.634 ±0.000
RANDOM	0.517 ±0.008	3.516 ±0.023	0.140 ±0.001	0.384 ±0.007	0.663 ±0.002	0.504 ±0.015	9.816 ±0.040	0.166 ±0.002	0.493 ±0.014	0.632 ±0.002
CLMBR-T (Wornow et al., 2023)	0.761 ±0.007	2.952 ±0.022	0.130 ±0.002	0.477 ±0.018	0.760 ±0.010	0.670 ±0.010	7.777 ±0.250	0.159 ±0.003	0.565 ±0.019	0.658 ±0.009
GARLIC (Wang et al., 2026)	0.550 ±0.041	3.158 ±0.080	0.126 ±0.002	0.399 ±0.028	0.673 ±0.009	0.621 ±0.023	8.509 ±0.175	0.163 ±0.002	0.549 ±0.011	0.650 ±0.012
KEDGN (Luo et al., 2024)	0.281 ±0.015	4.373 ±0.692	**0.110** ±0.005	0.397 ±0.011	0.371 ±0.048	0.639 ±0.027	16.610 ±1.327	**0.138** ±0.008	0.516 ±0.024	0.462 ±0.029
MIRA (Li et al., 2026)	0.658 ±0.010	2.974 ±0.030	0.128 ±0.001	0.426 ±0.031	0.806 ±0.004	0.669 ±0.005	7.781 ±0.150	0.161 ±0.002	0.499 ±0.040	0.700 ±0.003
TabICL (Qu et al., 2025)	0.768 ±0.006	3.566 ±0.108	0.146 ±0.003	0.419 ±0.009	0.679 ±0.004	0.758 ±0.007	9.316 ±0.356	0.191 ±0.006	0.549 ±0.011	0.700 ±0.008
MEWS (Subbe et al., 2001)	0.783 ±0.005	2.853 ±0.016	0.129 ±0.001	0.424 ±0.020	0.729 ±0.002	0.679 ±0.004	7.240 ±0.034	0.155 ±0.002	0.498 ±0.013	0.709 ±0.001
SOFA (Vincent et al., 1996)	0.804 ±0.003	2.865 ±0.011	0.133 ±0.001	0.443 ±0.025	0.855 ±0.004	0.733 ±0.007	7.045 ±0.082	0.155 ±0.002	0.533 ±0.009	**0.824** ±0.003
Markdown (Hegselmann et al., 2026)	0.769 ±0.009	2.758 ±0.057	0.136 ±0.003	0.462 ±0.012	0.854 ±0.012	0.680 ±0.006	**6.632** ±0.181	0.159 ±0.004	0.574 ±0.007	0.571 ±0.019
Record2Vec (Ji et al., 2026)	0.820 ±0.006	2.941 ±0.039	0.136 ±0.001	0.442 ±0.004	0.865 ±0.002	0.758 ±0.006	8.535 ±0.200	0.156 ±0.001	0.556 ±0.009	0.768 ±0.004
Template (Demirel et al., 2026)	0.757 ±0.005	2.766 ±0.079	0.136 ±0.002	**0.493** ±0.021	0.904 ±0.004	0.736 ±0.003	7.487 ±0.195	0.160 ±0.005	**0.593** ±0.005	0.673 ±0.012
QA	0.796 ±0.001	3.224 ±0.023	0.135 ±0.001	0.435 ±0.009	0.821 ±0.005	0.733 ±0.007	8.487 ±0.206	0.160 ±0.001	0.533 ±0.007	0.741 ±0.005
Treatment	0.805 ±0.010	2.910 ±0.036	0.137 ±0.000	0.428 ±0.037	0.858 ±0.004	0.739 ±0.013	8.203 ±0.218	0.159 ±0.001	0.542 ±0.009	0.777 ±0.004
Organ-severity trajectory	0.813 ±0.008	3.332 ±0.089	0.137 ±0.002	0.432 ±0.018	0.874 ±0.007	0.717 ±0.007	10.774 ±0.305	0.164 ±0.002	0.558 ±0.024	0.790 ±0.007
PORTER	**0.826** ±0.003	**2.749** ±0.029	0.141 ±0.001	0.435 ±0.008	**0.914** ±0.002	**0.766** ±0.005	8.336 ±0.052	0.159 ±0.002	0.571 ±0.006	0.781 ±0.005

Method	MIMIC → HIRID					MIMIC → PPICU				
	mort AUROC↑	los MSE↓	forecast MSE↓	lab AUROC↑	drug AUROC↑	mort AUROC↑	los MSE↓	forecast MSE↓	lab AUROC↑	drug AUROC↑
ATTACK	0.500 ±0.001	0.749 ±0.003	0.125 ±0.000	0.408 ±0.026	0.697 ±0.001	0.500 ±0.000	9.863 ±0.189	0.186 ±0.001	0.470 ±0.038	0.634 ±0.000
RANDOM	0.505 ±0.014	0.807 ±0.004	0.128 ±0.001	0.409 ±0.008	0.695 ±0.002	0.494 ±0.008	9.588 ±0.067	**0.179** ±0.004	0.473 ±0.004	0.629 ±0.001
CLMBR-T (Wornow et al., 2023)	0.787 ±0.009	0.524 ±0.032	0.131 ±0.003	0.496 ±0.013	0.792 ±0.004	0.669 ±0.019	7.692 ±0.510	0.183 ±0.002	0.546 ±0.015	0.635 ±0.010
GARLIC (Wang et al., 2026)	0.595 ±0.015	0.591 ±0.020	**0.120** ±0.002	0.457 ±0.021	0.688 ±0.017	0.595 ±0.010	9.381 ±0.199	0.182 ±0.002	0.501 ±0.019	0.685 ±0.004
KEDGN (Luo et al., 2024)	0.626 ±0.022	1.064 ±0.175	0.143 ±0.003	0.452 ±0.019	0.597 ±0.057	0.517 ±0.042	8.034 ±0.231	0.190 ±0.006	0.516 ±0.009	0.655 ±0.030
MIRA (Li et al., 2026)	0.750 ±0.011	0.517 ±0.086	0.126 ±0.002	0.439 ±0.031	0.788 ±0.004	0.599 ±0.006	8.704 ±0.295	0.186 ±0.004	0.474 ±0.032	0.676 ±0.007
TabICL (Qu et al., 2025)	0.822 ±0.017	0.646 ±0.033	0.135 ±0.003	0.501 ±0.012	0.759 ±0.011	0.692 ±0.034	8.943 ±0.548	0.206 ±0.004	0.519 ±0.002	0.716 ±0.012
MEWS (Subbe et al., 2001)	0.840 ±0.007	0.509 ±0.017	0.136 ±0.001	0.457 ±0.017	0.749 ±0.004	0.723 ±0.003	7.846 ±0.099	0.188 ±0.001	0.506 ±0.022	0.708 ±0.002
SOFA (Vincent et al., 1996)	0.848 ±0.009	0.518 ±0.010	0.131 ±0.000	0.485 ±0.002	0.880 ±0.001	0.762 ±0.004	7.785 ±0.094	0.187 ±0.002	0.545 ±0.005	**0.843** ±0.002
Markdown (Hegselmann et al., 2026)	0.874 ±0.007	**0.376** ±0.005	0.137 ±0.003	0.538 ±0.014	0.839 ±0.012	0.680 ±0.012	6.752 ±0.290	0.195 ±0.006	0.551 ±0.013	0.638 ±0.024
Record2Vec (Ji et al., 2026)	0.891 ±0.003	0.530 ±0.019	0.133 ±0.002	0.515 ±0.006	0.854 ±0.003	0.776 ±0.005	8.373 ±0.195	0.186 ±0.003	0.541 ±0.004	0.759 ±0.003
Template (Demirel et al., 2026)	0.833 ±0.013	0.481 ±0.013	0.138 ±0.005	**0.559** ±0.015	0.856 ±0.006	0.717 ±0.009	**6.208** ±0.112	0.199 ±0.008	**0.567** ±0.013	0.727 ±0.019
QA	0.875 ±0.005	0.596 ±0.006	0.129 ±0.001	0.474 ±0.010	0.802 ±0.005	0.752 ±0.004	8.696 ±0.092	0.184 ±0.001	0.528 ±0.006	0.727 ±0.003
Treatment	0.877 ±0.004	0.623 ±0.021	0.132 ±0.002	0.472 ±0.006	0.857 ±0.004	0.767 ±0.004	8.778 ±0.153	0.186 ±0.004	0.535 ±0.012	0.793 ±0.004
Organ-severity trajectory	0.874 ±0.003	0.693 ±0.085	0.137 ±0.002	0.464 ±0.033	0.882 ±0.004	0.758 ±0.003	8.965 ±0.123	0.193 ±0.002	0.518 ±0.040	0.790 ±0.003
PORTER	**0.900** ±0.003	0.499 ±0.025	0.135 ±0.002	0.517 ±0.011	**0.915** ±0.002	**0.790** ±0.004	8.246 ±0.172	0.185 ±0.004	0.545 ±0.009	0.792 ±0.002

Method	PPICU → HIRID					PPICU → MIMIC				
	mort AUROC↑	los MSE↓	forecast MSE↓	lab AUROC↑	drug AUROC↑	mort AUROC↑	los MSE↓	forecast MSE↓	lab AUROC↑	drug AUROC↑
ATTACK	0.499 ±0.003	0.354 ±0.018	0.105 ±0.000	0.501 ±0.013	0.698 ±0.001	0.498 ±0.003	0.665 ±0.013	0.115 ±0.000	0.427 ±0.025	0.661 ±0.002
RANDOM	0.512 ±0.015	0.483 ±0.037	0.108 ±0.001	0.509 ±0.018	0.696 ±0.004	0.506 ±0.005	0.782 ±0.031	0.120 ±0.001	0.439 ±0.018	0.654 ±0.003
CLMBR-T (Wornow et al., 2023)	0.690 ±0.064	0.116 ±0.020	0.107 ±0.001	0.629 ±0.006	0.746 ±0.008	0.664 ±0.044	0.712 ±0.028	0.112 ±0.002	0.510 ±0.005	0.692 ±0.008
GARLIC (Wang et al., 2026)	0.317 ±0.004	0.263 ±0.059	0.100 ±0.001	0.572 ±0.027	0.469 ±0.015	0.514 ±0.023	0.465 ±0.004	0.111 ±0.001	0.463 ±0.018	0.660 ±0.008
KEDGN (Luo et al., 2024)	0.347 ±0.040	0.457 ±0.092	**0.085** ±0.002	0.540 ±0.021	0.449 ±0.049	0.697 ±0.059	1.476 ±0.539	**0.097** ±0.001	0.456 ±0.010	0.543 ±0.036
MIRA (Li et al., 2026)	0.795 ±0.002	**0.112** ±0.009	0.107 ±0.001	0.561 ±0.032	0.791 ±0.006	0.739 ±0.001	0.467 ±0.006	0.117 ±0.001	0.451 ±0.024	0.816 ±0.002
TabICL (Qu et al., 2025)	0.644 ±0.019	1.140 ±0.365	0.098 ±0.003	0.529 ±0.008	0.590 ±0.057	0.632 ±0.015	1.029 ±0.103	0.128 ±0.004	0.511 ±0.008	0.615 ±0.067
MEWS (Subbe et al., 2001)	0.815 ±0.012	0.125 ±0.006	0.111 ±0.001	0.519 ±0.017	0.760 ±0.005	0.806 ±0.002	0.474 ±0.005	0.121 ±0.000	0.453 ±0.017	0.752 ±0.003
SOFA (Vincent et al., 1996)	0.822 ±0.006	0.156 ±0.004	0.107 ±0.001	0.570 ±0.054	**0.861** ±0.002	0.828 ±0.002	0.510 ±0.005	0.121 ±0.000	0.509 ±0.010	**0.833** ±0.004
Markdown (Hegselmann et al., 2026)	**0.902** ±0.002	0.146 ±0.016	0.103 ±0.001	0.611 ±0.006	0.738 ±0.023	0.824 ±0.008	0.449 ±0.009	0.117 ±0.001	0.494 ±0.005	0.814 ±0.014
Record2Vec (Ji et al., 2026)	0.886 ±0.003	0.363 ±0.059	0.109 ±0.000	0.608 ±0.008	0.805 ±0.004	0.848 ±0.003	0.448 ±0.013	0.121 ±0.001	0.512 ±0.003	0.813 ±0.005
Template (Demirel et al., 2026)	0.884 ±0.004	0.113 ±0.011	0.104 ±0.001	**0.630** ±0.003	0.669 ±0.028	0.821 ±0.005	**0.439** ±0.005	0.120 ±0.002	0.505 ±0.006	0.793 ±0.007
QA	0.857 ±0.004	0.270 ±0.042	0.108 ±0.000	0.573 ±0.030	0.796 ±0.003	0.813 ±0.003	0.499 ±0.013	0.116 ±0.001	0.493 ±0.013	0.793 ±0.005
Treatment	0.842 ±0.006	0.432 ±0.069	0.109 ±0.001	0.569 ±0.026	0.818 ±0.003	0.812 ±0.007	0.441 ±0.014	0.119 ±0.001	0.492 ±0.012	0.818 ±0.006
Organ-severity trajectory	0.861 ±0.005	0.821 ±0.194	0.111 ±0.001	0.597 ±0.006	0.860 ±0.008	0.843 ±0.003	0.501 ±0.027	0.121 ±0.001	**0.512** ±0.006	0.829 ±0.009
PORTER	0.895 ±0.008	0.447 ±0.088	0.107 ±0.001	0.612 ±0.006	0.807 ±0.002	**0.859** ±0.005	0.442 ±0.018	0.121 ±0.001	0.509 ±0.005	0.822 ±0.003

All six directions			
Mean rank↓	Cumulative occurrences↑		
	Top-1	Top-2	Top-3
15 (12.6)	0	1	2
16 (13.4)	1	1	1
7 (7.5)	0	2	7
12 (10.2)	1	4	5
14 (12.1)	4	4	4
11 (9.1)	1	1	2
13 (11.0)	0	1	3
10 (9.0)	0	0	2
2 (5.6)	4	6	7
5 (6.3)	3	8	9
4 (5.9)	0	6	7
3 (5.8)	7	**10**	11
8 (7.6)	0	0	0
6 (7.0)	0	2	5
9 (8.2)	1	5	8
1 (4.7)	**8**	9	**17**

Table 20: Zero-shot transfer, all six directions, the Qwen3.6 arm at the xLSTM-Mixer head. The directions are carried over three blocks of two. Cells are the mean over the five cross-validation folds with one standard deviation. The columns to the right of the first block summarize all six directions. *Mean rank* is each method's position among the sixteen by its average rank over the thirty columns, with that average in parentheses, and *Cumulative occurrences* counts the columns, out of thirty, in which a method places in the top-1, 2, 3. Groups, metrics and marking as in Table 1.

Method	HIRID → MIMIC					HIRID → PPICU					All six directions			
	mort AUROC↑	los MSE↓	forecast MSE↓	lab AUROC↑	drug AUROC↑	mort AUROC↑	los MSE↓	forecast MSE↓	lab AUROC↑	drug AUROC↑	Mean rank↓	Cumulative occurrences ↑ Top-1	Top-2	Top-3
ATTACK	0.498 ±0.009	3.550 ±0.020	0.264 ±0.099	0.376 ±0.004	0.659 ±0.002	0.500 ±0.001	10.111 ±0.098	0.269 ±0.112	0.484 ±0.004	0.634 ±0.000	14 (12.8)	0	0	0
RANDOM	0.500 ±0.010	3.541 ±0.077	0.410 ±0.176	0.384 ±0.014	0.659 ±0.004	0.500 ±0.012	9.824 ±0.232	0.451 ±0.193	0.487 ±0.012	0.632 ±0.003	15 (13.9)	0	0	0
CLMBR-T (Wornow et al., 2023)	0.754 ±0.008	3.044 ±0.108	1.054 ±0.954	0.486 ±0.019	0.765 ±0.009	0.661 ±0.006	7.398 ±0.203	0.972 ±0.895	0.585 ±0.008	0.656 ±0.007	9 (9.1)	0	3	6
GARLIC (Wang et al., 2026)	0.584 ±0.013	3.098 ±0.085	0.321 ±0.227	0.419 ±0.024	0.681 ±0.007	0.641 ±0.008	8.192 ±0.178	0.314 ±0.214	0.543 ±0.030	0.696 ±0.014	13 (11.8)	0	0	0
KEDGN (Luo et al., 2024)	0.300 ±0.035	4.081 ±0.627	0.447 ±0.257	0.407 ±0.045	0.375 ±0.077	0.660 ±0.043	14.418 ±2.387	0.459 ±0.290	0.513 ±0.027	0.416 ±0.013	15 (13.9)	0	0	0
MIRA (Li et al., 2026)	0.646 ±0.023	3.035 ±0.045	0.773 ±0.642	0.404 ±0.028	0.778 ±0.059	0.649 ±0.019	8.096 ±0.248	0.849 ±0.670	0.509 ±0.024	0.691 ±0.022	12 (10.4)	0	1	1
TabICL (Qu et al., 2025)	0.759 ±0.015	3.725 ±0.148	**0.157** ±0.018	0.430 ±0.015	0.690 ±0.009	0.746 ±0.022	10.281 ±0.430	**0.160** ±0.015	0.537 ±0.025	0.711 ±0.009	10 (9.6)	2	2	4
MEWS (Subbe et al., 2001)	0.783 ±0.003	2.874 ±0.014	0.512 ±0.338	0.424 ±0.013	0.722 ±0.004	0.682 ±0.005	7.313 ±0.123	0.506 ±0.256	0.491 ±0.005	0.711 ±0.001	11 (9.6)	0	0	0
SOFA (Vincent et al., 1996)	0.801 ±0.008	2.873 ±0.042	0.346 ±0.349	0.465 ±0.016	0.849 ±0.006	0.729 ±0.005	7.180 ±0.065	0.382 ±0.422	0.537 ±0.007	**0.828** ±0.002	6 (6.3)	3	5	9
Markdown (Hegselmann et al., 2026)	0.767 ±0.010	2.777 ±0.118	0.269 ±0.113	0.468 ±0.013	0.838 ±0.025	0.682 ±0.004	**6.709** ±0.451	0.289 ±0.155	0.578 ±0.009	0.577 ±0.022	4 (5.7)	4	8	12
Record2Vec (Ji et al., 2026)	0.819 ±0.007	2.963 ±0.044	0.206 ±0.049	0.452 ±0.006	0.861 ±0.003	0.755 ±0.011	8.714 ±0.179	0.199 ±0.032	0.551 ±0.005	0.767 ±0.003	3 (4.9)	0	6	7
Template (Demirel et al., 2026)	0.751 ±0.011	2.835 ±0.049	0.233 ±0.104	**0.494** ±0.013	0.897 ±0.007	0.736 ±0.002	7.176 ±0.276	0.266 ±0.149	**0.596** ±0.004	0.650 ±0.012	2 (4.8)	7	10	13
QA	0.793 ±0.006	3.223 ±0.019	0.180 ±0.055	0.443 ±0.009	0.822 ±0.003	0.733 ±0.006	8.575 ±0.070	0.163 ±0.029	0.528 ±0.005	0.740 ±0.006	7 (6.5)	1	5	5
Treatment	0.803 ±0.008	2.944 ±0.069	0.187 ±0.061	0.443 ±0.008	0.857 ±0.004	0.736 ±0.008	8.636 ±0.228	0.178 ±0.041	0.543 ±0.015	0.777 ±0.002	5 (6.0)	1	3	6
Organ-severity trajectory	0.814 ±0.013	3.382 ±0.178	4.815 ±2.445	0.456 ±0.012	0.874 ±0.007	0.723 ±0.003	10.787 ±0.096	3.875 ±2.093	0.544 ±0.034	0.791 ±0.006	8 (7.3)	2	5	8
PORTER	**0.825** ±0.004	**2.773** ±0.012	0.225 ±0.116	0.447 ±0.011	**0.915** ±0.002	**0.765** ±0.003	8.613 ±0.056	0.215 ±0.101	0.576 ±0.012	0.779 ±0.008	**1 (3.4)**	**10**	**12**	**19**

Method	MIMIC → HIRID					MIMIC → PPICU				
	mort AUROC↑	los MSE↓	forecast MSE↓	lab AUROC↑	drug AUROC↑	mort AUROC↑	los MSE↓	forecast MSE↓	lab AUROC↑	drug AUROC↑
ATTACK	0.498 ±0.008	0.747 ±0.004	0.160 ±0.048	0.387 ±0.012	0.697 ±0.002	0.500 ±0.000	9.859 ±0.113	0.192 ±0.067	0.488 ±0.015	0.634 ±0.000
RANDOM	0.503 ±0.031	0.791 ±0.019	0.470 ±0.287	0.411 ±0.003	0.692 ±0.006	0.503 ±0.009	9.646 ±0.135	0.443 ±0.259	0.473 ±0.004	0.631 ±0.004
CLMBR-T (Wornow et al., 2023)	0.777 ±0.017	0.564 ±0.055	0.484 ±0.204	0.524 ±0.039	0.799 ±0.004	0.695 ±0.010	7.853 ±0.404	0.426 ±0.183	0.571 ±0.020	0.656 ±0.017
GARLIC (Wang et al., 2026)	0.586 ±0.019	0.610 ±0.050	0.490 ±0.720	0.466 ±0.004	0.727 ±0.017	0.598 ±0.011	9.302 ±0.139	0.464 ±0.603	0.482 ±0.021	0.683 ±0.006
KEDGN (Luo et al., 2024)	0.613 ±0.018	1.082 ±0.195	0.647 ±0.397	0.455 ±0.020	0.572 ±0.064	0.511 ±0.039	7.919 ±0.215	0.619 ±0.391	0.512 ±0.014	0.661 ±0.024
MIRA (Li et al., 2026)	0.749 ±0.033	0.604 ±0.108	0.341 ±0.216	0.418 ±0.038	0.765 ±0.043	0.619 ±0.018	9.049 ±0.455	0.332 ±0.162	0.500 ±0.028	0.674 ±0.005
TabICL (Qu et al., 2025)	0.834 ±0.027	0.630 ±0.038	0.154 ±0.047	0.506 ±0.016	0.763 ±0.011	0.732 ±0.020	9.464 ±0.258	0.204 ±0.061	0.530 ±0.005	0.726 ±0.006
MEWS (Subbe et al., 2001)	0.831 ±0.009	0.552 ±0.022	0.427 ±0.311	0.462 ±0.007	0.748 ±0.006	0.713 ±0.006	7.885 ±0.113	0.473 ±0.346	0.516 ±0.011	0.704 ±0.002
SOFA (Vincent et al., 1996)	0.852 ±0.009	0.539 ±0.015	1.202 ±0.478	0.486 ±0.020	0.875 ±0.005	0.770 ±0.007	7.860 ±0.083	1.114 ±0.381	0.542 ±0.016	**0.844** ±0.004
Markdown (Hegselmann et al., 2026)	0.871 ±0.015	**0.399** ±0.015	0.135 ±0.019	0.543 ±0.010	0.838 ±0.008	0.671 ±0.012	6.557 ±0.366	0.167 ±0.017	0.558 ±0.006	0.631 ±0.018
Record2Vec (Ji et al., 2026)	0.891 ±0.002	0.539 ±0.025	0.129 ±0.009	0.510 ±0.008	0.853 ±0.003	0.776 ±0.003	8.272 ±0.284	0.170 ±0.011	0.534 ±0.008	0.761 ±0.003
Template (Demirel et al., 2026)	0.843 ±0.012	0.481 ±0.049	0.135 ±0.009	**0.558** ±0.015	0.849 ±0.008	0.713 ±0.012	**6.274** ±0.228	0.170 ±0.018	**0.573** ±0.009	0.713 ±0.014
QA	0.876 ±0.005	0.605 ±0.014	0.131 ±0.009	0.482 ±0.010	0.800 ±0.005	0.752 ±0.004	8.677 ±0.229	0.163 ±0.012	0.533 ±0.010	0.728 ±0.004
Treatment	0.878 ±0.004	0.634 ±0.044	0.128 ±0.004	0.477 ±0.011	0.855 ±0.006	0.768 ±0.002	8.880 ±0.309	0.173 ±0.010	0.536 ±0.005	0.795 ±0.002
Organ-severity trajectory	0.873 ±0.004	0.670 ±0.023	0.570 ±0.663	0.513 ±0.041	0.881 ±0.005	0.760 ±0.002	8.923 ±0.195	0.385 ±0.365	0.529 ±0.041	0.785 ±0.005
PORTER	**0.901** ±0.002	0.525 ±0.040	**0.126** ±0.010	0.514 ±0.010	**0.916** ±0.001	**0.792** ±0.003	8.093 ±0.269	**0.160** ±0.020	0.547 ±0.005	0.791 ±0.005

Method	PPICU → HIRID					PPICU → MIMIC				
	mort AUROC↑	los MSE↓	forecast MSE↓	lab AUROC↑	drug AUROC↑	mort AUROC↑	los MSE↓	forecast MSE↓	lab AUROC↑	drug AUROC↑
ATTACK	0.502 ±0.006	0.324 ±0.049	0.121 ±0.010	0.513 ±0.010	0.697 ±0.002	0.505 ±0.007	0.643 ±0.038	0.135 ±0.016	0.421 ±0.010	0.661 ±0.003
RANDOM	0.517 ±0.006	0.416 ±0.053	0.428 ±0.145	0.508 ±0.004	0.699 ±0.003	0.508 ±0.010	0.718 ±0.042	0.454 ±0.180	0.434 ±0.006	0.661 ±0.003
CLMBR-T (Wornow et al., 2023)	0.781 ±0.026	0.116 ±0.003	0.285 ±0.280	0.591 ±0.061	0.750 ±0.011	0.730 ±0.031	0.850 ±0.089	0.231 ±0.186	0.491 ±0.034	0.690 ±0.005
GARLIC (Wang et al., 2026)	0.316 ±0.012	0.270 ±0.028	0.698 ±1.056	0.531 ±0.046	0.654 ±0.024	0.520 ±0.014	0.493 ±0.024	0.512 ±0.686	0.467 ±0.011	0.719 ±0.008
KEDGN (Luo et al., 2024)	0.379 ±0.067	0.388 ±0.107	0.646 ±0.622	0.520 ±0.010	0.463 ±0.077	0.666 ±0.106	0.948 ±0.136	0.592 ±0.614	0.471 ±0.027	0.589 ±0.078
MIRA (Li et al., 2026)	0.794 ±0.025	0.110 ±0.008	0.147 ±0.035	0.521 ±0.035	0.792 ±0.008	0.714 ±0.035	0.484 ±0.010	0.179 ±0.035	0.453 ±0.013	0.807 ±0.003
TabICL (Qu et al., 2025)	0.650 ±0.031	0.873 ±0.215	0.161 ±0.038	0.533 ±0.013	0.697 ±0.058	0.648 ±0.028	0.797 ±0.097	0.154 ±0.018	0.512 ±0.015	0.659 ±0.037
MEWS (Subbe et al., 2001)	0.813 ±0.010	0.146 ±0.013	0.317 ±0.162	0.548 ±0.014	0.752 ±0.008	0.804 ±0.003	0.497 ±0.011	0.287 ±0.156	0.487 ±0.003	0.749 ±0.004
SOFA (Vincent et al., 1996)	0.817 ±0.010	0.178 ±0.013	0.259 ±0.174	0.613 ±0.010	0.857 ±0.003	0.824 ±0.004	0.538 ±0.036	0.273 ±0.149	**0.526** ±0.005	0.823 ±0.004
Markdown (Hegselmann et al., 2026)	**0.899** ±0.003	0.148 ±0.036	**0.106** ±0.006	0.611 ±0.002	0.751 ±0.023	0.819 ±0.009	0.462 ±0.019	0.126 ±0.008	0.496 ±0.003	0.812 ±0.008
Record2Vec (Ji et al., 2026)	0.887 ±0.002	0.358 ±0.054	0.124 ±0.025	0.602 ±0.005	0.810 ±0.007	0.850 ±0.001	0.438 ±0.017	0.134 ±0.021	0.504 ±0.005	0.814 ±0.005
Template (Demirel et al., 2026)	0.890 ±0.004	**0.101** ±0.017	0.122 ±0.036	**0.629** ±0.010	0.657 ±0.021	0.813 ±0.006	0.441 ±0.011	0.128 ±0.009	0.510 ±0.008	0.807 ±0.005
QA	0.863 ±0.006	0.275 ±0.050	0.117 ±0.004	0.583 ±0.005	0.799 ±0.009	0.813 ±0.005	0.498 ±0.018	**0.122** ±0.004	0.500 ±0.004	0.792 ±0.008
Treatment	0.851 ±0.007	0.421 ±0.047	0.135 ±0.046	0.583 ±0.004	0.815 ±0.003	0.813 ±0.003	**0.426** ±0.010	0.132 ±0.017	0.497 ±0.005	0.814 ±0.003
Organ-severity trajectory	0.860 ±0.004	0.548 ±0.074	0.136 ±0.023	0.595 ±0.006	**0.861** ±0.004	0.843 ±0.002	0.457 ±0.025	0.143 ±0.017	0.517 ±0.012	**0.831** ±0.005
PORTER	0.894 ±0.005	0.479 ±0.064	0.120 ±0.014	0.612 ±0.008	0.817 ±0.006	**0.863** ±0.001	0.438 ±0.011	0.130 ±0.013	0.508 ±0.009	0.828 ±0.006

Table 21: Zero-shot transfer, all six directions, the Muse arm at the LSTM head. The directions are carried over three blocks of two. Cells are the mean over the five cross-validation folds with one standard deviation. The columns to the right of the first block summarize all six directions. *Mean rank* is each method's position among the sixteen by its average rank over the thirty columns, with that average in parentheses, and *Cumulative occurrences* counts the columns, out of thirty, in which a method places in the top-1, 2, 3. Groups, metrics and marking as in Table 1.

Method	HIRID → MIMIC					HIRID → PPICU					All six directions			
	mort AUROC↑	los MSE↓	forecast MSE↓	lab AUROC↑	drug AUROC↑	mort AUROC↑	los MSE↓	forecast MSE↓	lab AUROC↑	drug AUROC↑	Mean rank↓	Cumulative occurrences ↑ Top-1	Top-2	Top-3
ATTACK	0.498 ±0.009	3.483 ±0.006	0.127 ±0.001	0.373 ±0.034	0.661 ±0.001	0.499 ±0.000	9.746 ±0.039	0.161 ±0.001	0.475 ±0.027	0.633 ±0.000	15 (12.3)	0	0	2
RANDOM	0.507 ±0.010	3.588 ±0.079	0.135 ±0.002	0.394 ±0.010	0.663 ±0.003	0.493 ±0.010	10.094 ±0.213	0.162 ±0.001	0.491 ±0.018	0.634 ±0.003	16 (12.9)	0	1	2
CLMBR-T (Wornow et al., 2023)	0.753 ±0.011	2.934 ±0.043	0.132 ±0.001	**0.468** ±0.020	0.772 ±0.006	0.662 ±0.008	7.758 ±0.485	0.156 ±0.002	0.574 ±0.009	0.663 ±0.007	6 (7.3)	4	5	7
GARLIC (Wang et al., 2026)	0.550 ±0.028	3.284 ±0.043	0.127 ±0.001	0.401 ±0.006	0.690 ±0.006	0.623 ±0.017	9.025 ±0.213	0.162 ±0.001	0.531 ±0.024	0.673 ±0.012	13 (9.9)	1	2	3
KEDGN (Luo et al., 2024)	0.304 ±0.024	3.663 ±0.166	**0.117** ±0.003	0.430 ±0.024	0.329 ±0.024	0.629 ±0.031	11.271 ±0.147	**0.138** ±0.003	0.446 ±0.020	0.430 ±0.017	14 (10.8)	5	5	5
MIRA (Li et al., 2026)	0.639 ±0.014	3.003 ±0.045	0.128 ±0.001	0.369 ±0.037	0.797 ±0.003	0.663 ±0.005	7.919 ±0.316	0.164 ±0.001	0.500 ±0.014	0.693 ±0.004	11 (9.0)	0	1	1
TabICL (Qu et al., 2025)	0.751 ±0.017	3.825 ±0.030	0.140 ±0.002	0.424 ±0.009	0.695 ±0.010	0.743 ±0.024	10.934 ±0.207	0.175 ±0.007	0.541 ±0.025	0.716 ±0.011	12 (9.8)	0	3	4
MEWS (Subbe et al., 2001)	0.784 ±0.003	2.856 ±0.052	0.131 ±0.001	0.419 ±0.011	0.729 ±0.003	0.679 ±0.006	7.596 ±0.268	0.156 ±0.001	0.513 ±0.010	0.711 ±0.002	9 (8.2)	0	1	4
SOFA (Vincent et al., 1996)	0.807 ±0.008	2.928 ±0.089	0.134 ±0.001	0.460 ±0.006	0.857 ±0.003	0.738 ±0.015	7.312 ±0.216	0.156 ±0.001	0.557 ±0.006	**0.832** ±0.004	**1 (5.3)**	6	8	**11**
Markdown (Hegselmann et al., 2026)	0.766 ±0.005	2.776 ±0.072	0.136 ±0.001	0.455 ±0.006	0.853 ±0.013	0.704 ±0.008	**6.694** ±0.156	0.155 ±0.001	0.578 ±0.005	0.613 ±0.033	7 (7.3)	3	**9**	10
Record2Vec (Ji et al., 2026)	0.791 ±0.008	3.001 ±0.039	0.133 ±0.001	0.422 ±0.038	0.845 ±0.003	0.762 ±0.003	8.506 ±0.156	0.158 ±0.002	0.559 ±0.026	0.750 ±0.004	4 (7.0)	0	4	6
Template (Demirel et al., 2026)	0.747 ±0.008	**2.675** ±0.162	0.135 ±0.001	0.467 ±0.013	0.895 ±0.004	0.734 ±0.002	7.556 ±0.278	0.156 ±0.001	**0.587** ±0.007	0.658 ±0.009	5 (7.2)	3	6	**11**
QA	0.790 ±0.010	3.217 ±0.055	0.131 ±0.001	0.436 ±0.005	0.832 ±0.005	0.747 ±0.003	8.421 ±0.170	0.156 ±0.001	0.546 ±0.006	0.723 ±0.003	8 (8.0)	0	0	1
Treatment	0.793 ±0.009	3.091 ±0.027	0.134 ±0.002	0.412 ±0.034	0.865 ±0.005	0.741 ±0.006	7.771 ±0.097	0.158 ±0.001	0.528 ±0.048	0.780 ±0.005	3 (6.7)	0	2	4
Organ-severity trajectory	0.807 ±0.004	3.220 ±0.039	0.134 ±0.001	0.385 ±0.025	0.865 ±0.002	0.725 ±0.007	10.213 ±0.152	0.162 ±0.001	0.482 ±0.032	0.776 ±0.005	10 (8.4)	1	4	9
PORTER	**0.809** ±0.009	2.878 ±0.032	0.134 ±0.001	0.414 ±0.033	**0.913** ±0.002	**0.764** ±0.004	8.227 ±0.138	0.158 ±0.002	0.553 ±0.023	0.748 ±0.010	2 (5.9)	**7**	**9**	10

Method	MIMIC → HIRID					MIMIC → PPICU				
	mort AUROC↑	los MSE↓	forecast MSE↓	lab AUROC↑	drug AUROC↑	mort AUROC↑	los MSE↓	forecast MSE↓	lab AUROC↑	drug AUROC↑
ATTACK	0.498 ±0.006	0.760 ±0.003	0.125 ±0.001	0.418 ±0.023	0.697 ±0.001	0.500 ±0.001	9.587 ±0.036	0.186 ±0.001	0.475 ±0.023	0.633 ±0.000
RANDOM	0.499 ±0.015	0.841 ±0.026	0.125 ±0.001	0.392 ±0.005	0.696 ±0.002	0.499 ±0.010	10.159 ±0.108	0.180 ±0.001	0.476 ±0.004	0.632 ±0.001
CLMBR-T (Wornow et al., 2023)	0.787 ±0.012	0.527 ±0.043	0.128 ±0.001	0.449 ±0.012	0.801 ±0.003	0.682 ±0.017	8.079 ±0.314	0.186 ±0.002	0.530 ±0.007	0.654 ±0.005
GARLIC (Wang et al., 2026)	0.598 ±0.007	0.651 ±0.073	**0.121** ±0.002	0.442 ±0.018	0.702 ±0.026	0.596 ±0.013	9.710 ±0.204	0.184 ±0.001	0.490 ±0.007	0.682 ±0.005
KEDGN (Luo et al., 2024)	0.605 ±0.048	0.903 ±0.172	0.130 ±0.002	0.440 ±0.021	0.605 ±0.029	0.511 ±0.050	9.771 ±0.722	**0.179** ±0.001	0.487 ±0.024	0.675 ±0.007
MIRA (Li et al., 2026)	0.749 ±0.010	0.488 ±0.050	0.127 ±0.001	0.394 ±0.020	0.787 ±0.003	0.593 ±0.002	8.663 ±0.303	0.186 ±0.001	0.486 ±0.027	0.674 ±0.004
TabICL (Qu et al., 2025)	0.823 ±0.018	0.618 ±0.070	0.124 ±0.003	0.458 ±0.016	0.768 ±0.010	0.719 ±0.017	10.306 ±0.244	0.183 ±0.002	0.514 ±0.006	0.730 ±0.008
MEWS (Subbe et al., 2001)	0.853 ±0.004	0.526 ±0.008	0.135 ±0.001	0.453 ±0.014	0.760 ±0.005	0.719 ±0.005	8.377 ±0.246	0.190 ±0.001	0.520 ±0.010	0.712 ±0.003
SOFA (Vincent et al., 1996)	0.854 ±0.005	0.561 ±0.007	0.132 ±0.001	**0.473** ±0.007	0.875 ±0.002	0.763 ±0.009	8.309 ±0.165	0.188 ±0.001	**0.540** ±0.012	**0.846** ±0.001
Markdown (Hegselmann et al., 2026)	0.867 ±0.012	**0.389** ±0.008	0.139 ±0.001	0.438 ±0.040	0.822 ±0.014	0.672 ±0.007	6.665 ±0.268	0.194 ±0.002	0.486 ±0.040	0.624 ±0.016
Record2Vec (Ji et al., 2026)	0.866 ±0.011	0.486 ±0.038	0.134 ±0.002	0.414 ±0.026	0.855 ±0.002	0.764 ±0.005	8.834 ±0.195	0.188 ±0.002	0.475 ±0.023	0.748 ±0.001
Template (Demirel et al., 2026)	0.799 ±0.027	0.451 ±0.025	0.140 ±0.003	0.412 ±0.024	0.819 ±0.027	0.704 ±0.003	**6.389** ±0.097	0.195 ±0.002	0.476 ±0.024	0.711 ±0.014
QA	0.845 ±0.010	0.584 ±0.015	0.131 ±0.001	0.418 ±0.024	0.790 ±0.008	0.737 ±0.007	9.009 ±0.178	0.187 ±0.002	0.473 ±0.025	0.727 ±0.002
Treatment	0.854 ±0.003	0.484 ±0.011	0.134 ±0.001	0.418 ±0.025	0.861 ±0.002	0.760 ±0.009	8.581 ±0.212	0.187 ±0.002	0.475 ±0.025	0.789 ±0.003
Organ-severity trajectory	0.858 ±0.006	0.825 ±0.118	0.139 ±0.002	0.415 ±0.028	0.863 ±0.002	0.773 ±0.003	9.004 ±0.383	0.188 ±0.002	0.481 ±0.018	0.782 ±0.002
PORTER	**0.880** ±0.006	0.441 ±0.017	0.133 ±0.002	0.416 ±0.025	**0.904** ±0.003	**0.796** ±0.010	8.126 ±0.204	0.188 ±0.003	0.474 ±0.027	0.780 ±0.001

Method	PPICU → HIRID					PPICU → MIMIC				
	mort AUROC↑	los MSE↓	forecast MSE↓	lab AUROC↑	drug AUROC↑	mort AUROC↑	los MSE↓	forecast MSE↓	lab AUROC↑	drug AUROC↑
ATTACK	0.502 ±0.000	0.415 ±0.009	0.105 ±0.001	0.499 ±0.013	0.697 ±0.003	0.501 ±0.000	0.712 ±0.008	0.114 ±0.001	0.416 ±0.026	0.660 ±0.002
RANDOM	0.515 ±0.012	0.449 ±0.012	0.107 ±0.001	0.487 ±0.016	0.698 ±0.002	0.507 ±0.008	0.743 ±0.011	0.116 ±0.001	0.425 ±0.031	0.660 ±0.003
CLMBR-T (Wornow et al., 2023)	0.665 ±0.040	**0.092** ±0.008	0.108 ±0.001	**0.616** ±0.010	0.748 ±0.013	0.675 ±0.027	0.647 ±0.061	0.116 ±0.001	**0.502** ±0.008	0.679 ±0.009
GARLIC (Wang et al., 2026)	0.314 ±0.010	0.257 ±0.054	0.101 ±0.001	0.572 ±0.032	0.485 ±0.052	0.509 ±0.024	0.494 ±0.021	0.114 ±0.001	0.472 ±0.010	0.670 ±0.026
KEDGN (Luo et al., 2024)	0.317 ±0.074	0.129 ±0.033	**0.091** ±0.001	0.530 ±0.029	0.513 ±0.064	0.658 ±0.081	0.528 ±0.012	**0.103** ±0.001	0.449 ±0.016	0.580 ±0.066
MIRA (Li et al., 2026)	0.793 ±0.017	0.109 ±0.019	0.105 ±0.001	0.534 ±0.026	0.788 ±0.007	0.741 ±0.005	0.485 ±0.015	0.114 ±0.001	0.434 ±0.021	0.812 ±0.003
TabICL (Qu et al., 2025)	0.668 ±0.049	0.863 ±0.610	0.096 ±0.002	0.509 ±0.009	0.716 ±0.038	0.694 ±0.019	0.690 ±0.117	0.120 ±0.002	0.472 ±0.034	0.660 ±0.030
MEWS (Subbe et al., 2001)	0.829 ±0.009	0.104 ±0.006	0.109 ±0.001	0.505 ±0.029	0.760 ±0.010	0.804 ±0.005	0.458 ±0.005	0.121 ±0.000	0.450 ±0.034	0.753 ±0.004
SOFA (Vincent et al., 1996)	0.807 ±0.015	0.117 ±0.006	0.108 ±0.001	0.571 ±0.050	**0.869** ±0.003	0.827 ±0.003	0.459 ±0.020	0.121 ±0.000	0.475 ±0.045	**0.837** ±0.002
Markdown (Hegselmann et al., 2026)	**0.900** ±0.003	0.138 ±0.008	0.106 ±0.001	0.548 ±0.054	0.737 ±0.029	0.816 ±0.009	0.445 ±0.011	0.122 ±0.003	0.450 ±0.039	0.830 ±0.003
Record2Vec (Ji et al., 2026)	0.875 ±0.005	0.377 ±0.068	0.109 ±0.001	0.609 ±0.003	0.787 ±0.004	0.839 ±0.001	0.454 ±0.022	0.121 ±0.001	0.502 ±0.005	0.774 ±0.008
Template (Demirel et al., 2026)	0.876 ±0.020	0.106 ±0.017	0.106 ±0.000	0.599 ±0.043	0.706 ±0.047	0.818 ±0.004	0.436 ±0.008	0.124 ±0.001	0.486 ±0.031	0.820 ±0.008
QA	0.837 ±0.008	0.269 ±0.079	0.110 ±0.001	0.562 ±0.038	0.780 ±0.006	0.816 ±0.001	0.478 ±0.031	0.118 ±0.001	0.475 ±0.021	0.762 ±0.007
Treatment	0.830 ±0.009	0.224 ±0.035	0.108 ±0.001	0.575 ±0.031	0.792 ±0.007	0.824 ±0.002	0.457 ±0.025	0.119 ±0.001	0.483 ±0.028	0.829 ±0.009
Organ-severity trajectory	0.868 ±0.006	0.656 ±0.090	0.111 ±0.001	0.578 ±0.022	0.825 ±0.005	**0.843** ±0.007	0.531 ±0.012	0.121 ±0.001	0.487 ±0.031	0.805 ±0.004
PORTER	0.888 ±0.006	0.301 ±0.090	0.107 ±0.001	0.588 ±0.039	0.784 ±0.019	0.839 ±0.003	**0.413** ±0.018	0.123 ±0.001	0.485 ±0.031	0.807 ±0.005

Table 22: Zero-shot transfer, all six directions, the Muse arm at the linear head. The directions are carried over three blocks of two. Cells are the mean over the five cross-validation folds with one standard deviation. The columns to the right of the first block summarize all six directions. *Mean rank* is each method's position among the sixteen by its average rank over the thirty columns, with that average in parentheses, and *Cumulative occurrences* counts the columns, out of thirty, in which a method places in the top-1, 2, 3. Groups, metrics and marking as in Table 1.

Method	HIRID → MIMIC					HIRID → PPICU				
	mort AUROC↑	los MSE↓	forecast MSE↓	lab AUROC↑	drug AUROC↑	mort AUROC↑	los MSE↓	forecast MSE↓	lab AUROC↑	drug AUROC↑
ATTACK	0.500 ±0.001	3.527 ±0.021	0.127 ±0.000	0.390 ±0.032	0.661 ±0.001	0.500 ±0.001	9.992 ±0.103	0.162 ±0.001	0.500 ±0.023	0.633 ±0.000
RANDOM	0.517 ±0.008	3.516 ±0.023	0.140 ±0.001	0.384 ±0.007	0.663 ±0.002	0.504 ±0.015	9.816 ±0.040	0.166 ±0.002	0.493 ±0.014	0.632 ±0.002
CLMBR-T (Wornow et al., 2023)	0.761 ±0.007	2.952 ±0.022	0.130 ±0.002	0.477 ±0.018	0.760 ±0.010	0.670 ±0.010	7.777 ±0.250	0.159 ±0.003	0.565 ±0.019	0.658 ±0.009
GARLIC (Wang et al., 2026)	0.550 ±0.041	3.158 ±0.080	0.126 ±0.002	0.399 ±0.028	0.673 ±0.009	0.621 ±0.023	8.509 ±0.175	0.163 ±0.002	0.549 ±0.011	0.650 ±0.012
KEDGN (Luo et al., 2024)	0.281 ±0.015	4.373 ±0.692	**0.110** ±0.005	0.397 ±0.011	0.371 ±0.048	0.639 ±0.027	16.610 ±1.327	**0.138** ±0.008	0.516 ±0.024	0.462 ±0.029
MIRA (Li et al., 2026)	0.658 ±0.010	2.974 ±0.030	0.128 ±0.001	0.426 ±0.031	0.806 ±0.004	0.669 ±0.005	7.781 ±0.150	0.161 ±0.002	0.499 ±0.040	0.700 ±0.003
TabICL (Qu et al., 2025)	0.768 ±0.006	3.566 ±0.108	0.146 ±0.003	0.419 ±0.009	0.679 ±0.004	0.758 ±0.007	9.316 ±0.356	0.191 ±0.006	0.549 ±0.011	0.700 ±0.008
MEWS (Subbe et al., 2001)	0.783 ±0.005	2.853 ±0.016	0.129 ±0.001	0.424 ±0.020	0.729 ±0.002	0.679 ±0.004	7.240 ±0.034	0.155 ±0.002	0.498 ±0.013	0.709 ±0.001
SOFA (Vincent et al., 1996)	0.804 ±0.003	2.865 ±0.011	0.133 ±0.001	0.443 ±0.025	0.855 ±0.004	0.733 ±0.007	7.045 ±0.082	0.155 ±0.002	0.533 ±0.009	**0.824** ±0.003
Markdown (Hegselmann et al., 2026)	0.769 ±0.009	**2.758** ±0.057	0.136 ±0.003	0.462 ±0.012	0.854 ±0.012	0.680 ±0.006	**6.632** ±0.181	0.159 ±0.004	0.574 ±0.007	0.571 ±0.019
Record2Vec (Ji et al., 2026)	0.797 ±0.010	2.964 ±0.014	0.136 ±0.001	0.430 ±0.037	0.840 ±0.006	0.761 ±0.006	8.134 ±0.142	0.160 ±0.002	0.566 ±0.013	0.748 ±0.003
Template (Demirel et al., 2026)	0.757 ±0.005	2.766 ±0.079	0.136 ±0.002	**0.493** ±0.021	0.904 ±0.004	0.736 ±0.003	7.487 ±0.195	0.160 ±0.005	**0.593** ±0.005	0.673 ±0.012
QA	0.789 ±0.005	3.213 ±0.063	0.132 ±0.000	0.418 ±0.033	0.829 ±0.008	0.745 ±0.004	8.040 ±0.105	0.158 ±0.003	0.534 ±0.005	0.723 ±0.002
Treatment	0.787 ±0.010	3.099 ±0.062	0.135 ±0.001	0.429 ±0.037	0.864 ±0.005	0.729 ±0.006	7.621 ±0.132	0.159 ±0.002	0.552 ±0.009	0.777 ±0.004
Organ-severity trajectory	**0.808** ±0.009	3.185 ±0.055	0.135 ±0.001	0.431 ±0.021	0.870 ±0.006	0.728 ±0.009	10.156 ±0.202	0.162 ±0.002	0.551 ±0.018	0.777 ±0.004
PORTER	0.800 ±0.010	2.848 ±0.019	0.139 ±0.001	0.425 ±0.035	**0.914** ±0.001	**0.762** ±0.008	7.968 ±0.195	0.159 ±0.002	0.571 ±0.014	0.755 ±0.007

Method	MIMIC → HIRID					MIMIC → PPICU				
	mort AUROC↑	los MSE↓	forecast MSE↓	lab AUROC↑	drug AUROC↑	mort AUROC↑	los MSE↓	forecast MSE↓	lab AUROC↑	drug AUROC↑
ATTACK	0.498 ±0.003	0.748 ±0.003	0.125 ±0.000	0.408 ±0.026	0.696 ±0.001	0.500 ±0.001	9.863 ±0.188	0.186 ±0.001	0.470 ±0.038	0.633 ±0.000
RANDOM	0.505 ±0.014	0.807 ±0.004	0.128 ±0.001	0.409 ±0.008	0.695 ±0.002	0.494 ±0.008	9.588 ±0.067	**0.179** ±0.004	0.473 ±0.004	0.629 ±0.001
CLMBR-T (Wornow et al., 2023)	0.787 ±0.009	0.524 ±0.032	0.131 ±0.003	0.496 ±0.013	0.792 ±0.004	0.669 ±0.019	7.692 ±0.510	0.183 ±0.002	0.546 ±0.015	0.635 ±0.010
GARLIC (Wang et al., 2026)	0.595 ±0.015	0.591 ±0.020	**0.120** ±0.002	0.457 ±0.021	0.688 ±0.017	0.595 ±0.010	9.381 ±0.199	0.182 ±0.002	0.501 ±0.019	0.685 ±0.004
KEDGN (Luo et al., 2024)	0.626 ±0.022	1.064 ±0.175	0.143 ±0.003	0.452 ±0.019	0.597 ±0.057	0.517 ±0.042	8.034 ±0.231	0.190 ±0.006	0.516 ±0.009	0.655 ±0.030
MIRA (Li et al., 2026)	0.750 ±0.011	0.517 ±0.086	0.126 ±0.002	0.439 ±0.031	0.788 ±0.004	0.599 ±0.006	8.704 ±0.295	0.186 ±0.004	0.474 ±0.032	0.676 ±0.007
TabICL (Qu et al., 2025)	0.822 ±0.017	0.646 ±0.033	0.135 ±0.003	0.501 ±0.012	0.759 ±0.011	0.692 ±0.034	8.943 ±0.548	0.206 ±0.004	0.519 ±0.002	0.716 ±0.012
MEWS (Subbe et al., 2001)	0.840 ±0.007	0.509 ±0.017	0.136 ±0.001	0.457 ±0.017	0.749 ±0.004	0.723 ±0.003	7.846 ±0.099	0.188 ±0.001	0.506 ±0.022	0.708 ±0.002
SOFA (Vincent et al., 1996)	0.848 ±0.009	0.518 ±0.010	0.131 ±0.000	0.485 ±0.002	0.880 ±0.001	0.762 ±0.004	7.785 ±0.094	0.187 ±0.002	0.545 ±0.005	**0.843** ±0.002
Markdown (Hegselmann et al., 2026)	0.874 ±0.007	**0.376** ±0.005	0.137 ±0.003	0.538 ±0.014	0.839 ±0.012	0.680 ±0.012	6.752 ±0.290	0.195 ±0.006	0.551 ±0.013	0.638 ±0.024
Record2Vec (Ji et al., 2026)	0.859 ±0.021	0.482 ±0.022	0.130 ±0.002	0.534 ±0.004	0.854 ±0.003	0.758 ±0.004	8.457 ±0.141	0.182 ±0.004	0.558 ±0.006	0.747 ±0.001
Template (Demirel et al., 2026)	0.833 ±0.013	0.481 ±0.013	0.138 ±0.005	**0.559** ±0.015	0.856 ±0.006	0.717 ±0.009	**6.208** ±0.112	0.199 ±0.008	**0.567** ±0.013	0.727 ±0.019
QA	0.849 ±0.011	0.595 ±0.009	0.128 ±0.001	0.478 ±0.005	0.791 ±0.004	0.734 ±0.006	8.450 ±0.137	0.185 ±0.003	0.525 ±0.009	0.724 ±0.004
Treatment	0.857 ±0.006	0.490 ±0.014	0.132 ±0.002	0.490 ±0.004	0.850 ±0.004	0.758 ±0.006	8.062 ±0.143	0.187 ±0.004	0.533 ±0.007	0.781 ±0.004
Organ-severity trajectory	0.856 ±0.005	0.648 ±0.054	0.139 ±0.002	0.465 ±0.037	0.864 ±0.004	0.775 ±0.006	9.549 ±0.336	0.193 ±0.002	0.520 ±0.035	0.782 ±0.002
PORTER	**0.880** ±0.010	0.472 ±0.020	0.134 ±0.002	0.526 ±0.013	**0.907** ±0.003	**0.789** ±0.013	7.957 ±0.116	0.184 ±0.004	0.554 ±0.009	0.776 ±0.002

Method	PPICU → HIRID					PPICU → MIMIC				
	mort AUROC↑	los MSE↓	forecast MSE↓	lab AUROC↑	drug AUROC↑	mort AUROC↑	los MSE↓	forecast MSE↓	lab AUROC↑	drug AUROC↑
ATTACK	0.503 ±0.004	0.354 ±0.016	0.105 ±0.000	0.501 ±0.013	0.698 ±0.002	0.503 ±0.004	0.664 ±0.012	0.115 ±0.000	0.427 ±0.025	0.662 ±0.001
RANDOM	0.512 ±0.015	0.483 ±0.037	0.108 ±0.001	0.509 ±0.018	0.696 ±0.004	0.506 ±0.005	0.782 ±0.031	0.120 ±0.001	0.439 ±0.018	0.654 ±0.003
CLMBR-T (Wornow et al., 2023)	0.690 ±0.064	0.116 ±0.020	0.107 ±0.001	0.629 ±0.006	0.746 ±0.008	0.664 ±0.044	0.712 ±0.028	0.112 ±0.002	0.510 ±0.005	0.692 ±0.008
GARLIC (Wang et al., 2026)	0.317 ±0.004	0.263 ±0.059	0.100 ±0.001	0.572 ±0.027	0.469 ±0.015	0.514 ±0.023	0.465 ±0.004	0.111 ±0.001	0.463 ±0.018	0.660 ±0.008
KEDGN (Luo et al., 2024)	0.347 ±0.040	0.457 ±0.092	**0.085** ±0.002	0.540 ±0.021	0.449 ±0.049	0.697 ±0.059	1.476 ±0.539	**0.097** ±0.001	0.456 ±0.010	0.543 ±0.036
MIRA (Li et al., 2026)	0.795 ±0.002	**0.112** ±0.009	0.107 ±0.001	0.561 ±0.032	0.791 ±0.006	0.739 ±0.001	0.467 ±0.006	0.117 ±0.001	0.451 ±0.024	0.816 ±0.002
TabICL (Qu et al., 2025)	0.644 ±0.019	1.140 ±0.365	0.098 ±0.003	0.529 ±0.008	0.590 ±0.057	0.632 ±0.015	1.029 ±0.103	0.128 ±0.004	**0.511** ±0.008	0.615 ±0.067
MEWS (Subbe et al., 2001)	0.815 ±0.012	0.125 ±0.006	0.111 ±0.001	0.519 ±0.017	0.760 ±0.005	0.806 ±0.002	0.474 ±0.005	0.121 ±0.000	0.453 ±0.017	0.752 ±0.003
SOFA (Vincent et al., 1996)	0.822 ±0.006	0.156 ±0.004	0.107 ±0.001	0.570 ±0.054	**0.861** ±0.002	0.828 ±0.002	0.510 ±0.005	0.121 ±0.000	0.509 ±0.010	**0.833** ±0.004
Markdown (Hegselmann et al., 2026)	**0.902** ±0.002	0.146 ±0.016	0.103 ±0.001	0.611 ±0.006	0.738 ±0.023	0.824 ±0.008	0.449 ±0.009	0.117 ±0.001	0.494 ±0.005	0.814 ±0.014
Record2Vec (Ji et al., 2026)	0.875 ±0.008	0.393 ±0.049	0.108 ±0.001	0.620 ±0.005	0.784 ±0.003	0.837 ±0.003	0.457 ±0.010	0.119 ±0.001	0.507 ±0.002	0.775 ±0.002
Template (Demirel et al., 2026)	0.884 ±0.004	0.113 ±0.011	0.104 ±0.001	**0.630** ±0.003	0.669 ±0.028	0.821 ±0.005	0.439 ±0.005	0.120 ±0.002	0.505 ±0.006	0.793 ±0.007
QA	0.838 ±0.002	0.272 ±0.052	0.108 ±0.001	0.590 ±0.008	0.775 ±0.008	0.819 ±0.006	0.500 ±0.029	0.117 ±0.001	0.498 ±0.004	0.765 ±0.009
Treatment	0.835 ±0.009	0.243 ±0.020	0.106 ±0.001	0.602 ±0.006	0.788 ±0.004	0.825 ±0.002	0.458 ±0.015	0.117 ±0.001	0.502 ±0.006	0.826 ±0.005
Organ-severity trajectory	0.867 ±0.006	0.727 ±0.234	0.111 ±0.001	0.592 ±0.006	0.829 ±0.005	**0.844** ±0.004	0.523 ±0.032	0.121 ±0.001	0.501 ±0.004	0.806 ±0.006
PORTER	0.883 ±0.006	0.357 ±0.050	0.107 ±0.001	0.617 ±0.003	0.784 ±0.007	0.836 ±0.005	**0.434** ±0.014	0.120 ±0.001	0.500 ±0.006	0.805 ±0.009

All six directions			
Mean rank↓	Cumulative occurrences↑		
	Top-1	Top-2	Top-3
15 (12.6)	0	1	2
16 (13.5)	1	1	1
7 (7.7)	0	3	6
12 (10.4)	1	4	5
14 (12.1)	4	4	4
10 (9.0)	1	1	4
13 (11.0)	1	2	3
11 (9.1)	0	0	2
2 (5.6)	4	8	10
6 (6.3)	4	8	10
3 (5.8)	0	3	7
3 (5.8)	**6**	**11**	12
8 (7.9)	0	0	0
5 (6.1)	0	1	3
9 (8.1)	2	6	8
1 (5.1)	**6**	7	**13**

Table 23: Zero-shot transfer, all six directions, the Muse arm at the xLSTM-Mixer head. The directions are carried over three blocks of two. Cells are the mean over the five cross-validation folds with one standard deviation. The columns to the right of the first block summarize all six directions. *Mean rank* is each method's position among the sixteen by its average rank over the thirty columns, with that average in parentheses, and *Cumulative occurrences* counts the columns, out of thirty, in which a method places in the top-1, 2, 3. Groups, metrics and marking as in Table 1.

Method	HIRID → MIMIC					HIRID → PPICU				
	mort AUROC↑	los MSE↓	forecast MSE↓	lab AUROC↑	drug AUROC↑	mort AUROC↑	los MSE↓	forecast MSE↓	lab AUROC↑	drug AUROC↑
ATTACK	0.506 ±0.007	3.551 ±0.028	0.227 ±0.088	0.376 ±0.004	0.660 ±0.002	0.500 ±0.001	10.112 ±0.135	0.221 ±0.091	0.482 ±0.005	0.634 ±0.000
RANDOM	0.500 ±0.010	3.541 ±0.077	0.410 ±0.176	0.384 ±0.014	0.659 ±0.004	0.500 ±0.012	9.824 ±0.232	0.451 ±0.193	0.487 ±0.012	0.632 ±0.003
CLMBR-T (Wornow et al., 2023)	0.754 ±0.008	3.044 ±0.108	1.054 ±0.954	0.486 ±0.019	0.765 ±0.009	0.661 ±0.006	7.398 ±0.203	0.972 ±0.895	0.585 ±0.008	0.656 ±0.007
GARLIC (Wang et al., 2026)	0.584 ±0.013	3.098 ±0.085	0.321 ±0.227	0.419 ±0.024	0.681 ±0.007	0.641 ±0.008	8.192 ±0.178	0.314 ±0.214	0.543 ±0.030	0.696 ±0.014
KEDGN (Luo et al., 2024)	0.300 ±0.035	4.081 ±0.627	0.447 ±0.257	0.407 ±0.045	0.375 ±0.077	0.660 ±0.043	14.418 ±2.387	0.459 ±0.290	0.513 ±0.027	0.416 ±0.013
MIRA (Li et al., 2026)	0.646 ±0.023	3.035 ±0.045	0.773 ±0.642	0.404 ±0.028	0.778 ±0.059	0.649 ±0.019	8.096 ±0.248	0.849 ±0.670	0.509 ±0.024	0.691 ±0.022
TabICL (Qu et al., 2025)	0.759 ±0.015	3.725 ±0.148	0.157 ±0.018	0.430 ±0.015	0.690 ±0.009	0.746 ±0.022	10.281 ±0.430	0.160 ±0.015	0.537 ±0.025	0.711 ±0.009
MEWS (Subbe et al., 2001)	0.783 ±0.003	2.874 ±0.014	0.512 ±0.338	0.424 ±0.013	0.722 ±0.004	0.682 ±0.005	7.313 ±0.123	0.506 ±0.256	0.491 ±0.005	0.711 ±0.001
SOFA (Vincent et al., 1996)	0.801 ±0.008	2.873 ±0.042	0.346 ±0.349	0.465 ±0.016	0.849 ±0.006	0.729 ±0.005	7.180 ±0.065	0.382 ±0.422	0.537 ±0.007	**0.828** ±0.002
Markdown (Hegselmann et al., 2026)	0.767 ±0.010	**2.777** ±0.118	0.269 ±0.113	0.468 ±0.013	0.838 ±0.025	0.682 ±0.004	**6.709** ±0.451	0.289 ±0.155	0.578 ±0.009	0.577 ±0.022
Record2Vec (Ji et al., 2026)	0.797 ±0.010	2.948 ±0.020	0.209 ±0.039	0.459 ±0.006	0.840 ±0.003	**0.762** ±0.004	8.236 ±0.341	0.195 ±0.055	0.577 ±0.011	0.751 ±0.004
Template (Demirel et al., 2026)	0.751 ±0.011	2.835 ±0.049	0.233 ±0.104	**0.494** ±0.013	0.897 ±0.007	0.736 ±0.002	7.176 ±0.276	0.266 ±0.149	**0.596** ±0.004	0.650 ±0.012
QA	0.789 ±0.006	3.185 ±0.087	0.179 ±0.070	0.441 ±0.002	0.832 ±0.003	0.748 ±0.004	8.179 ±0.206	0.198 ±0.088	0.536 ±0.007	0.722 ±0.004
Treatment	0.785 ±0.012	3.086 ±0.079	0.224 ±0.095	0.457 ±0.007	0.871 ±0.005	0.725 ±0.013	7.552 ±0.178	0.232 ±0.095	0.558 ±0.011	0.776 ±0.005
Organ-severity trajectory	0.803 ±0.026	3.237 ±0.125	4.806 ±2.435	0.452 ±0.016	0.864 ±0.006	0.726 ±0.019	10.267 ±0.383	3.867 ±2.089	0.583 ±0.011	0.773 ±0.003
PORTER	**0.806** ±0.013	2.860 ±0.029	**0.137** ±0.004	0.457 ±0.006	**0.911** ±0.001	0.760 ±0.009	8.147 ±0.263	**0.159** ±0.032	0.582 ±0.003	0.741 ±0.007

Method	MIMIC → HIRID					MIMIC → PPICU				
	mort AUROC↑	los MSE↓	forecast MSE↓	lab AUROC↑	drug AUROC↑	mort AUROC↑	los MSE↓	forecast MSE↓	lab AUROC↑	drug AUROC↑
ATTACK	0.502 ±0.008	0.746 ±0.003	**0.124** ±0.009	0.388 ±0.011	0.698 ±0.001	0.500 ±0.001	9.874 ±0.092	**0.159** ±0.011	0.489 ±0.016	0.634 ±0.000
RANDOM	0.503 ±0.031	0.791 ±0.019	0.470 ±0.287	0.411 ±0.003	0.692 ±0.006	0.503 ±0.009	9.646 ±0.135	0.443 ±0.259	0.473 ±0.004	0.631 ±0.004
CLMBR-T (Wornow et al., 2023)	0.777 ±0.017	0.564 ±0.055	0.484 ±0.204	0.524 ±0.039	0.799 ±0.004	0.695 ±0.010	7.853 ±0.404	0.426 ±0.183	0.571 ±0.020	0.656 ±0.017
GARLIC (Wang et al., 2026)	0.586 ±0.019	0.610 ±0.050	0.490 ±0.720	0.466 ±0.004	0.727 ±0.017	0.598 ±0.011	9.302 ±0.139	0.464 ±0.603	0.482 ±0.021	0.683 ±0.006
KEDGN (Luo et al., 2024)	0.613 ±0.018	1.082 ±0.195	0.647 ±0.397	0.455 ±0.020	0.572 ±0.064	0.511 ±0.039	7.919 ±0.215	0.619 ±0.391	0.512 ±0.014	0.661 ±0.024
MIRA (Li et al., 2026)	0.749 ±0.033	0.604 ±0.108	0.341 ±0.216	0.418 ±0.038	0.765 ±0.043	0.619 ±0.018	9.049 ±0.455	0.332 ±0.162	0.500 ±0.028	0.674 ±0.005
TabICL (Qu et al., 2025)	0.834 ±0.027	0.630 ±0.038	0.154 ±0.047	0.506 ±0.016	0.763 ±0.011	0.732 ±0.020	9.464 ±0.258	0.204 ±0.061	0.530 ±0.005	0.726 ±0.006
MEWS (Subbe et al., 2001)	0.831 ±0.009	0.552 ±0.022	0.427 ±0.311	0.462 ±0.007	0.748 ±0.006	0.713 ±0.006	7.885 ±0.113	0.473 ±0.346	0.516 ±0.011	0.704 ±0.002
SOFA (Vincent et al., 1996)	0.852 ±0.009	0.539 ±0.015	1.202 ±0.478	0.486 ±0.020	0.875 ±0.005	0.770 ±0.007	7.860 ±0.083	1.114 ±0.381	0.542 ±0.016	**0.844** ±0.004
Markdown (Hegselmann et al., 2026)	0.871 ±0.015	**0.399** ±0.015	0.135 ±0.019	0.543 ±0.010	0.838 ±0.008	0.671 ±0.012	6.557 ±0.366	0.167 ±0.017	0.558 ±0.006	0.631 ±0.018
Record2Vec (Ji et al., 2026)	0.873 ±0.007	0.480 ±0.027	0.138 ±0.009	0.536 ±0.012	0.853 ±0.004	0.766 ±0.005	8.752 ±0.315	0.177 ±0.014	0.562 ±0.004	0.746 ±0.001
Template (Demirel et al., 2026)	0.843 ±0.012	0.481 ±0.049	0.135 ±0.009	**0.558** ±0.015	0.849 ±0.008	0.713 ±0.012	**6.274** ±0.228	0.170 ±0.018	**0.573** ±0.009	0.713 ±0.014
QA	0.845 ±0.008	0.607 ±0.024	0.132 ±0.005	0.482 ±0.006	0.788 ±0.004	0.736 ±0.006	8.666 ±0.227	0.179 ±0.006	0.522 ±0.003	0.723 ±0.003
Treatment	0.855 ±0.004	0.490 ±0.013	0.130 ±0.006	0.501 ±0.005	0.853 ±0.006	0.763 ±0.004	8.380 ±0.259	0.175 ±0.008	0.538 ±0.005	0.782 ±0.004
Organ-severity trajectory	0.854 ±0.002	0.813 ±0.138	0.174 ±0.045	0.485 ±0.034	0.862 ±0.003	0.772 ±0.003	8.991 ±0.212	0.227 ±0.060	0.530 ±0.032	0.779 ±0.003
PORTER	**0.881** ±0.006	0.472 ±0.057	0.130 ±0.005	0.531 ±0.017	**0.908** ±0.002	**0.794** ±0.007	8.091 ±0.444	0.170 ±0.009	0.561 ±0.003	0.778 ±0.002

Method	PPICU → HIRID					PPICU → MIMIC				
	mort AUROC↑	los MSE↓	forecast MSE↓	lab AUROC↑	drug AUROC↑	mort AUROC↑	los MSE↓	forecast MSE↓	lab AUROC↑	drug AUROC↑
ATTACK	0.501 ±0.008	0.322 ±0.048	0.127 ±0.009	0.512 ±0.010	0.699 ±0.003	0.501 ±0.009	0.641 ±0.036	0.131 ±0.010	0.422 ±0.008	0.661 ±0.003
RANDOM	0.517 ±0.006	0.416 ±0.053	0.428 ±0.145	0.508 ±0.004	0.699 ±0.003	0.508 ±0.010	0.718 ±0.042	0.454 ±0.180	0.434 ±0.006	0.661 ±0.003
CLMBR-T (Wornow et al., 2023)	0.781 ±0.026	0.116 ±0.003	0.285 ±0.280	0.591 ±0.061	0.750 ±0.011	0.730 ±0.031	0.850 ±0.089	0.231 ±0.186	0.491 ±0.034	0.690 ±0.005
GARLIC (Wang et al., 2026)	0.316 ±0.012	0.270 ±0.028	0.698 ±1.056	0.531 ±0.046	0.654 ±0.024	0.520 ±0.014	0.493 ±0.024	0.512 ±0.686	0.467 ±0.011	0.719 ±0.008
KEDGN (Luo et al., 2024)	0.379 ±0.067	0.388 ±0.107	0.646 ±0.622	0.520 ±0.010	0.463 ±0.077	0.666 ±0.106	0.948 ±0.136	0.592 ±0.614	0.471 ±0.027	0.589 ±0.078
MIRA (Li et al., 2026)	0.794 ±0.025	0.110 ±0.008	0.147 ±0.035	0.521 ±0.035	0.792 ±0.008	0.714 ±0.035	0.484 ±0.010	0.179 ±0.035	0.453 ±0.013	0.807 ±0.003
TabICL (Qu et al., 2025)	0.650 ±0.031	0.873 ±0.215	0.161 ±0.038	0.533 ±0.013	0.697 ±0.058	0.648 ±0.028	0.797 ±0.097	0.154 ±0.018	0.512 ±0.015	0.659 ±0.037
MEWS (Subbe et al., 2001)	0.813 ±0.010	0.146 ±0.013	0.317 ±0.162	0.548 ±0.014	0.752 ±0.008	0.804 ±0.003	0.497 ±0.011	0.287 ±0.156	0.487 ±0.003	0.749 ±0.004
SOFA (Vincent et al., 1996)	0.817 ±0.010	0.178 ±0.013	0.259 ±0.174	0.613 ±0.010	**0.857** ±0.003	0.824 ±0.004	0.538 ±0.036	0.273 ±0.149	**0.526** ±0.005	0.823 ±0.004
Markdown (Hegselmann et al., 2026)	**0.899** ±0.003	0.148 ±0.036	**0.106** ±0.006	0.611 ±0.002	0.751 ±0.023	0.819 ±0.009	0.462 ±0.019	0.126 ±0.008	0.496 ±0.003	0.812 ±0.008
Record2Vec (Ji et al., 2026)	0.880 ±0.006	0.350 ±0.073	0.125 ±0.027	0.622 ±0.004	0.780 ±0.009	0.837 ±0.001	0.452 ±0.031	0.126 ±0.011	0.512 ±0.007	0.771 ±0.005
Template (Demirel et al., 2026)	0.890 ±0.004	**0.101** ±0.017	0.122 ±0.036	**0.629** ±0.010	0.657 ±0.021	0.813 ±0.006	0.441 ±0.011	0.128 ±0.009	0.510 ±0.008	0.807 ±0.005
QA	0.839 ±0.004	0.239 ±0.029	0.114 ±0.009	0.582 ±0.010	0.780 ±0.004	0.816 ±0.002	0.485 ±0.020	**0.122** ±0.003	0.490 ±0.005	0.760 ±0.008
Treatment	0.834 ±0.009	0.204 ±0.042	0.114 ±0.009	0.603 ±0.004	0.785 ±0.010	0.824 ±0.001	0.446 ±0.020	0.125 ±0.007	0.507 ±0.006	**0.825** ±0.006
Organ-severity trajectory	0.870 ±0.003	0.490 ±0.112	0.263 ±0.250	0.595 ±0.007	0.829 ±0.009	**0.846** ±0.003	0.474 ±0.022	0.231 ±0.202	0.505 ±0.005	0.808 ±0.008
PORTER	0.888 ±0.005	0.391 ±0.091	0.113 ±0.003	0.614 ±0.006	0.780 ±0.009	0.841 ±0.002	**0.427** ±0.019	0.125 ±0.001	0.502 ±0.010	0.802 ±0.005

All six directions			
Mean rank↓	Cumulative occurrences ↑		
	Top-1	Top-2	Top-3
14 (12.3)	2	2	2
16 (14.0)	0	0	0
9 (9.1)	0	3	5
13 (12.1)	0	0	0
15 (13.9)	0	0	0
12 (10.3)	0	1	2
10 (9.6)	0	3	3
11 (9.7)	0	0	0
6 (6.1)	4	6	9
5 (5.6)	5	8	11
3 (5.0)	1	3	9
2 (4.8)	7	12	13
7 (7.2)	1	1	4
4 (5.1)	1	3	7
8 (7.6)	1	4	8
1 (3.6)	**8**	**14**	**17**

Table 24: Few-shot transfer on sixteen labeled records of the target site. The two transfer directions into PPICU are given one above the other. Cells are the mean over five independently sampled support sets of 16 records, with one standard deviation across the sets, with an MLP head. Cells are the mean over the five source-trained folds with one standard deviation, with an MLP head. Best in each column in **bold**, second underlined; *Rank* and '1st/2nd' are over all ten columns of the table, so they read the same in both blocks. *PORTER* is the schema-guided representation reported in Table 1, and *PORTER + policy* augments it with distilled target-site context.

Method	HIRID → PPICU					Rank ↓	1st/2nd ↑
	mort AUROC↑	los MSE↓	forecast MSE↓	lab AUROC↑	drug AUROC↑		
CLMBR-T (Wornow et al., 2023)	0.671 ±0.026	7.60 ±0.37	0.104 ±0.001	0.645 ±0.020	0.686 ±0.003	7 (7.1)	0/1
GARLIC (Wang et al., 2026)	0.634 ±0.004	7.87 ±0.19	0.094 ±0.002	0.599 ±0.015	0.694 ±0.003	9 (7.7)	0/0
KEDGN (Luo et al., 2024)	0.638 ±0.038	10.70 ±0.34	0.125 ±0.002	0.567 ±0.042	0.541 ±0.050	12 (10.8)	0/0
MIRA (Li et al., 2026)	0.681 ±0.002	8.43 ±0.19	0.096 ±0.002	0.604 ±0.015	0.730 ±0.003	8 (7.6)	0/0
TabICL (Qu et al., 2025)	0.717 ±0.011	12.04 ±0.83	0.127 ±0.004	0.581 ±0.024	0.706 ±0.010	10 (8.9)	0/0
MEWS (Subbe et al., 2001)	0.709 ±0.007	8.77 ±0.11	0.126 ±0.001	0.529 ±0.008	0.731 ±0.001	11 (9.0)	0/0
SOFA (Vincent et al., 1996)	0.754 ±0.007	8.53 ±0.15	0.127 ±0.001	0.565 ±0.005	**0.845** ±0.004	5 (6.1)	2/0
Markdown (Hegselmann et al., 2026)	0.763 ±0.005	7.62 ±0.19	0.088 ±0.002	0.634 ±0.006	0.722 ±0.007	3 (3.9)	0/3
Record2Vec (Ji et al., 2026)	0.722 ±0.006	8.27 ±0.18	0.090 ±0.001	0.569 ±0.012	0.741 ±0.004	5 (6.1)	0/0
Record2Vec + policy	0.733 ±0.002	8.06 ±0.08	0.086 ±0.002	0.594 ±0.015	0.722 ±0.010	4 (5.3)	0/1
PORTER	0.753 ±0.004	9.06 ±0.28	0.080 ±0.005	0.651 ±0.045	0.802 ±0.007	2 (3.5)	0/6
PORTER + policy	**0.777** ±0.004	**7.24** ±0.16	**0.044** ±0.006	**0.758** ±0.019	0.795 ±0.007	**1 (1.4)**	8/0

Method	MIMIC → PPICU					Rank ↓	1st/2nd ↑
	mort AUROC↑	los MSE↓	forecast MSE↓	lab AUROC↑	drug AUROC↑		
CLMBR-T (Wornow et al., 2023)	0.700 ±0.008	6.82 ±0.20	0.120 ±0.001	0.596 ±0.008	0.662 ±0.007	7 (7.1)	0/1
GARLIC (Wang et al., 2026)	0.561 ±0.003	7.20 ±0.18	0.106 ±0.003	0.576 ±0.017	0.683 ±0.003	9 (7.7)	0/0
KEDGN (Luo et al., 2024)	0.503 ±0.025	8.46 ±0.35	0.160 ±0.001	0.557 ±0.010	0.574 ±0.026	12 (10.8)	0/0
MIRA (Li et al., 2026)	0.648 ±0.007	7.20 ±0.28	0.103 ±0.003	0.543 ±0.016	0.708 ±0.007	8 (7.6)	0/0
TabICL (Qu et al., 2025)	0.754 ±0.018	7.98 ±0.52	0.137 ±0.002	0.552 ±0.008	0.703 ±0.021	10 (8.9)	0/0
MEWS (Subbe et al., 2001)	0.744 ±0.004	7.22 ±0.14	0.156 ±0.001	0.534 ±0.003	0.725 ±0.003	11 (9.0)	0/0
SOFA (Vincent et al., 1996)	0.784 ±0.003	7.15 ±0.20	0.156 ±0.002	0.572 ±0.007	**0.856** ±0.002	5 (6.1)	2/0
Markdown (Hegselmann et al., 2026)	0.740 ±0.004	5.48 ±0.28	0.098 ±0.001	0.611 ±0.018	0.762 ±0.010	3 (3.9)	0/3
Record2Vec (Ji et al., 2026)	0.745 ±0.006	6.02 ±0.41	0.108 ±0.001	0.547 ±0.011	0.753 ±0.003	5 (6.1)	0/0
Record2Vec + policy	0.757 ±0.003	5.46 ±0.06	0.106 ±0.003	0.547 ±0.007	0.729 ±0.007	4 (5.3)	0/1
PORTER	0.787 ±0.016	6.74 ±0.26	0.098 ±0.003	0.591 ±0.028	0.798 ±0.007	2 (3.5)	0/6
PORTER + policy	**0.813** ±0.007	**4.88** ±0.06	**0.064** ±0.009	**0.624** ±0.005	0.790 ±0.005	**1 (1.4)**	8/0

Table 25: Few-shot transfer on sixteen labeled records of the target site, with a linear head throughout. Read exactly as Table 24: cells are the mean over the five sampled support sets with one standard deviation across the sets, best in each column in **bold** and second underlined, and *Rank* and '1st/2nd' are over all ten columns of this table.

Method	HIRID → PPICU					Rank ↓	1st/2nd ↑
	mort AUROC↑	los MSE↓	forecast MSE↓	lab AUROC↑	drug AUROC↑		
CLMBR-T (Wornow et al., 2023)	0.639 ±0.030	7.66 ±0.05	0.100 ±0.002	0.604 ±0.007	0.682 ±0.003	8 (7.4)	0/0
GARLIC (Wang et al., 2026)	0.627 ±0.012	8.07 ±0.13	0.091 ±0.002	0.584 ±0.017	0.693 ±0.003	9 (8.0)	0/0
KEDGN (Luo et al., 2024)	0.621 ±0.022	14.63 ±1.59	0.118 ±0.003	0.524 ±0.025	0.491 ±0.013	12 (11.1)	0/0
MIRA (Li et al., 2026)	0.687 ±0.002	8.35 ±0.10	0.085 ±0.003	0.572 ±0.012	0.732 ±0.002	7 (7.0)	0/0
TabICL (Qu et al., 2025)	0.715 ±0.022	11.12 ±0.36	0.123 ±0.002	0.559 ±0.007	0.708 ±0.004	11 (9.2)	0/0
MEWS (Subbe et al., 2001)	0.707 ±0.006	8.76 ±0.10	0.126 ±0.001	0.505 ±0.012	0.728 ±0.002	10 (8.9)	0/0
SOFA (Vincent et al., 1996)	0.756 ±0.006	8.45 ±0.12	0.128 ±0.001	0.538 ±0.005	**0.847** ±0.001	6 (6.2)	2/0
Markdown (Hegselmann et al., 2026)	0.762 ±0.003	**7.06** ±0.02	0.079 ±0.004	0.623 ±0.004	0.716 ±0.006	3 (3.6)	1/2
Record2Vec (Ji et al., 2026)	0.726 ±0.004	8.19 ±0.09	0.085 ±0.003	0.549 ±0.016	0.745 ±0.002	5 (6.1)	0/0
Record2Vec + policy	0.735 ±0.001	7.91 ±0.07	0.077 ±0.004	0.576 ±0.010	0.709 ±0.003	4 (5.3)	0/0
PORTER	0.750 ±0.009	8.76 ±0.43	0.061 ±0.008	0.652 ±0.014	0.806 ±0.004	2 (3.4)	0/7
PORTER + policy	**0.777** ±0.008	7.19 ±0.06	**0.039** ±0.006	**0.783** ±0.084	0.798 ±0.007	**1 (1.5)**	7/1

Method	MIMIC → PPICU					Rank ↓	1st/2nd ↑
	mort AUROC↑	los MSE↓	forecast MSE↓	lab AUROC↑	drug AUROC↑		
CLMBR-T (Wornow et al., 2023)	0.696 ±0.007	6.65 ±0.10	0.118 ±0.001	0.565 ±0.003	0.663 ±0.005	8 (7.4)	0/0
GARLIC (Wang et al., 2026)	0.564 ±0.006	7.44 ±0.14	0.103 ±0.001	0.547 ±0.005	0.675 ±0.008	9 (8.0)	0/0
KEDGN (Luo et al., 2024)	0.504 ±0.019	8.09 ±0.49	0.158 ±0.001	0.542 ±0.009	0.590 ±0.022	12 (11.1)	0/0
MIRA (Li et al., 2026)	0.657 ±0.004	6.43 ±0.11	0.096 ±0.004	0.533 ±0.015	0.703 ±0.005	7 (7.0)	0/0
TabICL (Qu et al., 2025)	0.740 ±0.020	7.91 ±0.24	0.135 ±0.002	0.527 ±0.007	0.721 ±0.008	11 (9.2)	0/0
MEWS (Subbe et al., 2001)	0.749 ±0.002	6.81 ±0.04	0.153 ±0.001	0.519 ±0.005	0.733 ±0.003	10 (8.9)	0/0
SOFA (Vincent et al., 1996)	0.785 ±0.004	6.87 ±0.08	0.152 ±0.001	0.552 ±0.004	**0.856** ±0.002	6 (6.2)	2/0
Markdown (Hegselmann et al., 2026)	0.743 ±0.007	5.44 ±0.21	0.096 ±0.002	0.597 ±0.012	0.759 ±0.004	3 (3.6)	1/2
Record2Vec (Ji et al., 2026)	0.746 ±0.003	6.11 ±0.26	0.105 ±0.003	0.536 ±0.006	0.752 ±0.002	5 (6.1)	0/0
Record2Vec + policy	0.757 ±0.005	5.75 ±0.15	0.103 ±0.002	0.540 ±0.006	0.730 ±0.003	4 (5.3)	0/0
PORTER	0.786 ±0.009	6.74 ±0.17	0.079 ±0.007	0.647 ±0.013	0.802 ±0.006	2 (3.4)	0/7
PORTER + policy	**0.808** ±0.003	**4.90** ±0.05	**0.059** ±0.011	**0.719** ±0.016	0.792 ±0.004	**1 (1.5)**	7/1

Table 26: Few-shot transfer on sixteen labeled records of the target site, with a LSTM head throughout. Read exactly as Table 24: cells are the mean over the five source-trained folds with one standard deviation, best in each column in **bold** and second underlined, and *Rank* and '1st/2nd' are over all ten columns of this table.

Method	HIRID → PPICU					Rank ↓	1st/2nd ↑
	mort AUROC↑	los MSE↓	forecast MSE↓	lab AUROC↑	drug AUROC↑		
CLMBR-T (Wornow et al., 2023)	0.676 ±0.009	7.57 ±0.19	0.105 ±0.002	0.594 ±0.008	0.684 ±0.002	7 (6.7)	0/1
GARLIC (Wang et al., 2026)	0.627 ±0.011	7.90 ±0.23	0.093 ±0.001	0.572 ±0.017	0.696 ±0.001	8 (7.8)	0/1
KEDGN (Luo et al., 2024)	0.641 ±0.029	13.21 ±0.15	0.099 ±0.001	0.531 ±0.024	0.500 ±0.100	12 (10.1)	0/1
MIRA (Li et al., 2026)	0.682 ±0.002	9.36 ±0.16	0.113 ±0.007	0.558 ±0.014	0.729 ±0.002	11 (9.4)	0/0
TabICL (Qu et al., 2025)	0.725 ±0.014	12.41 ±0.33	0.100 ±0.002	0.568 ±0.017	0.723 ±0.009	9 (7.9)	0/0
MEWS (Subbe et al., 2001)	0.706 ±0.003	8.96 ±0.34	0.110 ±0.001	0.518 ±0.007	0.730 ±0.003	10 (8.9)	0/0
SOFA (Vincent et al., 1996)	0.751 ±0.013	8.71 ±0.44	0.111 ±0.002	0.555 ±0.004	**0.848** ±0.005	6 (6.1)	2/1
Markdown (Hegselmann et al., 2026)	0.755 ±0.008	**7.25** ±0.32	0.092 ±0.001	0.578 ±0.029	0.735 ±0.007	2 (3.5)	2/1
Record2Vec (Ji et al., 2026)	0.725 ±0.003	8.70 ±0.49	0.098 ±0.002	0.554 ±0.010	0.744 ±0.003	5 (5.8)	0/0
Record2Vec + policy	0.730 ±0.006	8.20 ±0.14	0.094 ±0.001	0.571 ±0.011	0.723 ±0.007	4 (5.5)	0/1
PORTER	0.752 ±0.005	8.69 ±0.37	0.089 ±0.001	0.591 ±0.020	0.810 ±0.002	3 (3.6)	0/3
PORTER + policy	**0.775** ±0.005	7.35 ±0.13	**0.061** ±0.004	**0.611** ±0.056	0.798 ±0.006	**1 (1.8)**	6/1

Method	MIMIC → PPICU					Rank ↓	1st/2nd ↑
	mort AUROC↑	los MSE↓	forecast MSE↓	lab AUROC↑	drug AUROC↑		
CLMBR-T (Wornow et al., 2023)	0.702 ±0.004	7.04 ±0.34	0.114 ±0.001	0.557 ±0.006	0.663 ±0.003	7 (6.7)	0/1
GARLIC (Wang et al., 2026)	0.563 ±0.004	7.46 ±0.19	0.110 ±0.003	0.528 ±0.019	0.680 ±0.004	8 (7.8)	0/1
KEDGN (Luo et al., 2024)	0.515 ±0.014	8.37 ±0.38	0.137 ±0.002	0.562 ±0.012	0.616 ±0.029	12 (10.1)	0/1
MIRA (Li et al., 2026)	0.649 ±0.003	7.18 ±0.09	0.120 ±0.002	0.503 ±0.007	0.690 ±0.004	11 (9.4)	0/0
TabICL (Qu et al., 2025)	0.759 ±0.015	7.98 ±0.25	0.117 ±0.001	0.540 ±0.002	0.718 ±0.014	9 (7.9)	0/0
MEWS (Subbe et al., 2001)	0.746 ±0.003	7.09 ±0.14	0.147 ±0.001	0.523 ±0.003	0.731 ±0.003	10 (8.9)	0/0
SOFA (Vincent et al., 1996)	0.781 ±0.003	7.31 ±0.17	0.143 ±0.002	0.548 ±0.009	**0.859** ±0.002	6 (6.1)	2/1
Markdown (Hegselmann et al., 2026)	0.740 ±0.003	6.22 ±0.28	0.114 ±0.002	**0.599** ±0.011	0.763 ±0.002	2 (3.5)	2/1
Record2Vec (Ji et al., 2026)	0.749 ±0.005	6.29 ±0.33	0.114 ±0.002	0.544 ±0.007	0.751 ±0.003	5 (5.8)	0/0
Record2Vec + policy	0.750 ±0.002	6.05 ±0.15	0.114 ±0.003	0.530 ±0.009	0.744 ±0.004	4 (5.5)	0/1
PORTER	0.780 ±0.008	7.09 ±0.07	0.111 ±0.002	0.544 ±0.014	0.801 ±0.004	3 (3.6)	0/3
PORTER + policy	**0.811** ±0.005	**4.90** ±0.04	**0.081** ±0.005	0.552 ±0.005	0.792 ±0.005	**1 (1.8)**	6/1

Table 27: Few-shot transfer on an outlier support set of sixteen labeled records of the target site. The two transfer directions are given one above the other. Cells are the mean over the five source-trained folds with one standard deviation, with an MLP head. Best in each column in **bold** and second underlined. *Rank* and '1st/2nd' are over all ten columns of the table, so they read the same in both blocks. CLMBR-T produces no `forecast` output (–), so its *Rank* is over the remaining eight columns.

Method	HIRID → PPICU					Rank ↓	1st/2nd ↑
	`mort` AUROC↑	`los` MSE↓	`forecast` MSE↓	`lab` AUROC↑	`drug` AUROC↑		
CLMBR-T (Wornow et al., 2023)	0.671±0.034	9.96±3.92	–	0.601±0.021	0.669±0.006	8 (7.0)	0/0
GARLIC (Wang et al., 2026)	0.633±0.006	9.05±0.44	0.154±0.002	0.562±0.029	0.666±0.020	9 (7.7)	0/1
KEDGN (Luo et al., 2024)	0.664±0.017	10.36±0.34	0.147±0.003	0.538±0.030	0.465±0.036	11 (9.4)	0/0
MIRA (Li et al., 2026)	0.675±0.004	9.25±0.56	0.157±0.005	0.472±0.026	0.669±0.011	12 (9.6)	0/0
TabICL (Qu et al., 2025)	0.720±0.014	11.73±0.80	0.171±0.005	0.542±0.018	0.700±0.015	10 (7.9)	0/0
MEWS (Subbe et al., 2001)	0.713±0.007	7.89±0.17	0.150±0.001	0.508±0.008	0.724±0.003	7 (6.6)	0/0
SOFA (Vincent et al., 1996)	0.753±0.007	7.61±0.09	0.151±0.001	0.543±0.003	**0.853**±0.004	2 (3.9)	2/1
Markdown (Hegselmann et al., 2026)	0.742±0.017	15.14±1.49	0.142±0.002	0.626±0.006	0.644±0.021	4 (5.7)	0/5
Record2Vec (Ji et al., 2026)	0.724±0.003	9.86±1.54	0.152±0.003	0.565±0.021	0.694±0.013	5 (6.1)	0/0
Record2Vec + policy	0.724±0.007	9.04±0.60	0.154±0.001	0.558±0.011	0.680±0.027	6 (6.4)	0/0
PORTER	0.768±0.007	8.88±0.47	0.155±0.006	0.560±0.023	0.801±0.005	3 (4.7)	0/3
PORTER + policy	**0.778**±0.004	**7.27**±0.12	**0.038**±0.007	**0.832**±0.007	0.795±0.005	1 (1.3)	8/1

Method	MIMIC → PPICU					Rank ↓	1st/2nd ↑
	`mort` AUROC↑	`los` MSE↓	`forecast` MSE↓	`lab` AUROC↑	`drug` AUROC↑		
CLMBR-T (Wornow et al., 2023)	0.671±0.008	6.50±0.81	–	0.566±0.010	0.647±0.009	8 (7.0)	0/0
GARLIC (Wang et al., 2026)	0.557±0.004	6.75±0.28	0.175±0.002	0.529±0.011	0.682±0.006	9 (7.7)	0/1
KEDGN (Luo et al., 2024)	0.543±0.023	7.92±0.64	0.184±0.001	0.546±0.007	0.552±0.031	11 (9.4)	0/0
MIRA (Li et al., 2026)	0.633±0.007	7.72±0.85	0.201±0.011	0.491±0.030	0.684±0.011	12 (9.6)	0/0
TabICL (Qu et al., 2025)	0.750±0.017	7.80±0.69	0.179±0.002	0.532±0.009	0.661±0.026	10 (7.9)	0/0
MEWS (Subbe et al., 2001)	0.731±0.007	6.59±0.18	0.184±0.001	0.517±0.004	0.736±0.002	7 (6.6)	0/0
SOFA (Vincent et al., 1996)	0.768±0.002	6.54±0.17	0.183±0.002	0.553±0.003	**0.857**±0.001	2 (3.9)	2/1
Markdown (Hegselmann et al., 2026)	0.654±0.012	5.54±0.48	0.175±0.005	0.573±0.004	0.564±0.017	4 (5.7)	0/5
Record2Vec (Ji et al., 2026)	0.735±0.004	5.64±0.23	0.184±0.004	0.520±0.009	0.728±0.005	5 (6.1)	0/0
Record2Vec + policy	0.737±0.007	5.69±0.48	0.186±0.002	0.521±0.012	0.730±0.002	6 (6.4)	0/0
PORTER	0.769±0.018	6.78±0.30	0.178±0.004	0.534±0.012	0.771±0.027	3 (4.7)	0/3
PORTER + policy	**0.814**±0.002	**4.87**±0.08	**0.072**±0.008	**0.617**±0.004	0.789±0.003	1 (1.3)	8/1

Table 28: Prompt ablations under all four summarizers. Inside every run of an arm the seven prompts are ranked against one another, as in Figure 5; ties take the better rank. *Mean* is that rank averaged over the 120 runs of the arm, and *1st* and *7th* count the runs the prompt took the best and the worst rank in. Lowest mean in each arm in **bold**.

Prompt	MedGemma			Qwen3.6			Gemma-4			Muse		
	Mean ↓	1st	7th	Mean ↓	1st	7th	Mean ↓	1st	7th	Mean ↓	1st	7th
static	3.08	31	6	**2.87**	51	10	**3.08**	37	9	**3.33**	38	17
no-course	**2.79**	43	5	3.54	32	17	3.39	37	19	3.76	25	18
PORTER (full)	3.97	12	7	3.84	8	10	4.09	11	6	4.33	8	14
reword	4.15	12	11	4.57	8	14	3.92	8	16	3.76	14	4
bare	4.12	14	12	3.98	14	2	4.28	14	13	4.35	14	19
generic	4.40	15	19	4.34	7	16	4.34	8	15	3.98	15	11
free-form	5.48	5	58	4.86	12	49	4.89	17	40	4.50	18	35

Table 29: Zero-shot transfer between EHRSHOT and INSPECT with the MLP head, one column per context window. Each cell is a representation's mean rank among the six, taken inside a run and averaged over the ten runs of that window. The number in parentheses is how many of these runs it wins outright. A run is one transfer direction and one task. The last column averages over all forty runs. RANDOM and Demographics do not use the context window. Their metric values are the same at every window. Best in each column in **bold**.

	Last			Full	All
Method	30 days	180 days	365 days	record	windows
RANDOM	5.20 (0)	5.40 (0)	5.30 (1)	5.50 (0)	5.35 (1)
Demographics	3.40 (0)	3.70 (0)	3.90 (0)	4.50 (0)	3.88 (0)
Count	4.50 (0)	4.60 (0)	4.20 (0)	4.30 (0)	4.40 (0)
CLMBR-T (Wornow et al., 2023)	4.10 (0)	4.30 (0)	4.20 (0)	3.40 (0)	4.00 (0)
Direct embedding	2.10 (2)	1.80 (2)	1.80 (3)	1.90 (3)	1.90 (10)
PORTER	**1.70 (8)**	**1.20 (8)**	**1.60 (6)**	**1.40 (7)**	**1.48 (29)**

Table 30: Zero-shot transfer between EHRSHOT and INSPECT with the MLP head, the same mean rank restricted to one task at a time. Each cell averages the two runs of a task at one context window, one run per transfer direction. Best in each column in **bold**.

	Last 30 days					Last 180 days					Last 365 days					Full record				
Method	Hyperlip.	Hypert.	Malig.	Creat.	Admit.	Hyperlip.	Hypert.	Malig.	Creat.	Admit.	Hyperlip.	Hypert.	Malig.	Creat.	Admit.	Hyperlip.	Hypert.	Malig.	Creat.	Admit.
RANDOM	6.0	6.0	6.0	4.0	4.0	6.0	6.0	6.0	4.5	4.5	6.0	6.0	6.0	5.0	3.5	6.0	6.0	6.0	5.5	4.0
Demographics	3.5	3.0	3.0	4.0	3.5	3.5	3.0	3.5	5.0	3.5	4.0	3.0	4.0	5.0	3.5	5.0	4.0	4.5	5.5	3.5
Count	3.5	4.5	5.0	4.0	5.5	5.0	4.5	5.0	4.0	4.5	5.0	5.0	4.5	3.0	3.5	4.0	5.0	4.5	4.0	4.0
CLMBR-T (Wornow et al., 2023)	2.5	4.5	4.0	5.0	4.5	3.5	4.5	3.5	4.5	5.5	3.0	4.0	3.5	5.0	5.5	2.5	3.0	3.0	3.0	5.5
Direct embedding	**1.0**	2.0	2.0	3.0	2.5	**1.0**	2.0	2.0	2.0	2.0	**1.0**	2.0	2.0	2.0	**2.0**	2.0	**1.5**	2.0	2.0	**2.0**
PORTER	4.5	**1.0**	**1.0**	**1.0**	**1.0**	2.0	**1.0**	**1.0**	**1.0**	**1.0**	2.0	**1.0**	**1.0**	**1.0**	3.0	**1.5**	**1.5**	**1.0**	**1.0**	**2.0**

Table 31: Zero-shot transfer between EHRSHOT and INSPECT with the MLP head, for direct embedding and PORTER under two renderings of the record. *All values* lists every value recorded for each code, and *latest value* keeps only the latest value of each code. Each cell is the mean over the four context windows. The ± is one standard deviation over 20 values, the five source-trained folds at each of the four context windows. The better rendering for each method and column in **bold**.

	EHRSHOT → INSPECT					INSPECT → EHRSHOT				
Method	Hyperlip. AUROC↑	Hypert. AUROC↑	Malig. AUROC↑	Creat. MSE↓	Admit. MSE↓	Hyperlip. AUROC↑	Hypert. AUROC↑	Malig. AUROC↑	Creat. MSE↓	Admit. MSE↓
Direct embedding (all values)	**0.641** ±0.022	**0.647** ±0.022	0.606 ±0.011	**0.396** ±0.095	1.089 ±0.045	**0.667** ±0.012	**0.708** ±0.010	**0.655** ±0.019	**1.813** ±0.148	0.942 ±0.016
Direct embedding (latest value)	0.628 ±0.021	0.635 ±0.017	**0.609** ±0.010	0.421 ±0.033	**1.059** ±0.029	0.646 ±0.008	0.687 ±0.014	0.647 ±0.026	1.957 ±0.081	**0.926** ±0.015
PORTER (all values)	**0.629** ±0.029	**0.660** ±0.016	**0.619** ±0.013	**0.309** ±0.035	1.052 ±0.026	**0.654** ±0.018	**0.718** ±0.015	**0.668** ±0.026	**1.458** ±0.182	0.962 ±0.043
PORTER (latest value)	0.596 ±0.028	0.622 ±0.018	0.576 ±0.019	0.335 ±0.042	**1.037** ±0.022	0.605 ±0.019	0.663 ±0.011	0.622 ±0.032	1.593 ±0.224	**0.917** ±0.006

Table 32: Zero-shot transfer between EHRSHOT and INSPECT with the MLP prediction head, one block per context window. Within each block, *Mean rank* is each method's position among the six by its average rank over the ten columns, with that average in parentheses, and the cumulative occurrences count the columns in which a method places in the top-1, 2, 3. Bold and underlined values are the best and second-best results in each block. Values are the mean over the five source-trained folds ± one standard deviation.

Method	EHRSHOT → INSPECT					INSPECT → EHRSHOT					Mean rank↓	Cumulative occurrences ↑		
	Hyperlip. AUROC↑	Hypert. AUROC↑	Malig. AUROC↑	Creat. MSE↓	Admit. MSE↓	Hyperlip. AUROC↑	Hypert. AUROC↑	Malig. AUROC↑	Creat. MSE↓	Admit. MSE↓		Top-1	Top-2	Top-3
Last 30 days														
RANDOM	0.507 ±0.007	0.510 ±0.005	0.493 ±0.004	0.455 ±0.002	1.075 ±0.023	0.516 ±0.012	0.512 ±0.009	0.480 ±0.025	2.494 ±0.054	1.169 ±0.033	6 (5.2)	0	2	2
Demographics	0.586 ±0.009	0.603 ±0.019	0.585 ±0.007	0.462 ±0.011	1.115 ±0.044	0.636 ±0.010	0.677 ±0.009	0.615 ±0.015	2.268 ±0.013	0.976 ±0.009	3 (3.4)	0	0	7
Count	0.564 ±0.014	0.585 ±0.013	0.531 ±0.016	0.507 ±0.037	1.571 ±0.168	0.646 ±0.020	0.617 ±0.017	0.587 ±0.017	2.048 ±0.097	1.124 ±0.027	5 (4.5)	0	1	2
CLMBR-T (Wornow et al., 2023)	0.593 ±0.009	0.568 ±0.008	0.558 ±0.004	0.617 ±0.050	1.283 ±0.032	0.641 ±0.020	0.637 ±0.015	0.605 ±0.020	2.267 ±0.043	0.987 ±0.024	4 (4.1)	0	1	2
Direct embedding	**0.611** ±0.013	0.627 ±0.009	0.606 ±0.016	0.486 ±0.155	1.101 ±0.040	**0.663** ±0.010	0.706 ±0.002	0.641 ±0.004	1.956 ±0.073	0.932 ±0.014	2 (2.1)	2	**8**	**9**
PORTER	0.585 ±0.011	**0.649** ±0.008	**0.607** ±0.011	**0.357** ±0.016	**1.048** ±0.020	0.632 ±0.015	**0.708** ±0.012	**0.658** ±0.010	**1.622** ±0.050	**0.914** ±0.008	**1 (1.7)**	**8**	**8**	8
Last 180 days														
RANDOM	0.507 ±0.007	0.510 ±0.005	0.493 ±0.004	0.455 ±0.002	1.075 ±0.023	0.516 ±0.012	0.512 ±0.009	0.480 ±0.025	2.494 ±0.054	1.169 ±0.033	6 (5.4)	0	0	2
Demographics	0.586 ±0.009	0.603 ±0.019	0.585 ±0.007	0.462 ±0.011	1.115 ±0.044	0.636 ±0.010	0.677 ±0.009	0.615 ±0.015	2.268 ±0.013	0.976 ±0.009	3 (3.7)	0	0	5
Count	0.574 ±0.012	0.589 ±0.010	0.554 ±0.013	0.457 ±0.022	1.163 ±0.044	0.624 ±0.018	0.613 ±0.023	0.589 ±0.034	2.248 ±0.067	1.054 ±0.056	5 (4.6)	0	0	0
CLMBR-T (Wornow et al., 2023)	0.581 ±0.013	0.578 ±0.008	0.577 ±0.007	0.521 ±0.015	1.270 ±0.022	0.641 ±0.029	0.623 ±0.020	0.627 ±0.018	2.181 ±0.081	1.148 ±0.030	4 (4.3)	0	0	3
Direct embedding	**0.643** ±0.003	0.633 ±0.008	0.606 ±0.010	0.358 ±0.019	1.050 ±0.031	**0.671** ±0.011	0.696 ±0.006	0.646 ±0.011	1.800 ±0.098	0.947 ±0.011	2 (1.8)	2	**10**	**10**
PORTER	0.624 ±0.008	**0.648** ±0.008	**0.616** ±0.004	**0.309** ±0.012	**1.035** ±0.006	0.652 ±0.007	**0.708** ±0.009	**0.662** ±0.029	**1.415** ±0.062	**0.929** ±0.008	**1 (1.2)**	**8**	**10**	**10**
Last 365 days														
RANDOM	0.507 ±0.007	0.510 ±0.005	0.493 ±0.004	0.455 ±0.002	**1.075** ±0.023	0.516 ±0.012	0.512 ±0.009	0.480 ±0.025	2.494 ±0.054	1.169 ±0.033	6 (5.3)	1	1	1
Demographics	0.586 ±0.009	0.603 ±0.019	0.585 ±0.007	0.462 ±0.011	1.115 ±0.044	0.636 ±0.010	0.677 ±0.009	0.615 ±0.015	2.268 ±0.013	0.976 ±0.009	3 (3.9)	0	0	4
Count	0.572 ±0.009	0.590 ±0.010	0.536 ±0.021	0.442 ±0.023	1.122 ±0.022	0.605 ±0.013	0.639 ±0.019	0.618 ±0.027	2.025 ±0.118	0.964 ±0.030	4 (4.2)	0	1	3
CLMBR-T (Wornow et al., 2023)	0.601 ±0.012	0.593 ±0.009	0.570 ±0.011	0.509 ±0.010	1.282 ±0.024	0.649 ±0.028	0.658 ±0.007	0.631 ±0.018	2.131 ±0.095	1.155 ±0.035	4 (4.2)	0	0	3
Direct embedding	**0.648** ±0.007	0.648 ±0.006	0.604 ±0.007	0.372 ±0.022	1.096 ±0.044	**0.676** ±0.009	0.718 ±0.006	0.654 ±0.013	1.773 ±0.103	**0.957** ±0.014	2 (1.8)	3	**9**	**10**
PORTER	0.646 ±0.009	**0.662** ±0.006	**0.620** ±0.010	**0.306** ±0.013	1.083 ±0.031	0.670 ±0.008	**0.719** ±0.007	**0.670** ±0.014	**1.417** ±0.101	1.003 ±0.011	**1 (1.6)**	**6**	**9**	9
Full record														
RANDOM	0.507 ±0.007	0.510 ±0.005	0.493 ±0.004	0.455 ±0.002	1.075 ±0.023	0.516 ±0.012	0.512 ±0.009	0.480 ±0.025	2.494 ±0.054	1.169 ±0.033	6 (5.5)	0	1	1
Demographics	0.586 ±0.009	0.603 ±0.019	0.585 ±0.007	0.462 ±0.011	1.115 ±0.044	0.636 ±0.010	0.677 ±0.009	0.615 ±0.015	2.268 ±0.013	0.976 ±0.009	5 (4.5)	0	1	1
Count	0.593 ±0.012	0.594 ±0.005	0.578 ±0.014	0.444 ±0.009	1.113 ±0.032	0.645 ±0.019	0.654 ±0.028	0.634 ±0.040	2.120 ±0.149	1.010 ±0.046	4 (4.3)	0	0	0
CLMBR-T (Wornow et al., 2023)	0.630 ±0.011	0.625 ±0.010	0.592 ±0.015	0.438 ±0.017	1.283 ±0.021	0.663 ±0.018	0.691 ±0.013	0.637 ±0.020	2.005 ±0.124	1.166 ±0.046	3 (3.4)	0	1	8
Direct embedding	**0.662** ±0.018	**0.681** ±0.006	0.607 ±0.011	0.371 ±0.019	1.108 ±0.038	0.657 ±0.006	0.714 ±0.002	0.678 ±0.019	1.723 ±0.178	**0.931** ±0.007	2 (1.9)	3	8	**10**
PORTER	0.659 ±0.005	0.681 ±0.012	**0.633** ±0.008	**0.265** ±0.008	**1.043** ±0.006	**0.664** ±0.007	**0.738** ±0.011	**0.683** ±0.034	**1.378** ±0.281	1.002 ±0.021	**1 (1.4)**	**7**	**9**	**10**