Separating Laboratory Values and Care Events in ICU Representation Learning

Anonymous authors
Paper under double-blind review

Abstract

Self-supervised models of health records typically merge laboratory results and care events (laboratory orders and medication starts) into a single sequence read by one transformer, mixing signals about the patient's physiology with signals about the care process. Because care processes differ across hospitals, we study whether these two types of information are better represented separately. We propose a coupled model in which laboratory values and care events are encoded by separate encoders and aligned with a contrastive loss. Across three ICU databases (eICU, MIMIC-IV, HiRID), ten seeds, and eight downstream tasks, the coupled model achieves significantly higher AUROC than the single-stream baseline in 67 of 85 comparisons and significantly lower AUROC in only one, with a mean improvement of +0.019, including at held-out sites and after cross-database transfer. The coupled model also preserves more information about measured laboratory values: linear probes recover laboratory values substantially better from its representation, with R^2 gains of +0.042 to +0.108 in distribution, +0.053 and +0.072 at held-out sites, and +0.075 to +0.238 after transfer. Ablations show that separating the encoders, using a set encoder for laboratory values, and adding a laboratory–vital contrastive loss improve laboratory-value recovery, whereas the cross-encoder contrastive loss primarily improves downstream prediction and can reduce recovery on one database. These results suggest that separately encoding laboratory values and care events, while aligning their representations contrastively, can improve both predictive performance and transfer of ICU representations across sites and databases.

1 Introduction

An electronic health record (EHR) is not a direct reading of the patient. Hripcsak & Albers (2013) describe raw EHR data as "an indirect reflection of the true patient state due to the recording process" and call for models of the healthcare process itself. That process carries its own signal: for 118 of 174 laboratory tests, the timing of the test predicted three-year survival better than its result (Agniel et al., 2018). Because what is recorded depends on why it was recorded, analyses that ignore the process are biased (Goldstein et al., 2016; Haneuse & Daniels, 2016), and site-specific care practices are one reason clinical models fail to transport between sites; Lasko et al. (2024) propose isolating the imprint of those practices from the patterns of disease. This suggests encoding what was measured separately from the record of how care was delivered.

Foundation models for EHRs instead serialize the record into one token sequence and train a single transformer over it (Pang et al., 2021; Steinberg et al., 2024; Renc et al., 2024; Kraljevic et al., 2024); most encode diagnoses, orders and timestamps as codes, and those that carry laboratory results quantize them into the same stream (Renc et al., 2024; Im et al., 2025). These tokens include laboratory values and care events. Laboratory values record what was measured and the numerical values of test results, whereas care events record clinical actions, such as ordering a test or starting a medication. Laboratory values and care events are related but only partially redundant: on eICU, a ridge regression from care-event features (which analytes were ordered, and when) recovers laboratory values at

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.

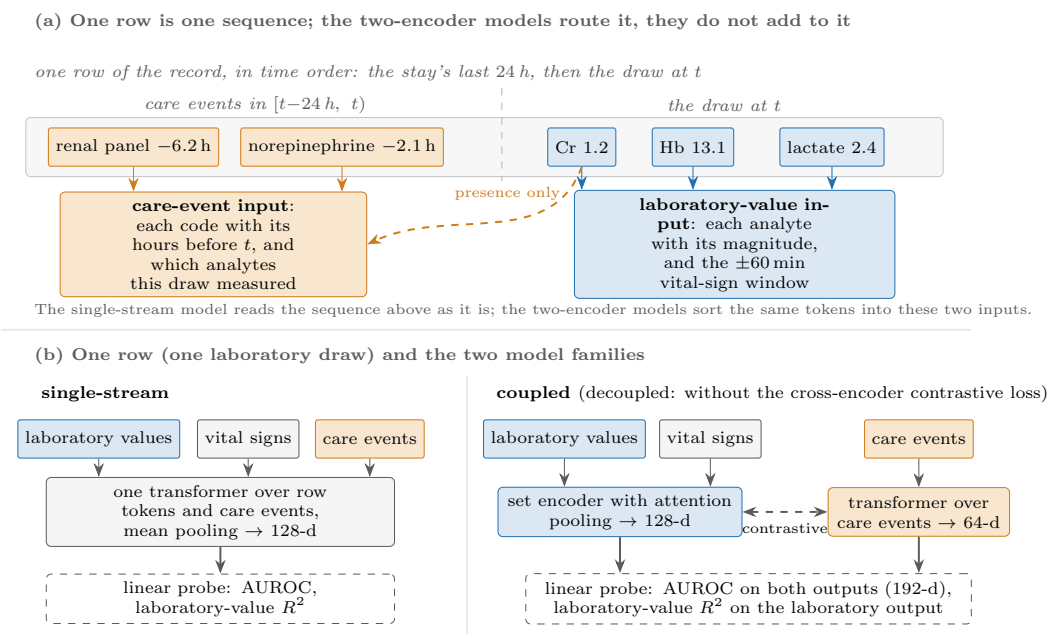

Figure 1: **One laboratory draw, and the two ways to encode it.** (a) One row of the record is a single sequence in time order, and all three models receive all of it: the two-encoder models sort these tokens rather than add to them. A measured analyte sends its magnitude to the laboratory-value tokens and only a presence bit to the care-event input, which carries in addition every laboratory order and medication start of the previous 24 h with its hours before t; the draw's own order is not among them. Medications never reach the laboratory-value tokens (§3). (b) The two model families. Both read every token of (a); they differ in whether one encoder or two read them, and the vital-sign window enters the laboratory-value side through a recurrent layer.

R^2=0.05, whereas a ridge from the other laboratory values in the same row reaches 0.59 (App. A).

Motivated by this distinction, we study whether laboratory values and care events benefit from separate representation learning. Specifically, we compare three self-supervised architectures for ICU records: a *single-stream model*, which encodes laboratory values and care events jointly with one transformer; a *decoupled model*, which assigns them separate encoders; and a *coupled model*, which uses the same separate encoders but aligns their representations with a contrastive loss. We train all three models on eICU, MIMIC-IV, and HiRID (Figure 1; §3) and evaluate them in distribution, at held-out sites, and under cross-database transfer. We assess both downstream prediction with linear AUROC probes and how much information about measured laboratory values is retained in the frozen representation, using linear recovery probes. To our knowledge, this is the first direct comparison of shared and separate encoders for laboratory values and care events in EHR representation learning.

Contributions.

1. **Downstream tasks.** On mortality, prolonged stay, sepsis and five preregistered labels, coupled minus single-stream AUROC is above zero in 67 of 85 database–task–setting cells and below zero in one (mean +0.019; decoupled model +0.006; §4).
2. **Laboratory-value recovery.** A linear probe recovers laboratory values better from both two-encoder models on all three databases, at held-out sites and after cross-database transfer, with every seed-paired interval excluding zero (§5).

3. **Decomposition.** Separating laboratory values from care events, the set encoder and a contrastive loss between laboratory values and vital signs each raise laboratory-value recovery; the contrastive loss between the two encoders adds downstream AUROC instead (§6).

2 Related work

EHR foundation models and their evaluation. EHR foundation models pass codes and times to one encoder (Pang et al., 2021; Steinberg et al., 2024; Kraljevic et al., 2024); those that carry laboratory values quantize them into the same stream (Renc et al., 2024; Zhang et al., 2026), and LabTOP predicts laboratory results with one autoregressive model over the record (Im et al., 2025). These models are evaluated mainly on outcome prediction (Wornow et al., 2023; McDermott et al., 2021), as are the ICU benchmarks on the databases we use (Yèche et al., 2021; van de Water et al., 2024). We add a second evaluation, linear-probe recovery of laboratory values from the frozen representation, following the use of linear probes to read what a representation retains (Alain & Bengio, 2017).

The observation process in clinical data. The informatics work that motivates our design is summarized in §1; Subbaswamy & Saria (2020) argue for models that stay stable when practice changes. ICU time-series models often encode when and what was measured: GRU-D uses masks and the time since each variable's last measurement (Che et al., 2018), SeFT and STraTS treat each measurement as an element of a set or a (time, variable, value) triple (Horn et al., 2020; Tipirneni & Reddy, 2022), and Lipton et al. (2016) feeds missingness indicators alongside the values, and Jeanselme et al. (2025) models the observation and missingness processes jointly. These models read values and the observation pattern with one encoder; we ask whether the two should have separate encoders.

Shared and separate encoders. Multimodal models that tokenize every modality into one sequence use one encoder (Reed et al., 2022; Lu et al., 2023; Mizrahi et al., 2023; Chameleon Team, 2024; Zhou et al., 2025). Vision–language models often use separate image and text encoders (Radford et al., 2021; Jia et al., 2021; Zhang et al., 2022; Li et al., 2023; Wu et al., 2025), and CLIP-style models align them with a contrastive (InfoNCE) loss (Radford et al., 2021; Jia et al., 2021; Zhang et al., 2022; van den Oord et al., 2018), as our coupled model does. Our laboratory-value encoder is a set encoder with attention pooling (Zaheer et al., 2017; Lee et al., 2019). Joint training can under-use a modality (Wang et al., 2020; Huang et al., 2022), with remedies from gradient surgery (Yu et al., 2020; Wang et al., 2021) to modulation (Peng et al., 2022); this work instead asks which token types to separate.

3 Setup

Input. The records store laboratory results as (analyte, value, time) entries. A row is one laboratory draw during one ICU stay: the results of at least five distinct analytes timestamped in the same minute (App. A). No model sees the laboratory values of earlier draws. There are 46 laboratory-value tokens, one per analyte slot shared by the three databases, each carrying the analyte's value and a presence indicator; analytes not measured in the draw are marked absent. One vital-sign token summarizes the vital signs from 60 minutes before to 60 minutes after the draw (24 five-minute steps, 8 channels) with a recurrent layer. The care-event input has two parts. A 57-dimensional care vector describes the draw itself: the 46 presence indicators, the fraction of the 46 analytes measured, which of the 8 vital-sign channels were recorded in the window and their fraction, and the elapsed time since ICU admission. A *care-event sequence* gives the stay's recent history: up to 128 laboratory orders and medication starts from the 24 hours before the draw, from a vocabulary of 135 codes shared by the three databases, each embedded from its code and from the logarithm of the hours between it and the draw. Every model receives all of this: the single-stream transformer reads 176 tokens (46 laboratory values, one vital-sign summary, the care vector and 128 care events).

Models. The *single-stream model* passes all 176 tokens of a row through one two-layer transformer and mean-pools them into a 128-dimensional representation. The two-encoder models split the row. A *laboratory-value encoder* reads the laboratory-value tokens with a set encoder (each token is processed separately, then pooled with learned attention over the measured analytes), adds the vital-sign token, and outputs 128 dimensions. A *care-event encoder* reads the care vector and the care-event sequence with a two-layer transformer (d=128, four heads), mean-pooled to 64 dimensions. All models reconstruct a random 30% of the row's measured values, hidden at the input, together with the vital-sign summary and the 46 presence indicators, and predict 15% of the care events, hidden at the input. Both two-encoder models add a contrastive (InfoNCE) loss between the laboratory-value and vital-sign parts of the laboratory-value encoder (the *laboratory–vital contrastive loss*). The *decoupled model* trains the two encoders without further interaction; the *coupled model* adds a contrastive loss that aligns the two encoders' outputs for the same row (the *cross-encoder contrastive loss*; Figure 1b).

Downstream and recovery models. In the models used for downstream prediction (*downstream models*), the single-stream model's decoders read its 128-dimensional representation, while each two-encoder decoder reads the layer below the output we probe, so that output is trained only by the contrastive losses. For laboratory-value recovery we retrain every model with its decoders reading the probed output, and without dropout, so that the probed layer is trained by reconstruction in every model (*recovery models*).

Differences between the models. All models share data, tokenization, optimizer (AdamW, learning rate 2×10^{-3}, batch 2048) and 60 training epochs. Only the single-stream downstream model uses dropout; only the two-encoder models L_2-normalize their outputs. Trained-parameter counts per model are in Table 5 (App. B). The downstream representation is 128-dimensional for the single-stream model and 192-dimensional (both outputs concatenated) for the two-encoder models; §4 tests whether this width matters.

Databases and evaluation settings. We use eICU v2.0 (Pollard et al., 2018; 2019), MIMIC-IV v2.2 (Johnson et al., 2023b;a) and HiRID v1.1.1 (Hyland et al., 2020; Faltys et al., 2021), all distributed through PhysioNet (Pollard et al., 2026; Goldberger et al., 2000). eICU has 199 hospitals and 5.6 rows per stay (median 3). MIMIC-IV comes from one hospital with 17 ICU care units and has 12.0 rows per stay (median 5). HiRID comes from one center, covers the first 24h of each stay, and has 5.3 rows per stay (median 5). Mortality is ICU mortality on eICU, in-hospital mortality on MIMIC-IV and discharge status on HiRID. *In distribution* means rows from stays of the encoder's own database that were not used for training. *Held-out sites* are eICU hospitals and MIMIC-IV care units that no encoder saw in training. *Cross-database transfer* applies a frozen encoder trained on one database to another database's rows and fits a new probe there. A *cell* is one task, or one set of scored analytes, evaluated in one setting with encoders from one source database; in-distribution and held-out-site cells together are the *within-database* cells.

Probes and statistics. *Laboratory-value recovery* fits a ridge regression (α=10) from a frozen representation to each of the 20 analytes measured in the most rows, predicts it out of fold over five folds, and averages R^2 over analytes. It reads the single-stream representation, or the laboratory-value encoder output of a two-encoder model; the care-event output is not used. The *recovery gap* is coupled minus single-stream recovery, and the *downstream margin* is coupled minus single-stream AUROC, unless another model is named. Folds are grouped by stay (*stay-grouped folds*; 4% of eICU and 17% of MIMIC-IV patients have more than one stay). After cross-database transfer, folds are grouped by the target database's stays. *Downstream probes* are logistic regressions on the full representation, fitted on training-split rows and scored by AUROC on evaluation rows. Each contrast is paired over ten seeds with a 95% t-interval across seeds (a *seed-paired* interval), and we call a difference *significant* when that interval excludes zero. These intervals are over training seeds, not over patients, hospitals or databases, cells that share encoders are not independent tests, and we apply no correction across cells.

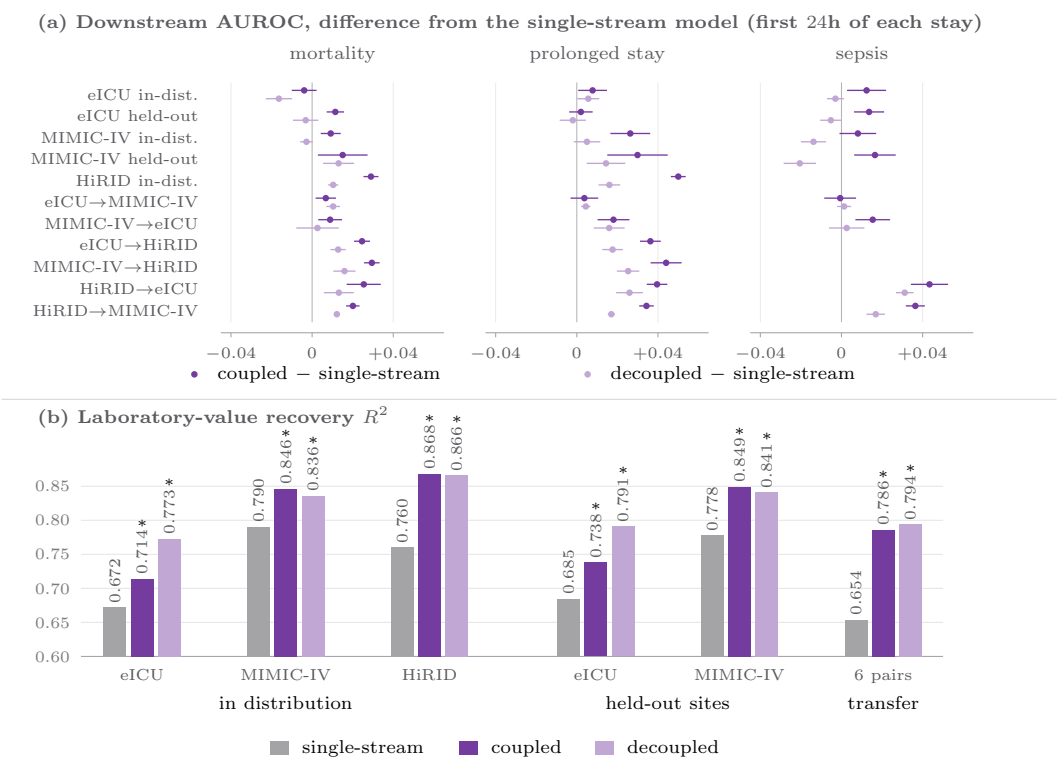

Figure 2: **Separate encoders for laboratory values and care events improve downstream prediction and laboratory-value recovery.** (a) Each of the 30 cells of the three standard tasks: the difference in AUROC from the single-stream model, for a logistic probe on the frozen representation fitted and scored on rows from the first 24h of each stay. Dot: the mean over ten seeds; bar: the seed-paired 95% t-interval; vertical line: zero. Rows are in the order of Table 6, and sepsis has no HiRID cells, so three of its rows are empty. (b) Laboratory-value recovery: R^2 of a ridge probe from the frozen representation to each of the 20 most-measured analytes, averaged over analytes and ten seeds (recovery models, stay-grouped folds; Table 2); transfer is the mean over the six source–target pairs. The R^2 axis starts at 0.60, so bar heights are not proportional to the values. $*$: the seed-paired 95% interval of the difference from the single-stream model lies above zero (ten seeds; for transfer, each seed's difference is first averaged over the six pairs).

4 Downstream tasks

The coupled model is ahead on all three standard tasks. The three *standard tasks* are mortality, prolonged stay (> 3 days) and sepsis; we fit and score them on rows from the first 24h of each stay. Elapsed time is part of the care vector, so without this restriction every prolonged-stay row after day three is positive by construction (App. E); within the first 24h, elapsed time alone gives 0.535 AUROC for prolonged stay on eICU. Coupled minus single-stream AUROC is above zero in 25 of 30 cells and below zero in none (Table 1, Figure 2a). Averaged over the cells that use each source database's encoders (t-intervals over cells), it is $+0.010$ $[+0.003, +0.018]$ for eICU, $+0.020$ $[+0.013, +0.028]$ for MIMIC-IV and $+0.035$ $[+0.027, +0.043]$ for HiRID encoders (App. D). Figure 2a gives every cell: the five whose interval includes zero involve only eICU and MIMIC-IV, the two databases with the smallest margins, and every cell involving HiRID, in distribution and transferred in either direction, lies above zero.

The margin holds on the five preregistered labels. Before probing, we preregistered three further tasks with five labels: death within 24h, vasopressor start within 12h and within 24h, and acute kidney injury (AKI) within 48h under the creatinine criterion and

Table 1: **Downstream AUROC of the downstream models.** Each task has 11 cells: in distribution on each database (3), held-out sites on eICU and MIMIC-IV (2) and six cross-database transfers; sepsis has 8, because HiRID has no comparable sepsis label. Columns: range of in-distribution coupled AUROC across databases; mean paired ΔAUROC over cells, ten seeds per cell (no interval, because cells share encoders and are not independent); and +/−/n.s., the number of cells whose ten-seed 95% t-interval lies above zero, lies below zero, and includes zero, which sum to the cell count. The three standard tasks are fitted and scored on rows from the first 24h of each stay (§4). The last five rows are the preregistered labels (App. C); AKI is acute kidney injury within 48h. Per-cell values for all eight tasks, the unrestricted standard-task values and the elapsed-time-only baselines are in App. D.

			coupled−single-stream		decoupled−single-stream	
Task	cells	coupled AUROC	mean Δ	+/−/n.s.	mean Δ	+/−/n.s.
Mortality (first 24h)	11	0.80–0.88	+0.016	10/0/1	+0.006	7/1/3
Prolonged stay (first 24h)	11	0.66–0.80	+0.027	9/0/2	+0.013	9/0/2
Sepsis (first 24h)	8	0.78–0.84	+0.018	6/0/2	+0.001	2/2/4
Death within 24h	11	0.86–0.92	+0.013	8/1/2	+0.001	4/3/4
Vasopressor 12h	11	0.79–0.89	+0.028	10/0/1	+0.010	7/0/4
Vasopressor 24h	11	0.78–0.87	+0.026	10/0/1	+0.008	6/1/4
AKI, creatinine	11	0.73–0.82	+0.010	6/0/5	+0.004	7/2/2
AKI, urine output	11	0.68–0.70	+0.011	8/0/3	+0.001	5/2/4
All	85		+0.019	67/1/17	+0.006	47/11/27

under the urine-output criterion of KDIGO (Kidney Disease: Improving Global Outcomes; App. C). Coupled minus single-stream is above zero in 20 of the 22 vasopressor cells, 8 of 11 cells of AKI by urine output, 6 of 11 cells of AKI by creatinine and 8 of 11 cells of death within 24h, and below zero in one cell, death within 24h on eICU in distribution (-0.011 $[-0.016, -0.006]$, 278 positive rows). Creatinine is also an encoder input, so the preregistration stated that an advantage under the creatinine criterion only would indicate carry-over of the input value. The coupled model is ahead under both criteria. Over all 85 cells the coupled model is ahead in 67 and behind in one (mean +0.019); the decoupled model is ahead in 47 and behind in 11 (mean +0.006).

Removing elapsed time from the care vector leaves the coupled model's margin unchanged. Elapsed time alone reaches 0.71–0.79 AUROC in distribution on vasopressor start, so the care-event encoder could be supplying mainly a clock. Retraining all three models with the elapsed time removed from the care vector leaves the coupled model's margin over the decoupled model unchanged: +0.008, +0.019 and +0.018 (eICU, MIMIC-IV, HiRID; in-distribution means over the eight tasks, seven on HiRID, which has no sepsis label, against +0.007, +0.016 and +0.018 with the clock), significant in 5 of 8, 8 of 8 and 7 of 7 of those cells. Its margin over the single-stream model also survives (+0.002, +0.019, +0.027).

The cross-encoder loss, not the readout width, carries the within-database margin. Separating the encoders is not enough on its own: over the 39 within-database cells of all eight tasks, the decoupled model is above zero in 12 and below in 9, and its mean margin is −0.001, while its gains concentrate in the transfer cells (35 of 46 above zero, mean +0.011; App. D). On the first-24h standard tasks it is ahead of the single-stream model in 18 of 30 cells and behind in 3 (mean +0.007, against +0.020 for the coupled model). The cross-encoder contrastive loss is what puts the two-encoder design above the single-stream model within a database. Readout width does not explain the margin either: against a single-stream model retrained at 192 dimensions, coupled minus single-stream is above zero in 24 of the 39 cells and below zero in one (App. B).

Table 2: **Laboratory-value recovery of the three recovery models.** Out-of-fold R^2 of a ridge probe (α=10) fitted from each frozen representation to each of the 20 most-measured analytes and averaged over analytes, with five folds grouped by stay. *Recovery models* are retrained with every decoder reading the probed output, under standard training, so that the probed layer is shaped by reconstruction in every model. *Single-stream*: one transformer over all of a row's tokens. *Coupled* and *decoupled*: separate laboratory-value and care-event encoders, with and without the cross-encoder contrastive loss. Levels are means over 10 seeds; **bold** marks the highest level in the row, which is not a significance claim. Differences are seed-paired, with 95% t-intervals and the number of the 10 seeds above zero; *: the interval excludes zero. *In distribution*: unseen stays of the encoder's own database. *Held-out sites*: eICU hospitals or MIMIC-IV care units no encoder saw in training. *Transfer*: the encoder is frozen, applied to the target database, and a new probe is fitted there.

database	evaluation	single-stream / coupled / decoupled	coupled − single-stream	decoupled − single-stream
eICU	in distribution	0.672 / 0.714 / **0.773**	+0.042* [+0.016, +0.068] 9/10	+0.101* [+0.088, +0.114] 10/10
eICU	held-out sites	0.685 / 0.738 / **0.791**	+0.053* [+0.029, +0.078] 9/10	+0.106* [+0.094, +0.117] 10/10
MIMIC-IV	in distribution	0.790 / **0.846** / 0.836	+0.056* [+0.046, +0.065] 10/10	+0.046* [+0.043, +0.049] 10/10
MIMIC-IV	held-out sites	0.778 / **0.849** / 0.841	+0.072* [+0.061, +0.083] 10/10	+0.063* [+0.059, +0.067] 10/10
HiRID	in distribution	0.760 / **0.868** / 0.866	+0.108* [+0.092, +0.124] 10/10	+0.106* [+0.089, +0.123] 10/10
eICU → MIMIC-IV	transfer	0.701 / 0.776 / **0.810**	+0.075* [+0.055, +0.096] 10/10	+0.109* [+0.097, +0.121] 10/10
eICU → HiRID	transfer	0.788 / 0.877 / **0.900**	+0.089* [+0.077, +0.101] 10/10	+0.112* [+0.105, +0.119] 10/10
MIMIC-IV → eICU	transfer	0.680 / **0.761** / **0.761**	+0.081* [+0.062, +0.100] 10/10	+0.082* [+0.066, +0.097] 10/10
MIMIC-IV → HiRID	transfer	0.802 / 0.896 / **0.901**	+0.095* [+0.084, +0.105] 10/10	+0.100* [+0.087, +0.112] 10/10
HiRID → eICU	transfer	0.434 / **0.672** / 0.669	+0.238* [+0.212, +0.263] 10/10	+0.235* [+0.211, +0.259] 10/10
HiRID → MIMIC-IV	transfer	0.518 / **0.737** / 0.726	+0.220* [+0.196, +0.244] 10/10	+0.208* [+0.184, +0.232] 10/10

5 Laboratory-value recovery

The downstream margin says the coupled representation predicts better on the eight labels we chose. It does not say what the representation holds, and for a frozen encoder that is the more durable question: whatever it discarded is unavailable to every task chosen later. The laboratory values are the natural thing to ask for back. They are the part of a row that measures the patient rather than the care delivered around them (§1) — creatinine for kidney function, lactate for perfusion — and they are quantitative, so a probe can be scored against the recorded value rather than against a label someone assigned. We therefore measure how much of a draw's measured values a linear probe can read back. What this measures is linear recoverability: a higher R^2 says the values are more accessible to a linear readout of the frozen representation, not that the representation is more faithful, more clinically meaningful or more disentangled.

Both two-encoder models recover more of the measured values in every cell. Both recover more than the single-stream model in every cell of Table 2, with every seed-paired interval excluding zero (Figure 2b). In distribution, coupled minus single-stream is +0.042 on eICU, +0.056 on MIMIC-IV and +0.108 on HiRID; at held-out sites it is +0.053 and +0.072. After cross-database transfer it is +0.075 to +0.238. The largest transfer gaps come from HiRID encoders, where the single-stream encoder recovers least after transfer (0.434 on eICU and 0.518 on MIMIC-IV, against 0.672 and 0.737 for the coupled encoder).

Hiding the scored analytes from the input and the loss leaves the gap in place. The scored values are part of each model's input, so the probe could reward copying them. Retraining every model with the scored analytes hidden from the input and from the loss (*hidden-target training*; App. F) leaves the gap in place and in fact larger: +0.089 on eICU, +0.062 on MIMIC-IV and +0.132 on HiRID in distribution, every interval excluding zero.

The gap survives when the probe reads the contrastively trained layer. In the downstream models, whose probed two-encoder layer is trained only by the contrastive losses, coupled minus single-stream recovery is +0.075 on HiRID and +0.024 on MIMIC-IV in distribution, and +0.027 and +0.037 at eICU and MIMIC-IV held-out sites; only the eICU in-distribution cell, at +0.006, has an interval including zero (App. D).

Table 3: **The recovery gap, decomposed.** Recovery models, stay-grouped folds, ten seeds. First and last rows: recovery R^2 of the single-stream and coupled models. Middle rows: seed-paired change as each difference is added in turn; the model after the third step is the decoupled model. *: 95% interval excludes zero.

	eICU		MIMIC-IV		HiRID
	in distribution	held-out sites	in distribution	held-out sites	in distribution
single-stream model	0.672	0.685	0.790	0.778	0.760
+ shared set encoder	+0.023*	+0.026*	+0.010	+0.014*	+0.026*
+ separate encoders	+0.052*	+0.055*	+0.015*	+0.025*	+0.057*
+ laboratory–vital contrastive loss	+0.026*	+0.024*	+0.022*	+0.023*	+0.024*
+ cross-encoder contrastive loss	−0.059*	−0.052*	+0.010	+0.009	+0.002
coupled model	0.714	0.738	0.846	0.849	0.868

6 Decomposition of the recovery gap

Table 3 adds the differences between the single-stream and coupled models one at a time.

- **Set encoder over all tokens.** Replacing the transformer by one set encoder over the same tokens adds +0.023 on eICU and +0.026 on HiRID, and +0.010 on MIMIC-IV (not significant in distribution).
- **Separate encoders.** Giving laboratory values and care events their own encoders is the largest step on two of the three databases: +0.052 on eICU, +0.057 on HiRID and +0.015 on MIMIC-IV, every interval excluding zero. Routing by token type is what matters, not the split itself: at a matched 192-dimensional representation, giving the laboratory-value encoder the analyte values beats a size-matched random split of the analytes for the decoupled model on all three databases (+0.064, +0.051, +0.081) and for the coupled model on two.
- **Laboratory–vital contrastive loss.** Adds +0.022 to +0.026 in every cell. Giving the same loss to the single-stream model does not close the gap: that model then recovers less on eICU (−0.018) and HiRID (−0.028) and is unchanged on MIMIC-IV (−0.004, not significant), and the coupled model stays ahead of it by +0.060, +0.136 and +0.060.
- **Cross-encoder contrastive loss.** Lowers recovery on eICU (−0.059 in distribution, −0.052 at held-out sites) and changes it by at most +0.010, not significantly, elsewhere. Its contribution is downstream, where it is the step that puts the two-encoder design above the single-stream model within a database (§4).

7 Discussion

Separate encoders for laboratory values and care events, joined by a cross-encoder contrastive loss, give frozen ICU representations with higher downstream AUROC than a single-stream transformer (+0.019 over 85 cells) and, with or without that loss, representations from which a linear probe recovers more of the measured values on all three databases, at held-out sites and after transfer.

Different objectives shape different properties. The decomposition does not point one way. The separation itself, the set encoder and the laboratory–vital contrastive loss carry the recovery gap, and the cross-encoder loss contributes nothing to it — on eICU it costs −0.059. What the cross-encoder loss buys is downstream: it is the step that puts the two-encoder design above the single-stream model within a database (§4). A representation is not simply better or worse, then; which objective is added decides which property improves, and an evaluation that reports only outcome metrics would have recorded the second effect and missed the first.

Implications for encoder design. The informatics work in §1 describes the record as a mix of measurements of the patient and traces of the care process that produced them.

Models of a laboratory draw usually put both in one encoder. Our results support a different default: encode laboratory values and care events separately and align the two encoders with a contrastive loss. The laboratory-value representation keeps more of the measured values, and the loss that aligns the two encoders is what converts the separation into downstream AUROC. The recovery gap after transfer is larger than on the source database for every source–target pair (Table 2), so the design matters most when a model trained at one institution is used at another. That is better transfer, not site invariance: the site stays predictable from every representation we report (Ethics statement). The design is also a factorization with an alignment term, not a separation of two independent processes — the cross-encoder loss deliberately pulls the two outputs together — and we do not claim the two are identifiable. A whole record serialized into one token sequence, as EHR foundation models do, is a different architecture; we do not test it.

Limitations. All three databases are ICU records, so the design is untested where ordering follows fewer protocols. Besides the separation, the two-encoder models differ from the single-stream model in encoder type, the laboratory–vital contrastive loss and output normalization; Table 3 separates the first two. The coupled model also carries 33,024 more trained parameters than the decoupled model (App. B), and we do not separate the cross-encoder loss from that difference.

Next steps. The laboratory-value encoder still reads one draw at a time; giving it the stay's earlier values, as the care-event encoder already reads earlier orders, tests whether the separation survives when both streams carry history. Records outside the ICU, where ordering follows fewer protocols, test whether the care process is separable where it is less routine.

The two processes in the record are not interchangeable, and a model that merges them at the input keeps less of what was measured. That cost is easy to miss when representations are compared on outcome metrics alone: over our cells the two designs differ by +0.019 in AUROC, and by as much as +0.238 in how much of a draw a linear probe can read back. Representations like these are trained once and reused, at another institution or for a task chosen later, and on that use the boundary between measurement and care process is worth drawing deliberately.

Reproducibility statement

Code for training and probing the three models, the run configurations for the reported results, the per-seed values behind every cell of the tables computed from the submitted run set, and the scripts that recompute those tables from them, including the stay-grouped recovery refits, the hidden-target analyses (App. F) and the preregistered task labels, will be released in a public repository. During the review period we will provide an anonymized copy to reviewers on request through the discussion forum. The three databases are credentialed, so the release will run end to end on synthetic data without credentialed access. Before each new analysis we reproduced a previously reported value: stay-grouped refits reproduce each run's row-level recovery to 10^{-5}, and downstream re-probes reproduce each run's AUROC to 2×10^{-3}, the tolerance set by the logistic solver. The preregistered labels and their data-mapping amendments were fixed before any probe was fitted (App. C). Results carried over from the earlier models with a multilayer-perceptron care encoder are labeled where they appear.

Ethics statement

This work uses de-identified electronic health records under each source's data-use agreement, which the citations in §3 follow; no new data were collected. The site can be predicted from our representations: on eICU, a classifier of the hospital reaches an AUROC of 0.90 to 0.94 from the frozen representation of every model we report (ten seeds each). None of these representations should be treated as site-anonymized.

AI use statement

In this work, we used generative AI tools to proofread and edit the text, to write the code that generates the figures and tables, and to run the analyses that produce the reported numbers. We did not use generative AI tools to generate synthetic data sets, to develop theoretical models or conceptual frameworks, or to formulate or prove mathematical claims. Every reported number is produced by the released analysis code and traced to a stored results file; the authors verified them and are responsible for the contents of this submission.

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.

Guillaume Alain and Yoshua Bengio. Understanding intermediate layers using linear classifier probes. In *International Conference on Learning Representations (ICLR), Workshop Track*, 2017. arXiv:1610.01644.

Chameleon Team. Chameleon: Mixed-modal early-fusion foundation models. *arXiv preprint arXiv:2405.09818*, 2024.

Zhengping Che, Sanjay Purushotham, Kyunghyun Cho, David Sontag, and Yan Liu. Recurrent neural networks for multivariate time series with missing values. *Scientific Reports*, 8(1):6085, 2018. doi: 10.1038/s41598-018-24271-9.

Martin Faltys, Marc Zimmermann, Xinrui Lyu, Matthias Hüser, Stephanie Hyland, Gunnar Rätsch, and Tobias Merz. HiRID, a high time-resolution ICU dataset (version 1.1.1). PhysioNet, 2021. doi: 10.13026/nkwc-js72.

Ary L. Goldberger, Luis A. N. Amaral, Leon Glass, Jeffrey M. Hausdorff, Plamen Ch. Ivanov, Roger G. Mark, Joseph E. Mietus, George B. Moody, Chung-Kang Peng, and H. Eugene Stanley. PhysioBank, PhysioToolkit, and PhysioNet: components of a new research resource for complex physiologic signals. *Circulation*, 101(23):e215–e220, 2000. doi: 10.1161/01.CIR.101.23.e215.

Benjamin A. Goldstein, Nrupen A. Bhavsar, Matthew Phelan, and Michael J. Pencina. Controlling for informed presence bias due to the number of health encounters in an electronic health record. *American Journal of Epidemiology*, 184(11):847–855, 2016. doi: 10.1093/aje/kww112.

Sebastien Haneuse and Michael Daniels. A general framework for considering selection bias in EHR-based studies: what data are observed and why? *eGEMs (Generating Evidence & Methods to improve patient outcomes)*, 4(1):16, 2016. doi: 10.13063/2327-9214.1203.

Max Horn, Michael Moor, Christian Bock, Bastian Rieck, and Karsten Borgwardt. Set functions for time series. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 4353–4363, 2020.

George Hripcsak and David J. Albers. Next-generation phenotyping of electronic health records. *Journal of the American Medical Informatics Association*, 20(1):117–121, 2013. doi: 10.1136/amiajnl-2012-001145.

Yu Huang, Junyang Lin, Chang Zhou, Hongxia Yang, and Longbo Huang. Modality competition: What makes joint training of multi-modal network fail in deep learning? (provably). In *ICML*, 2022. arXiv:2203.12221.

Stephanie L. Hyland, Martin Faltys, Matthias Hüser, Xinrui Lyu, Thomas Gumbsch, Cristóbal Esteban, Christian Bock, Max Horn, Michael Moor, Bastian Rieck, Marc Zimmermann, Dean Bodenham, Karsten Borgwardt, Gunnar Rätsch, and Tobias M. Merz. Early prediction of circulatory failure in the intensive care unit using machine learning. *Nature Medicine*, 26(3):364–373, 2020. doi: 10.1038/s41591-020-0789-4.

Sujeong Im, Jungwoo Oh, and Edward Choi. LabTOP: A unified model for lab test outcome prediction on electronic health records. In *Proceedings of the Conference on Health, Inference, and Learning*, volume 287 of *Proceedings of Machine Learning Research*, pp. 817–843, 2025.

Vincent Jeanselme, Glen Martin, Matthew Sperrin, Niels Peek, Brian Tom, and Jessica Barrett. Prediction of survival outcomes under clinical presence shift: A joint neural network architecture. *arXiv preprint arXiv:2508.05472*, 2025.

Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc V. Le, Yunhsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In *ICML*, 2021.

Alistair Johnson, Lucas Bulgarelli, Tom Pollard, Steven Horng, Leo A. Celi, and Roger Mark. MIMIC-IV (version 2.2). PhysioNet, 2023a. doi: 10.13026/6mm1-ek67.

Alistair E. W. Johnson, Lucas Bulgarelli, Lu Shen, Alvin Gayles, Ayad Shammout, Steven Horng, Tom J. Pollard, Sicheng Hao, Benjamin Moody, Brian Gow, Li-wei H. Lehman, Leo A. Celi, and Roger G. Mark. MIMIC-IV, a freely accessible electronic health record dataset. *Scientific Data*, 10(1):1, 2023b. doi: 10.1038/s41597-022-01899-x.

Zeljko Kraljevic, Dan Bean, Anthony Shek, Rebecca Bendayan, Harry Hemingway, Joshua Au Yeung, Alexander Deng, Alfred Balston, Jack Ross, Esther Idowu, James T. Teo, and Richard J. B. Dobson. Foresight—a generative pretrained transformer for modelling of patient timelines using electronic health records: a retrospective modelling study. *The Lancet Digital Health*, 6(4):e281–e290, 2024. doi: 10.1016/S2589-7500(24)00025-6.

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. doi: 10.1038/s41746-024-01037-4.

Juho Lee, Yoonho Lee, Jungtaek Kim, Adam R. Kosiorek, Seungjin Choi, and Yee Whye Teh. Set transformer: A framework for attention-based permutation-invariant neural networks. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 3744–3753, 2019.

Chunyuan Li, Cliff Wong, Sheng Zhang, Naoto Usuyama, Haotian Liu, Jianwei Yang, Tristan Naumann, Hoifung Poon, and Jianfeng Gao. LLaVA-Med: Training a large language-and-vision assistant for biomedicine in one day. In *Advances in Neural Information Processing Systems (Datasets and Benchmarks Track)*, volume 36, pp. 28541–28564, 2023. doi: 10.52202/075280-1240.

Zachary C. Lipton, David C. Kale, and Randall Wetzel. Directly modeling missing data in sequences with RNNs: Improved classification of clinical time series. In *Proceedings of the 1st Machine Learning for Healthcare Conference (MLHC)*, volume 56 of *PMLR*, pp. 253–270, 2016.

Jiasen Lu, Christopher Clark, Rowan Zellers, Roozbeh Mottaghi, and Aniruddha Kembhavi. Unified-IO: A unified model for vision, language, and multi-modal tasks. In *International Conference on Learning Representations (ICLR)*, 2023. arXiv:2206.08916.

Matthew B. A. McDermott, Bret Nestor, Evan Kim, Wancong Zhang, Anna Goldenberg, Peter Szolovits, and Marzyeh Ghassemi. A comprehensive EHR timeseries pre-training benchmark. In *Proceedings of the Conference on Health, Inference, and Learning (CHIL)*, pp. 257–278, 2021. doi: 10.1145/3450439.3451877.

David Mizrahi, Roman Bachmann, Oguzhan Fatih Kar, Teresa Yeo, Mingfei Gao, Afshin Dehghan, and Amir Zamir. 4M: Massively multimodal masked modeling. In *Advances in Neural Information Processing Systems*, volume 36, pp. 58363–58408, 2023. doi: 10.52202/075280-2544.

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 *Proceedings of the Machine Learning for Health (ML4H) Symposium*, volume 158 of *Proceedings of Machine Learning Research*, pp. 239–260, 2021. arXiv:2111.08585.

Xiaokang Peng, Yake Wei, Andong Deng, Dong Wang, and Di Hu. Balanced multimodal learning via on-the-fly gradient modulation. In *CVPR*, 2022. arXiv:2203.15332.

Tom Pollard, Alistair Johnson, Jesse Raffa, Leo A. Celi, Omar Badawi, and Roger Mark. eICU collaborative research database (version 2.0). PhysioNet, 2019. doi: 10.13026/C2WM1R.

Tom Pollard, Benjamin E. Moody, Li-wei H. Lehman, Brian J. Gow, Chrystinne Fernandes, Chen Xie, Alistair Johnson, Roger G. Mark, and Thomas Heldt. PhysioNet as a global platform for biomedical research. *Nature Health*, 1(8):792–795, 2026. doi: 10.1038/s44360-026-00096-z.

Tom J. Pollard, Alistair E. W. Johnson, Jesse D. Raffa, Leo A. Celi, Roger G. Mark, and Omar Badawi. The eICU collaborative research database, a freely available multi-center database for critical care research. *Scientific Data*, 5(1):180178, 2018. doi: 10.1038/sdata.2018.178.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *ICML*, 2021.

Scott Reed, Konrad Zolna, Emilio Parisotto, Sergio Gomez Colmenarejo, Alexander Novikov, Gabriel Barth-Maron, Mai Gimenez, Yury Sulsky, Jackie Kay, Jost Tobias Springenberg, Tom Eccles, Jake Bruce, Ali Razavi, Ashley Edwards, Nicolas Heess, Yutian Chen, Raia Hadsell, Oriol Vinyals, Mahyar Bordbar, and Nando de Freitas. A generalist agent. *Transactions on Machine Learning Research*, 2022. arXiv:2205.06175.

Pawel Renc, Yugang Jia, Anthony E. Samir, Jaroslaw Was, Quanzheng Li, David W. Bates, and Arkadiusz Sitek. Zero shot health trajectory prediction using transformer. *npj Digital Medicine*, 7:256, 2024. doi: 10.1038/s41746-024-01235-0.

Ethan Steinberg, Jason Fries, Yizhe Xu, and Nigam Shah. MOTOR: A time-to-event foundation model for structured medical records. In *International Conference on Learning Representations (ICLR)*, 2024. arXiv:2301.03150.

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. doi: 10.1093/biostatistics/kxz041.

Sindhu Tipirneni and Chandan K. Reddy. Self-supervised transformer for sparse and irregularly sampled multivariate clinical time-series. *ACM Transactions on Knowledge Discovery from Data*, 16(6):1–17, 2022. doi: 10.1145/3516367.

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*, 2024. URL https://openreview.net/forum?id=ox2ATRM90I.

Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018.

Weiyao Wang, Du Tran, and Matt Feiszli. What makes training multi-modal classification networks hard? In *CVPR*, 2020. arXiv:1905.12681.

Zirui Wang, Yulia Tsvetkov, Orhan Firat, and Yuan Cao. Gradient vaccine: Investigating and improving multi-task optimization in massively multilingual models. In *ICLR*, 2021.

Michael Wornow, Rahul Thapa, Ethan Steinberg, Jason A. Fries, and Nigam H. Shah. EHRSHOT: An EHR benchmark for few-shot evaluation of foundation models. In *Advances in Neural Information Processing Systems (Datasets and Benchmarks Track)*, volume 36, pp. 67125–67137, 2023. doi: 10.52202/075280-2933.

Chaoyi Wu, Xiaoman Zhang, Ya Zhang, Yanfeng Wang, and Weidi Xie. Towards generalist foundation model for radiology by leveraging web-scale 2D and 3D medical data. *Nature Communications*, 16:7866, 2025. doi: 10.1038/s41467-025-62385-7.

Hugo Yèche, Rita Kuznetsova, Marc Zimmermann, Matthias Hüser, Xinrui Lyu, Martin Faltys, and Gunnar Rätsch. HiRID-ICU-Benchmark – a comprehensive machine learning benchmark on high-resolution ICU data. In *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*, volume 1, 2021.

Tianhe Yu, Saurabh Kumar, Abhishek Gupta, Sergey Levine, Karol Hausman, and Chelsea Finn. Gradient surgery for multi-task learning. In *NeurIPS*, 2020.

Manzil Zaheer, Satwik Kottur, Siamak Ravanbakhsh, Barnabás Póczos, Ruslan Salakhutdinov, and Alexander Smola. Deep sets. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

Andrew Zhang, Tong Ding, Sophia J. Wagner, Caiwei Tian, Ming Y. Lu, Rowland Pettit, Joshua E. Lewis, Alexandre Misrahi, Dandan Mo, Long Phi Le, and Faisal Mahmood. A multimodal and temporal foundation model for virtual patient representations at healthcare system scale. *arXiv preprint arXiv:2604.18570*, 2026.

Yuhao Zhang, Hang Jiang, Yasuhide Miura, Christopher D. Manning, and Curtis P. Langlotz. Contrastive learning of medical visual representations from paired images and text. In *Machine Learning for Healthcare (MLHC)*, 2022. arXiv:2010.00747.

Chunting Zhou, Lili Yu, Arun Babu, Kushal Tirumala, Michihiro Yasunaga, Leonid Shamis, Jacob Kahn, Xuezhe Ma, Luke Zettlemoyer, and Omer Levy. Transfusion: Predict the next token and diffuse images with one multi-modal model. In *International Conference on Learning Representations (ICLR)*, 2025. arXiv:2408.11039.

A Probe definitions, fold scheme and recovery detail

Models, input, training regimes and probes are defined in §3. Every laboratory-value recovery number uses the readout below on a frozen representation.

Row construction. Laboratory entries are grouped by stay and by timestamp rounded to the minute; a group with results for at least five distinct analytes is one draw, and isolated tests are not rows. The 46 analyte slots are eICU's laboratory vocabulary; MIMIC-IV and HiRID results are mapped to them through a crosswalk (with unit conversion for HiRID), and all three databases use eICU's training-set median and scaled median absolute deviation to standardize values. A slot not measured in the draw holds 0 with presence 0. The vital-sign window is resampled to 5-minute bins over $[t-60, t+60]$ minutes around the draw time t, with its own presence mask; it is identical input for all three models. On eICU a stay contributes at most 30 draws, evenly spaced, and a draw needs at least one vital-sign channel in its window; HiRID draws are those between 12h before and 24h after admission.

Readout. A ridge regression (α=10) is fitted from the representation to each scored analyte with at least 300 observed rows, scored by out-of-fold R^2 over five folds, and averaged over analytes. In-distribution cells score rows from stays of the encoder's own database that were not used in training. Held-out-site cells score eICU hospitals or MIMIC-IV care units that no encoder saw in training. Cross-database cells apply the frozen encoder to another database and fit a new ridge there.

Table 4: **Stay-grouped against row-level folds**, in distribution, ten seeds. These are the *downstream* models, whose two-encoder probed layer is shaped only by the contrastive losses, so their levels differ from the recovery models of Table 2. Hidden-target rows score the held-out analytes of hidden-target-trained encoders; standard rows score the top-20 analytes of standard-trained encoders. Levels are row-level→grouped R^2. The last column is the grouped recovery gap and, in parentheses, the grouped gap as a percentage of the row-level gap. **Bold**: the highest grouped level in the row, which is not a significance claim.

training	database	single-stream	coupled	decoupled	recovery gap (grouped/row-level)
hidden-target	eICU	0.278 → 0.226	0.368 → **0.321**	0.354 → 0.305	+0.095 (105%)
	MIMIC-IV	0.445 → 0.433	0.483 → **0.474**	0.477 → 0.467	+0.041 (109%)
	HiRID	0.350 → 0.341	0.454 → 0.448	0.458 → **0.453**	+0.108 (103%)
standard	eICU	0.718 → 0.701	0.759 → **0.745**	0.736 → 0.720	+0.044 (108%)
	MIMIC-IV	0.825 → **0.821**	0.799 → 0.796	0.748 → 0.745	−0.025 (96%)
	HiRID	0.775 → 0.772	0.830 → **0.828**	0.829 → 0.826	+0.056 (103%)

Fold scheme. In-distribution and held-out-site cells use five folds grouped by stay, so the rows of one stay (5.3 to 12.0 per stay on average) never fall in both the ridge's training and scoring folds. Table 4 compares them with row-level folds on identical representations, rows and analytes: grouping lowers recovery by up to 0.05 (eICU) and leaves each gap at 96 to 109% of its row-level size. That comparison was run on the earlier models with a multilayer-perceptron care encoder; it is a property of the fold scheme rather than of the encoder, and we did not repeat it on the models we report. Hidden-target cross-database cells group folds by target stay.

Care-event features and laboratory values (§1). The introduction compares two model-free ridge regressions on eICU. A ridge from care-event features (ordering and timing) to each laboratory value gives R^2=0.05 in distribution; a ridge from the other observed laboratory values of the same row gives 0.59. Both are averaged over the 20 most-measured analytes, with folds grouped by stay; over all 46 analytes they are 0.05 and 0.52. The second quantity predicts each analyte from all other observed values, on every row on which it was measured, with no analyte hidden. It is therefore not the recovery of the hidden-target probe, which predicts only the held-out analytes from what hidden-target training leaves visible. Permuting which stay's care events are paired with which stay's laboratory values (1000 permutations) gives null recoveries centered at −0.083 with maximum −0.059, so the observed 0.05 exceeds every permutation ($p \leq 0.001$).

Probe conventions. The conventions below differ in whether the scored analytes' values enter the input, whether they enter the reconstruction loss, where each decoder attaches relative to the probed representation.

1. **Visible-target probe** (§5). Standard training: values in input and loss; scored on the top-20 analytes. The single-stream decoder attaches at the probed representation; each two-encoder decoder attaches one layer below it (the code release gives the code locations).
2. **Hidden-target probe** (§F.1). Hidden-target training removes the 23 held-out analytes' values from the input, excludes them from the loss, and removes them again when representations are computed, including when a frozen encoder is applied to another database. Their presence indicators still enter every model. Scored on the held-out analytes.
3. **Loss-only training**. The held-out analytes are excluded from the loss, but their values stay in every model's input. Compared with hidden-target training on the same analytes (separate run sets, row-level folds), leaving the values in the input raises recovery for every model and raises the two-encoder models more: the recovery gap is +0.124/+0.071/+0.114 (eICU/MIMIC-IV/HiRID), against +0.090/+0.038/+0.104 under hidden-target training (Table 4, row-level levels). Both figures come from the earlier models with a multilayer-perceptron care encoder; the convention in item 4 below, not this one, is what Tables 2 and 3 use.

4. **Decoder attachment A.** Standard training, values in input and loss as in (1), with every model's decoder attached at the probed representation. This is the convention of the recovery models (§3), and it is the one under which Table 2 and Table 3 are computed.

Conventions 1 and 4 differ only in where the two-encoder decoders attach, and the paper reports 4 as its recovery convention because it treats every model alike.

Held-out analytes per database. The coverage split is computed within each database, so the held-out analytes differ. eICU: Hct, MCV, RBC, −monos, −polys, BUN, O_2 sat, PT-INR, PTT, RDW, albumin, alkaline phosphatase, anion gap, bicarbonate, chloride, magnesium, $paCO_2$, paO_2, phosphate, sodium, total protein, troponin-I, urine specific gravity. MIMIC-IV: Hgb, MCHC, RBC, −eos, −monos, AST, BUN, FiO_2, MPV, O_2 sat, PT-INR, PTT, RDW, anion gap, calcium, chloride, $paCO_2$, paO_2, platelets, sodium, total bilirubin, total protein, urine specific gravity. HiRID: Hgb, MCH, −basos, −eos, −monos, ALT, base excess, HCO_3, MPV, PTT, RDW, bedside glucose, calcium, creatinine, glucose, lactate, magnesium, $paCO_2$, paO_2, platelets, sodium, total protein, urine specific gravity. Abbreviations: Hct hematocrit; Hgb hemoglobin; RBC red blood cell count; MCV, MCH, MCHC mean corpuscular volume, hemoglobin and hemoglobin concentration; RDW red-cell distribution width; MPV mean platelet volume; −monos, −polys, −eos, −basos white-cell differential percentages (source variable names); BUN blood urea nitrogen; PT-INR prothrombin time as international normalized ratio; PTT partial thromboplastin time; AST, ALT aspartate and alanine aminotransferase; $paCO_2$, paO_2 arterial partial pressures; FiO_2 inspired oxygen fraction. On HiRID, HCO_3 and bicarbonate, and glucose and bedside glucose, are separate source variables. Analytes with fewer than 300 observed rows are not scored, leaving 20, 17 and 13 of the 23 in distribution.

Visible-target recovery per analyte. This breakdown was run on the earlier models with a multilayer-perceptron care encoder. Scored with stay-grouped folds on the standard-trained encoders, the analytes on which the single-stream model recovers more than the coupled model are, largest difference first (in parentheses, the per-analyte recovery gap summed over these analytes, against the sum over the other analytes): eICU MCHC, sodium, MCV, chloride, Hct, MCH, Hgb, RBC, anion gap, creatinine (10 of 20; −0.71 against +1.60); MIMIC-IV MCH, MCV, MCHC, sodium, chloride, anion gap, potassium, creatinine, BUN, RBC, Hct, Hgb, bicarbonate (13 of 20; −1.01 against +0.51); HiRID MCH, MCV, MCHC, base excess, pH, $paCO_2$, bicarbonate, creatinine (8 of 20; −0.55 against +1.68). Scored instead on the held-out analytes, the coupled model recovers more than the single-stream model on every database, while the decoupled model recovers less on eICU (−0.014) and MIMIC-IV (−0.028). At held-out sites, the recovery gap on those analytes is +0.025 (eICU, 9 of 10 seeds) and +0.011 (MIMIC-IV, 10 of 10).

B Representation width and parameter counts

Matched readout width. The two-encoder downstream representation is 192-dimensional and the single-stream one 128-dimensional. With the single-stream model retrained at 192 dimensions (ten paired seeds; the 39 within-database cells of all eight tasks), the coupled model's interval lies above the single-stream model's in 24 cells, below in 1 and includes zero in 14. The one cell below is death within 24h on eICU in distribution (−0.009), the same cell that is below at width 128. All seven HiRID cells are above zero, as are 12 of the 16 MIMIC-IV cells.

Parameter counts. Table 5 gives the counts behind §3. Each model is built as in training, one forward and backward pass is run, and only parameters that receive a gradient are counted. The single-stream implementation constructs 74,168 parameters it never uses; we exclude them, which lowers the single-stream count. The three models are within 6% of each other in trained parameters, so neither design is the larger one: the coupled model carries 34,050 more than the single-stream model and the decoupled model 1,026 more.

Table 5: **Effective (gradient-carrying) parameters per model.** Measured by one forward and backward pass on each model as the training script builds it. Raw counts every parameter constructed; effective counts only those that receive gradient; the remainder are constructed but never trained.

model	raw	effective	never trained
single-stream	733,988	659,820	74,168
coupled	693,870	693,870	0
decoupled	685,614	660,846	24,768

The coupled model has 33,024 more effective parameters than the decoupled model, not only the coupling head's 8,256: the coupling head is the only consumer of the care-event encoder's projection head, so adding it also trains 24,768 parameters that the decoupled model leaves untrained. A parameter-matched control would therefore need 33,024 extra parameters, which we do not run: the comparison between the coupled and decoupled models does not separate the cross-encoder loss from these parameters. The comparisons against the single-stream model, which carry the paper's claims, are not affected; all three models are within 6% of each other.

C Preregistered downstream tasks

Three tasks were preregistered: death within 24h, vasopressor start (primary horizon 24h, secondary 12h), and acute kidney injury (AKI) within 48h under two criteria of KDIGO (Kidney Disease: Improving Global Outcomes). They give the five labels of Table 8. The encoders, probe, statistics and reporting rule were written down before any label was built or any representation was probed on these tasks, and all five labels are reported. Amendments were restricted to data mapping and were logged before any probe was fitted (below).

Definitions. A label is attached to each row, that is, to one laboratory draw at time t; ineligible rows are dropped. **Death within 24h**: the stay's mortality flag (ICU, in-hospital or discharge status, by database; §3) is positive and death falls in $(t, t{+}24\text{h}]$; rows at or after the end of the stay are ineligible. **Vasopressor start** within horizon $H \in \{24\text{h}, 12\text{h}\}$: the stay's first vasopressor start falls in $(t, t{+}H]$; rows at or after a first start are ineligible; on eICU, only hospitals with medication data for more than half their stays are used, for fitting and evaluation alike. **AKI within 48h**, KDIGO stage ≥ 1, with the two criteria scored separately. Creatinine: $\geq 1.5\times$ the minimum over the prior 7 days, or a rise of $\geq 0.3\,\text{mg/dL}$ within 48h; stays whose first creatinine exceeds $4\,\text{mg/dL}$ are excluded. Urine output: $< 0.5\,\text{ml/kg/h}$ over 6h. For each criterion, rows at or after its first onset are ineligible. Every label is reported beside the same probe fitted on elapsed time alone. HiRID rows cover only the first 24h of a stay, but its labels can use later data. Because creatinine is an encoder input, we fixed in advance that a two-encoder advantage under the creatinine criterion but not under urine output would be read as carry-over of the input creatinine value. The coupled model's advantage appears under both criteria and under neither is any cell below zero: its interval lies above zero in 6 of the 11 creatinine cells and in 8 of the 11 urine-output cells. The decoupled model is above zero in 7 creatinine cells and 5 urine-output cells and below zero in 2 of each, so it has no consistent advantage under either criterion and the preregistered carry-over reading is not triggered for either model (Table 8).

Data mapping. Death time: eICU unit discharge offset for ICU deaths; MIMIC-IV hospital death time relative to ICU admission; HiRID length of stay for discharge status "dead". Vasopressors: norepinephrine, epinephrine, dopamine and phenylephrine on eICU and MIMIC-IV, norepinephrine and epinephrine on HiRID (vasopressin, preregistered, is absent from every processed medication table). Creatinine from each database's laboratory table (HiRID μmol/L /88.4). Urine output: MIMIC-IV charted urinary catheter (Foley), void, operating-room and post-anesthesia urine; eICU intake–output entries labeled as urine, Foley or void; HiRID's hourly urine rate. Weight: the source's recorded patient weight.

Amendments. (A1, before any label was built) Identifier and time-origin mapping. Rows with an observed creatinine coincide with a creatinine timestamp in the source table in 100% of sampled rows on all three databases. (A2, after the first label build and before any probe) Urine is charted intermittently as accumulated volume, so the first build's hourly bins read every charting gap as low urine output (onset in 79 to 90% of stays). Each charted volume is now spread over the interval since the previous record, and a 6h window counts only if it is fully covered with no interval longer than 12h. (A3, before AKI by urine output was probed) HiRID's hourly urine variable is a rate in ml/h, so its volume is rate times interval. After A2 and A3, urine-output onset occurs in 64%, 65% and 40% of evaluable stays (eICU, MIMIC-IV, HiRID). As a check only, never for labels, we compared stays with the AKI cohort of an independent ICU benchmark pipeline, which combines the creatinine and urine criteria: stay-level agreement is 85%, 72% and 87%. Because that cohort combines both criteria, this is not a like-for-like comparison with the urine-output label.

D Downstream results per cell

All cells use the downstream probe of §3 (logistic regression) on the full representation: the single-stream 128-dimensional representation, or the concatenated 128-dimensional laboratory-value and 64-dimensional care-event representations. Each cell has ten paired seeds and a 95% t-interval across seeds; * marks an interval excluding zero. Before any new task was probed, re-running the probe on each run's mortality label reproduced its originally reported AUROC to 10^{-6} in every model, seed and cell. Table 7 gives the three standard tasks on standard-trained encoders without the 24h restriction, so prolonged stay includes the elapsed-time shortcut of App. E. Table 8 gives the preregistered labels with the elapsed-time-only baseline. Without the 24h restriction the margins are smaller: on the 14 within-database cells of Table 7 the coupled model's interval lies above the single-stream model's in 7, below in 1 and includes zero in 6, and the decoupled model's lies above in 2 and below in 7. The restriction of Table 6 removes the elapsed-time shortcut that prolonged stay carries in these unrestricted cells.

Recovery from the downstream models. The recovery numbers of §5 come from the recovery models, whose decoders read the probed representation. Probing the downstream models instead (the encoders of this appendix; the two-encoder layer they expose is shaped only by the contrastive losses) gives coupled minus single-stream recovery of $+0.006$ $[-0.010, +0.022]$ on eICU, $+0.024$ $[+0.012, +0.035]$ on MIMIC-IV and $+0.075$ $[+0.067, +0.084]$ on HiRID in distribution, and $+0.027$ $[+0.012, +0.042]$ and $+0.037$ $[+0.018, +0.055]$ at eICU and MIMIC-IV held-out sites (ten paired seeds, stay-grouped folds, top-20 analytes, as in Table 2). The decoupled model behaves differently here: $+0.034$ on eICU and $+0.074$ on HiRID, but -0.042 on MIMIC-IV, the same asymmetry between the two two-encoder models that the within-database downstream cells show (§4).

Elapsed-time shortcut removed. Table 6 refits and scores the three standard tasks on rows within the first 24h of the ICU stay; rows scored on HiRID already end there. Its coupled-minus-single-stream prolonged-stay margins agree with those of App. E. The coupled model's interval lies above the single-stream model's in 25 of the 30 cells, below in none and includes zero in 5; for the decoupled model the counts are 18, 3 and 9.

E The elapsed-time shortcut in the prolonged-stay label

Downstream rows are laboratory draws, and the care vector includes elapsed time since ICU admission in days. The prolonged-stay label is positive when the stay exceeds three days, so every row drawn after day three is positive by construction. This appendix measures the shortcut and the downstream margin once it is removed.

Size of the shortcut. Table 9 fits the logistic probe of §3 directly on the care vector's features, on the same rows and splits.

Table 6: **Every cell behind the standard tasks.** Mortality, prolonged stay and sepsis on rows within the first 24h of the ICU stay, standard-trained encoders, probe refitted and scored on these rows only. Mean AUROC per model over ten seeds and seed-paired differences with 95% t-intervals; *: interval excludes zero. "in-dist.": in distribution; "held-out sites": eICU hospitals or MIMIC-IV care units unseen in training; "A→B": encoder trained on A, probe fitted and scored on B. HiRID has no comparable sepsis label. Interval bounds printed as ±0.000 are below 0.0005 in magnitude. **Bold**: the highest of the three model columns in the row, which is not a significance claim.

task	cell	single-stream	coupled	decoupled	coupled−single-stream	decoupled−single-stream
Mortality	eICU in-dist.	**0.803**	0.799	0.786	−0.004 [−0.010, +0.002]	−0.016 [−0.023, −0.010]*
	eICU held-out sites	0.835	**0.846**	0.832	+0.011 [+0.007, +0.016]*	−0.003 [−0.009, +0.003]
	MIMIC-IV in-dist.	0.839	**0.849**	0.837	+0.009 [+0.004, +0.014]*	−0.003 [−0.006, +0.000]
	MIMIC-IV held-out sites	0.665	**0.680**	0.678	+0.015 [+0.003, +0.027]*	+0.013 [+0.005, +0.021]*
	HiRID in-dist.	0.855	**0.884**	0.865	+0.029 [+0.025, +0.033]*	+0.010 [+0.008, +0.013]*
	eICU→MIMIC-IV	0.811	0.818	**0.822**	+0.007 [+0.002, +0.012]*	+0.010 [+0.007, +0.014]*
	MIMIC-IV→eICU	0.790	**0.799**	0.793	+0.009 [+0.003, +0.015]*	+0.003 [−0.008, +0.013]
	eICU→HiRID	0.822	**0.847**	0.835	+0.025 [+0.021, +0.029]*	+0.013 [+0.009, +0.017]*
	MIMIC-IV→HiRID	0.812	**0.841**	0.828	+0.029 [+0.026, +0.033]*	+0.016 [+0.011, +0.021]*
	HiRID→eICU	0.777	**0.802**	0.790	+0.025 [+0.017, +0.034]*	+0.013 [+0.006, +0.021]*
	HiRID→MIMIC-IV	0.808	**0.828**	0.821	+0.020 [+0.017, +0.023]*	+0.012 [+0.011, +0.014]*
Prolonged stay	eICU in-dist.	0.656	**0.664**	0.662	+0.008 [+0.001, +0.015]*	+0.006 [+0.000, +0.011]*
	eICU held-out sites	0.683	**0.685**	0.681	+0.002 [−0.004, +0.008]	−0.002 [−0.008, +0.004]
	MIMIC-IV in-dist.	0.701	**0.727**	0.706	+0.026 [+0.017, +0.036]*	+0.005 [−0.002, +0.012]
	MIMIC-IV held-out sites	0.657	**0.687**	0.672	+0.030 [+0.015, +0.045]*	+0.014 [+0.005, +0.024]*
	HiRID in-dist.	0.750	**0.800**	0.766	+0.050 [+0.046, +0.054]*	+0.016 [+0.011, +0.021]*
	eICU→MIMIC-IV	0.683	0.686	**0.687**	+0.004 [−0.003, +0.011]	+0.004 [+0.002, +0.007]*
	MIMIC-IV→eICU	0.649	**0.667**	0.665	+0.018 [+0.010, +0.026]*	+0.016 [+0.008, +0.024]*
	eICU→HiRID	0.721	**0.758**	0.739	+0.036 [+0.031, +0.041]*	+0.018 [+0.013, +0.023]*
	MIMIC-IV→HiRID	0.711	**0.755**	0.736	+0.044 [+0.036, +0.052]*	+0.025 [+0.020, +0.031]*
	HiRID→eICU	0.630	**0.670**	0.656	+0.040 [+0.035, +0.045]*	+0.026 [+0.019, +0.033]*
	HiRID→MIMIC-IV	0.669	**0.703**	0.686	+0.034 [+0.031, +0.038]*	+0.017 [+0.015, +0.019]*
Sepsis	eICU in-dist.	0.772	**0.785**	0.769	+0.012 [+0.003, +0.022]*	−0.003 [−0.007, +0.001]
	eICU held-out sites	0.779	**0.793**	0.774	+0.014 [+0.006, +0.021]*	−0.005 [−0.011, −0.000]
	MIMIC-IV in-dist.	0.831	**0.839**	0.818	+0.008 [−0.001, +0.017]	−0.014 [−0.020, −0.008]*
	MIMIC-IV held-out sites	0.765	**0.782**	0.745	+0.017 [+0.006, +0.027]*	−0.021 [−0.029, −0.013]*
	eICU→MIMIC-IV	0.806	0.806	**0.808**	−0.001 [−0.009, +0.007]	+0.001 [−0.002, +0.005]
	MIMIC-IV→eICU	0.763	**0.779**	0.766	+0.015 [+0.007, +0.024]*	+0.003 [−0.006, +0.011]
	HiRID→eICU	0.719	**0.763**	0.751	+0.043 [+0.034, +0.053]*	+0.031 [+0.027, +0.035]*
	HiRID→MIMIC-IV	0.785	**0.822**	0.802	+0.036 [+0.032, +0.041]*	+0.017 [+0.012, +0.021]*

Table 7: **Mortality, prolonged stay (> 3 days) and sepsis, standard-trained encoders, without the 24h restriction.** Mean AUROC per model over ten seeds, and seed-paired differences with 95% t-intervals; *: interval excludes zero. "in-dist.": in distribution; "held-out sites": eICU hospitals or MIMIC-IV care units unseen in training; "A→B": encoder trained on A, probe fitted and scored on B. HiRID has no sepsis label. Interval bounds printed as ±0.000 are below 0.0005 in magnitude. **Bold**: the highest of the three model columns in the row, which is not a significance claim.

task	cell	single-stream	coupled	decoupled	coupled−single-stream	decoupled−single-stream
Mortality	eICU in-dist.	0.784	**0.785**	0.774	+0.001 [−0.005, +0.006]	−0.011 [−0.017, −0.005]*
	eICU held-out sites	0.787	**0.798**	0.784	+0.012 [+0.005, +0.018]*	−0.003 [−0.009, +0.004]
	MIMIC-IV in-dist.	0.809	**0.816**	0.801	+0.006 [+0.002, +0.011]*	−0.008 [−0.012, −0.004]*
	MIMIC-IV held-out sites	**0.675**	0.667	0.671	−0.008 [−0.016, −0.001]*	−0.004 [−0.011, +0.004]
	HiRID in-dist.	0.856	**0.884**	0.865	+0.029 [+0.025, +0.032]*	+0.009 [+0.007, +0.012]*
	eICU→MIMIC-IV	0.779	**0.786**	0.783	+0.007 [−0.001, +0.014]	+0.004 [−0.000, +0.008]
	MIMIC-IV→eICU	0.783	**0.786**	0.779	+0.002 [−0.005, +0.010]	−0.004 [−0.010, +0.003]
	eICU→HiRID	0.822	**0.847**	0.835	+0.025 [+0.021, +0.029]*	+0.013 [+0.009, +0.017]*
	MIMIC-IV→HiRID	0.812	**0.841**	0.828	+0.029 [+0.026, +0.033]*	+0.016 [+0.011, +0.021]*
	HiRID→eICU	0.754	**0.791**	0.772	+0.037 [+0.029, +0.045]*	+0.019 [+0.014, +0.024]*
	HiRID→MIMIC-IV	0.769	**0.796**	0.787	+0.026 [+0.022, +0.030]*	+0.018 [+0.013, +0.022]*
Prolonged stay	eICU in-dist.	**0.814**	0.799	0.784	−0.015 [−0.034, +0.004]	−0.030 [−0.039, −0.022]*
	eICU held-out sites	**0.821**	0.805	0.787	−0.015 [−0.031, +0.001]	−0.033 [−0.043, −0.023]*
	MIMIC-IV in-dist.	0.859	**0.885**	0.854	+0.026 [+0.016, +0.036]*	−0.005 [−0.015, +0.006]
	MIMIC-IV held-out sites	0.887	**0.909**	0.883	+0.022 [+0.013, +0.031]*	−0.004 [−0.013, +0.006]
	HiRID in-dist.	0.749	**0.800**	0.767	+0.051 [+0.049, +0.053]*	+0.017 [+0.013, +0.021]*
	eICU→MIMIC-IV	0.861	0.855	**0.862**	−0.007 [−0.018, +0.004]	+0.001 [−0.009, +0.011]
	MIMIC-IV→eICU	0.798	**0.823**	0.786	+0.025 [+0.005, +0.045]*	−0.011 [−0.027, +0.004]
	eICU→HiRID	0.721	**0.758**	0.739	+0.036 [+0.031, +0.041]*	+0.018 [+0.013, +0.023]*
	MIMIC-IV→HiRID	0.711	**0.755**	0.736	+0.044 [+0.036, +0.052]*	+0.025 [+0.020, +0.031]*
	HiRID→eICU	0.801	**0.857**	0.844	+0.056 [+0.031, +0.081]*	+0.043 [+0.017, +0.070]*
	HiRID→MIMIC-IV	0.851	**0.893**	0.883	+0.042 [+0.022, +0.063]*	+0.033 [+0.013, +0.053]*
Sepsis	eICU in-dist.	0.732	**0.737**	0.733	+0.005 [−0.003, +0.013]	+0.001 [−0.005, +0.006]
	eICU held-out sites	0.741	**0.742**	0.726	+0.001 [−0.006, +0.008]	−0.015 [−0.021, −0.009]*
	MIMIC-IV in-dist.	0.816	**0.821**	0.797	+0.005 [−0.004, +0.015]	−0.019 [−0.027, −0.011]*
	MIMIC-IV held-out sites	0.751	**0.770**	0.730	+0.019 [+0.004, +0.034]*	−0.021 [−0.028, −0.014]*
	eICU→MIMIC-IV	0.779	**0.780**	**0.780**	+0.001 [−0.008, +0.010]	+0.001 [−0.003, +0.005]
	MIMIC-IV→eICU	0.725	**0.734**	0.724	+0.009 [+0.002, +0.015]*	−0.001 [−0.007, +0.006]
	HiRID→eICU	0.684	**0.722**	0.715	+0.037 [+0.029, +0.046]*	+0.030 [+0.023, +0.037]*
	HiRID→MIMIC-IV	0.764	**0.802**	0.783	+0.038 [+0.031, +0.044]*	+0.019 [+0.013, +0.026]*

Table 8: **Preregistered labels, standard-trained encoders.** Death within 24h; vasopressor start within 12h and within 24h; acute kidney injury (AKI) within 48h by the KDIGO creatinine and urine-output criteria (App. C). "time": the same probe fitted on elapsed time alone, on the evaluation rows. "positives/eligible": positive and eligible evaluation rows. Cells scored on the same rows can show slightly different time-only values for death within 24h: the 25,000 fitting rows are subsampled in a different order, and with one feature the AUROC depends only on the sign of the fitted coefficient, which varies when elapsed time carries almost no signal (≈ 0.50). Rows are grouped by the database the probe is scored on. Mean AUROC per model over ten seeds, and seed-paired differences with 95% t-intervals; *: the interval excludes zero. **Bold**: the highest of the three model columns in the row; the time-only baseline is excluded. "in-dist.": in distribution, unseen stays of the encoder's own database; "held-out sites": eICU hospitals or MIMIC-IV care units unseen in training; "A→B": encoder trained on A, probe fitted and scored on B.

task	cell	positives/eligible	time	single-stream	coupled	decoupled	coupled−single-stream	decoupled−single-stream
Death within 24h	eICU in-dist.	278/10,091	0.499	**0.867**	0.856	0.848	−0.011 [−0.016, −0.006]*	−0.019 [−0.024, −0.014]*
	eICU held-out sites	449/15,988	0.537	0.868	**0.870**	0.861	+0.002 [−0.003, +0.006]	−0.007 [−0.013, −0.001]*
	MIMIC-IV→eICU	278/10,091	0.499	**0.861**	0.859	0.850	−0.002 [−0.009, +0.006]	−0.011 [−0.017, −0.004]*
	HiRID→eICU	278/10,091	0.499	0.839	**0.851**	0.844	+0.013 [+0.006, +0.020]*	+0.005 [−0.005, +0.016]
	MIMIC-IV in-dist.	2,651/83,680	0.536	0.862	**0.878**	0.865	+0.016 [+0.012, +0.019]*	+0.003 [−0.001, +0.007]
	MIMIC-IV held-out sites	381/10,640	0.538	0.731	**0.745**	0.732	+0.014 [+0.000, +0.027]*	+0.001 [−0.012, +0.014]
	eICU→MIMIC-IV	2,651/83,680	0.545	0.853	**0.858**	0.855	+0.005 [+0.000, +0.010]*	+0.003 [−0.002, +0.007]
	HiRID→MIMIC-IV	2,651/83,680	0.545	0.849	**0.864**	0.856	+0.015 [+0.011, +0.020]*	+0.008 [+0.003, +0.012]*
	HiRID in-dist.	880/25,270	0.496	0.883	**0.918**	0.891	+0.035 [+0.030, +0.039]*	+0.008 [+0.000, +0.016]*
	eICU→HiRID	880/25,270	0.500	0.856	**0.881**	0.865	+0.025 [+0.019, +0.032]*	+0.009 [+0.006, +0.013]*
	MIMIC-IV→HiRID	880/25,270	0.500	0.838	**0.874**	0.854	+0.035 [+0.028, +0.043]*	+0.016 [+0.006, +0.026]*
Vasopressor 12h	eICU in-dist.	238/5,008	0.754	**0.790**	0.788	0.787	−0.002 [−0.012, +0.008]	−0.004 [−0.014, +0.006]
	eICU held-out sites	417/8,658	0.742	0.797	**0.822**	0.794	+0.025 [+0.017, +0.032]*	−0.003 [−0.010, +0.005]
	MIMIC-IV→eICU	238/5,008	0.754	0.780	**0.802**	0.776	+0.023 [+0.011, +0.034]*	−0.003 [−0.013, +0.006]
	HiRID→eICU	238/5,008	0.754	0.762	**0.796**	0.780	+0.033 [+0.025, +0.042]*	+0.018 [+0.005, +0.031]*
	MIMIC-IV in-dist.	2,877/38,186	0.790	0.867	**0.894**	0.874	+0.027 [+0.021, +0.033]*	+0.007 [+0.002, +0.012]*
	MIMIC-IV held-out sites	249/6,366	0.676	0.743	**0.802**	0.753	+0.060 [+0.043, +0.076]*	+0.010 [−0.005, +0.025]
	eICU→MIMIC-IV	2,877/38,186	0.790	0.837	0.857	**0.860**	+0.021 [+0.014, +0.027]*	+0.023 [+0.017, +0.029]*
	HiRID→MIMIC-IV	2,877/38,186	0.790	0.837	**0.878**	0.865	+0.041 [+0.034, +0.048]*	+0.027 [+0.019, +0.035]*
	HiRID in-dist.	2,011/16,042	0.737	0.815	**0.841**	0.827	+0.026 [+0.023, +0.028]*	+0.012 [+0.010, +0.014]*
	eICU→HiRID	2,011/16,042	0.737	0.804	**0.829**	0.815	+0.025 [+0.023, +0.027]*	+0.010 [+0.007, +0.014]*
	MIMIC-IV→HiRID	2,011/16,042	0.737	0.793	**0.821**	0.804	+0.029 [+0.024, +0.033]*	+0.012 [+0.005, +0.019]*
Vasopressor 24h	eICU in-dist.	312/5,008	0.725	0.779	**0.781**	0.774	+0.002 [−0.009, +0.014]	−0.005 [−0.018, +0.009]
	eICU held-out sites	551/8,658	0.724	0.772	**0.799**	0.771	+0.027 [+0.019, +0.035]*	−0.001 [−0.011, +0.009]
	MIMIC-IV→eICU	312/5,008	0.725	0.773	**0.787**	0.763	+0.014 [+0.001, +0.026]*	−0.010 [−0.019, −0.000]*
	HiRID→eICU	312/5,008	0.725	0.746	**0.780**	0.761	+0.034 [+0.028, +0.040]*	+0.015 [+0.000, +0.030]*
	MIMIC-IV in-dist.	3,535/38,186	0.760	0.843	**0.867**	0.846	+0.024 [+0.016, +0.031]*	+0.002 [−0.003, +0.008]
	MIMIC-IV held-out sites	383/6,366	0.656	0.736	**0.791**	0.741	+0.055 [+0.038, +0.072]*	+0.005 [−0.012, +0.023]
	eICU→MIMIC-IV	3,535/38,186	0.760	0.808	0.825	**0.831**	+0.017 [+0.011, +0.023]*	+0.022 [+0.017, +0.028]*
	HiRID→MIMIC-IV	3,535/38,186	0.760	0.809	**0.851**	0.834	+0.042 [+0.034, +0.050]*	+0.025 [+0.017, +0.033]*
	HiRID in-dist.	2,414/16,042	0.712	0.805	**0.828**	0.814	+0.022 [+0.019, +0.026]*	+0.009 [+0.007, +0.011]*
	eICU→HiRID	2,414/16,042	0.712	0.790	**0.816**	0.799	+0.026 [+0.023, +0.029]*	+0.009 [+0.005, +0.013]*
	MIMIC-IV→HiRID	2,414/16,042	0.712	0.778	**0.805**	0.789	+0.027 [+0.023, +0.032]*	+0.010 [+0.005, +0.016]*
AKI, creatinine	eICU in-dist.	695/7,103	0.566	**0.731**	0.727	0.721	−0.004 [−0.010, +0.002]	−0.010 [−0.015, −0.005]*
	eICU held-out sites	1,143/11,115	0.602	0.751	**0.753**	0.752	+0.002 [−0.001, +0.006]	+0.001 [−0.004, +0.006]
	MIMIC-IV→eICU	695/7,103	0.566	0.716	**0.724**	0.713	+0.009 [−0.002, +0.020]	−0.003 [−0.011, +0.006]
	HiRID→eICU	695/7,103	0.566	0.709	**0.733**	0.722	+0.024 [+0.020, +0.028]*	+0.013 [+0.006, +0.020]*
	MIMIC-IV in-dist.	7,510/42,193	0.566	0.728	**0.730**	0.720	+0.002 [−0.002, +0.007]	−0.008 [−0.011, −0.005]*
	MIMIC-IV held-out sites	752/5,954	0.511	0.666	**0.679**	0.672	+0.013 [+0.008, +0.018]*	+0.006 [+0.000, +0.013]*
	eICU→MIMIC-IV	7,510/42,193	0.566	0.707	0.706	**0.713**	−0.001 [−0.006, +0.004]	+0.006 [+0.003, +0.010]*
	HiRID→MIMIC-IV	7,510/42,193	0.566	0.693	**0.718**	0.709	+0.025 [+0.021, +0.029]*	+0.016 [+0.012, +0.020]*
	HiRID in-dist.	2,053/23,283	0.522	0.807	**0.823**	0.811	+0.016 [+0.012, +0.019]*	+0.004 [+0.002, +0.007]*
	eICU→HiRID	2,053/23,283	0.522	0.784	**0.794**	0.793	+0.010 [+0.003, +0.016]*	+0.009 [+0.003, +0.015]*
	MIMIC-IV→HiRID	2,053/23,283	0.522	0.775	**0.791**	0.789	+0.016 [+0.011, +0.020]*	+0.014 [+0.011, +0.017]*
AKI, urine output	eICU in-dist.	1,319/3,279	0.683	0.692	**0.695**	0.689	+0.003 [−0.005, +0.012]	−0.002 [−0.009, +0.005]
	eICU held-out sites	2,580/4,750	0.665	0.679	**0.681**	0.668	+0.002 [−0.007, +0.010]	−0.012 [−0.018, −0.005]*
	MIMIC-IV→eICU	1,319/3,279	0.683	0.689	**0.703**	0.686	+0.014 [+0.005, +0.022]*	−0.003 [−0.015, +0.008]
	HiRID→eICU	1,319/3,279	0.683	0.676	**0.699**	0.696	+0.023 [+0.017, +0.029]*	+0.020 [+0.010, +0.029]*
	MIMIC-IV in-dist.	14,670/29,019	0.663	0.690	**0.701**	0.689	+0.010 [+0.006, +0.014]*	−0.001 [−0.005, +0.003]
	MIMIC-IV held-out sites	1,399/3,905	0.716	0.706	**0.719**	0.691	+0.014 [+0.006, +0.021]*	−0.015 [−0.025, −0.005]*
	eICU→MIMIC-IV	14,670/29,019	0.663	0.684	0.684	**0.688**	−0.000 [−0.005, +0.005]	+0.004 [+0.000, +0.007]*
	HiRID→MIMIC-IV	14,670/29,019	0.663	0.680	**0.701**	0.695	+0.020 [+0.014, +0.026]*	+0.015 [+0.010, +0.020]*
	HiRID in-dist.	7,228/19,139	0.551	0.669	**0.679**	0.673	+0.010 [+0.007, +0.012]*	+0.003 [+0.001, +0.006]*
	eICU→HiRID	7,228/19,139	0.551	0.660	**0.671**	0.663	+0.011 [+0.009, +0.014]*	+0.003 [+0.001, +0.006]*
	MIMIC-IV→HiRID	7,228/19,139	0.551	0.651	**0.663**	0.655	+0.011 [+0.008, +0.015]*	+0.004 [−0.001, +0.009]

Table 9: **How much of the prolonged-stay label elapsed time alone explains.** The logistic probe of §3 fitted directly on the care vector's features, on the rows and splits used everywhere else. "rows after day 3": the share of rows drawn more than three days into the stay, for which the label is positive by construction. Columns 4–6 are in-distribution AUROC for prolonged stay. HiRID rows stop at 24h, so it has no rows after day three. The probe reads the care vector only, not the care-event sequence, so it measures the shortcut carried by the elapsed-time feature itself.

database	rows after day 3	P(prolonged | $t \geq 3$d)	elapsed time alone	care-event features	features, elapsed time zeroed
eICU	27.2%	1.000	0.830	0.850	0.592
MIMIC-IV	41.2%	1.000	0.864	0.880	0.632
HiRID	0.0%	n/a	0.555	0.631	0.613

The shortcut affects prolonged stay: under the logistic probe, elapsed time alone reaches 0.580 on eICU mortality, and zeroing it costs 0.027 on mortality against 0.258 on prolonged stay. On sepsis, elapsed time alone reaches 0.574 under the logistic probe. The eICU analysis below is for prolonged stay; the roll-ups of §4 and Table 6 restrict all three standard tasks to the first 24h on every database.

The margin with the shortcut removed. We restrict both the fitting and the evaluation rows; restricting evaluation alone would leave the shortcut in the fitted coefficients. Comparing Table 7 with Table 6 isolates what the shortcut contributes. HiRID is the internal control: its rows already end at 24h, and its in-distribution prolonged-stay margin is the same either way (+0.051 unrestricted, +0.050 restricted). MIMIC-IV in distribution is also unchanged (+0.026 both ways). On eICU, where 27.2% of rows fall after day three, the unrestricted margin is −0.015 (interval including zero) and the restricted one +0.008 [+0.001, +0.015]: with the shortcut present the single-stream model reads the clock as well as the prolonged-stay label allows, and the comparison measures that rather than the representations. The main text therefore quotes the first-24h margins throughout.

F Hidden-target training

Every model is also trained a second time, as separate runs. *Standard training* puts all 46 analytes in the input and the reconstruction loss. *Hidden-target training* splits the 46 analytes into two halves of 23 with similar coverage, the fraction of rows in which an analyte is measured, computed within each database. The values of one half, the *held-out analytes*, are removed from the input and the loss during training and when representations are computed. The other half, the *reconstructed analytes*, stays in input and loss. Whether a held-out analyte was measured still enters every model through its presence indicator on the laboratory-value token; its presence column in the care vector is set to zero.

F.1 With the scored values removed, both two-encoder models recover more

Hidden-target training retrains every model with the 23 held-out analytes absent from input and loss, and keeps them absent when representations are computed, so no model can copy a value it is scored on. Recovery levels are lower than under the visible-target probe (recovery models: single-stream 0.214/0.412/0.336, coupled 0.303/0.475/0.467 on eICU/MIMIC-IV/HiRID). The recovery gap is positive on every database and every seed: +0.089 [+0.082, +0.096] on eICU, +0.062 [+0.057, +0.068] on MIMIC-IV and +0.132 [+0.122, +0.141] on HiRID (10 of 10 seeds each), and +0.090 [+0.080, +0.100] and +0.093 [+0.086, +0.100] at held-out eICU hospitals and MIMIC-IV units. Decoupled minus single-stream is +0.093, +0.065 and +0.130: the decoupled model recovers as much as the coupled model, so the recovery gap does not need the contrastive loss between the encoders (§6). The gap is larger here than under the visible-target probe on every database, so the probe is not rewarding a copy of the scored value.

The held-out analytes differ by database because the coverage split is computed within each database (full lists in App. A). They include hematocrit (Hct), mean corpuscular volume (MCV) and red blood cell count (RBC) on eICU; hemoglobin (Hgb), mean corpuscular hemoglobin concentration (MCHC) and RBC on MIMIC-IV; and Hgb, mean corpuscular hemoglobin (MCH), base excess, HCO_3, $paCO_2$ and paO_2 on HiRID.