

A generative digital twin forecasts and simulates intensive care patients without task labels

Han Wang

Affiliated Hospital of Guangdong Medical University

Kaisheng Ye

The First Affiliated Hospital of Jinan University

Huaguo Yuan

Affiliated Hospital of Guangdong Medical University

Weibang Chen

Affiliated Hospital of Guangdong Medical University

Yao Cai

Affiliated Hospital of Guangdong Medical University

Ye Tian

Affiliated Hospital of Guangdong Medical University

Ting Liu

Affiliated Hospital of Guangdong Medical University

Yue Xu

Affiliated Hospital of Guangdong Medical University

Xiaoyu Liu

Affiliated Hospital of Guangdong Medical University

Lingpin Pang

Affiliated Hospital of Guangdong Medical University

Huafeng Liu

Guangdong Provincial Key Laboratory of Autophagy and Major Chronic Non-communicable Diseases

Yue Ma

East China Normal University

Zizhong Yu

Shanghai Xieai Ze'an Traditional Chinese Medicine Hospital

Xiaozheng Zou

The Fourth Affiliated Hospital of China Medical University

Junfeng Hao

Affiliated Hospital of Guangdong Medical University

Chunjie Tian

tianchunjie@gdmu.edu.cn

Article

Keywords:

Posted Date: October 16th, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-11382527/v1>

License: © ⓘ This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Additional Declarations: No competing interests reported.

A generative digital twin forecasts and simulates intensive care patients without task labels

Han Wang¹, Kaisheng Ye², Huaguo Yuan¹, Weibang Chen¹, Yao Cai¹, Ye Tian¹, Ting Liu¹, Yue Xu¹, Xiaoyu Liu¹, Lingpin Pang³, Huafeng Liu⁴, Yue Ma⁵, Zizhong Yu⁶, Xiaozheng Zou^{7,*}, Junfeng Hao^{8,9,*}, and Chunjie Tian^{1,8,10,*}

¹Department of Otorhinolaryngology, Affiliated Hospital of Guangdong Medical University, Zhanjiang 524000, Guangdong, China.

²Department of Anesthesiology, The First Affiliated Hospital of Jinan University, Guangzhou, Guangdong, China.

³Emergency Medical Center, Affiliated Hospital of Guangdong Medical University, Zhanjiang, Guangdong, China.

⁴Guangdong Provincial Key Laboratory of Autophagy and Major Chronic Non-communicable Diseases, Zhanjiang 524001, Guangdong, China.

⁵School of International Chinese Studies, East China Normal University, Shanghai 200062, China.

⁶Shanghai Xieai Ze'an Traditional Chinese Medicine Hospital, Shanghai, China.

⁷Department of Critical Care Medicine, The Fourth Affiliated Hospital of China Medical University, Shenyang 110032, Liaoning, China.

⁸Guangdong Provincial Key Laboratory of Autophagy and Major Chronic Non-communicable Diseases, Key Laboratory of Prevention and Management of Chronic Kidney Disease of Zhanjiang City, Institute of Nephrology, Affiliated Hospital of Guangdong Medical University, Zhanjiang 524001, Guangdong, China.

⁹Department of Family Medicine, Shengjing Hospital of China Medical University, Shenyang 110022, Liaoning, China.

¹⁰Otolaryngology Research Center, Affiliated Hospital of Guangdong Medical University, Zhanjiang 524000, Guangdong, China.

*Correspondence and requests for materials should be addressed to:

Xiaozheng Zou (zouxz_cmu@163.com);

Junfeng Hao (ygzhjf85@gmail.com); or

Chunjie Tian (tianchunjie@gdmu.edu.cn).

Abstract

Generative models that simulate a patient's futures could serve as digital twins for prognosis, trial design and teaching, but in intensive care they have been trained within single databases and rarely tested against observed treatment responses. We trained Didymos, a 123-million-parameter transformer, without task labels on 409,297 stays from seven databases in five countries. On held-out stays, Didymos forecast 13 primary endpoints with AUROCs of 0.83–0.98 from its simulated futures and ranked first among 20 predictors, including supervised LightGBM and TabPFN-3.5 (post hoc). Its step hazards of death were calibrated, whereas free generation postponed deaths. Simulated treatment responses agreed with eight of nine observational contrasts, membership inference stayed near chance ($AUC \leq 0.57$) and raters did not tell synthetic from real patients. Didymos risks cut simulated trial sizes by 14–37%. At two unseen hospitals, minutes of training on unlabelled local stays brought Didymos level with locally supervised models.

1 Introduction

2 Intensive care is anticipatory. Within hours, a clinician must judge whether a patient is likely to deteriorate, to
3 need invasive ventilation, vasopressors or kidney replacement, or to die, and how a treatment started now
4 might change that course. A patient digital twin, which we take here to mean a model that is conditioned on a
5 patient's record and can simulate that patient's possible futures, would let one model address many of these
6 questions, from prognosis to the design of trials and the teaching of clinicians.¹ Mechanistic twins of critically
7 ill patients encode physiology and treatment effects by hand. They have been verified prospectively on small
8 cohorts, but they cover few organ systems and treatments.²

9 Most prediction in intensive care rests instead on supervised models, each trained for one outcome with
10 labels defined at one site. Such models discriminate well where they were developed, but their area under the
11 receiver operating characteristic curve (AUROC) sometimes falls by 0.2 at new hospitals, and training on
12 several databases narrows the gap without closing it.³ Every new outcome or change of definition requires
13 new labels and a new model. Clinical scores such as the Sequential Organ Failure Assessment (SOFA), the
14 Simplified Acute Physiology Score (SAPS) II and the National Early Warning Score (NEWS) transfer across
15 sites and are widely understood, but they were designed for a few outcomes and predict others coarsely.⁴⁻⁶

16 Generative models of health records take a different route. Trained to continue sequences of medical events,
17 they learn the joint distribution of a patient's record and can answer a prediction question by sampling futures
18 and counting how often an event occurs. ETHOS made zero-shot predictions from tokenized timelines in
19 MIMIC-IV.⁷ Foresight forecast disorders from concepts extracted from free text in UK hospitals,⁸ Delphi-2M
20 modelled the incidence of more than 1,000 diseases in UK Biobank and transferred to Danish registries,⁹ and
21 CoMET models trained on 118 million US patients matched or exceeded task-specific supervised models
22 on 78 tasks.¹⁰ Large language models fine-tuned on patient trajectories forecast laboratory values, including
23 in intensive care,¹¹ and recent twins of intensive care unit (ICU) patients estimate risk continuously or plan
24 sepsis treatment within one health system.^{12,13} Yet the models among them that use intensive care data were
25 each trained within one database or health system and judged mainly by the accuracy of their predictions.
26 More broadly, a review of 84 foundation models of clinical records found that most were trained on narrow
27 datasets and evaluated on tasks that say little about their usefulness to health systems.¹⁴

28 Four questions that decide whether such a model can serve as a twin of intensive care patients therefore
29 remain largely open. First, can one model learn intensive care across databases whose recording schemas,
30 practices and populations differ, and carry what it learned to hospitals it has not seen? Second, are its forecasts
31 calibrated, and when they are not, why? Third, do its simulated treatment responses agree with responses
32 observed in data, beyond face validity? Fourth, can it be used safely and transparently, keeping the patients in
33 its training data private even when sampling is conditioned on real records, as every twin application requires,
34 and offering explanations that reflect what it learned?¹⁵⁻¹⁷

35 Here we present Didymos (Greek for "twin"), a generative model of intensive care trained once, without
36 task labels, on 409,297 stays from seven databases in the USA, Switzerland, the Netherlands, Austria and
37 South Korea. We evaluated it on held-out stays under a protocol fixed before any evaluation score was read,
38 with later analyses added as dated amendments. The comparators were 13 learned models trained with
39 labels, including TabPFN-3.5 and LightGBM, seven clinical scores, the databases' own severity scores and
40 single bedside values, and each analysis had a control. Didymos forecast outcomes level with or ahead of
41 the supervised models, helped by a hazard estimate that removes most of the Monte Carlo noise of sampling.
42 Separating its step hazards from free generation showed that the hazards of death were calibrated and that free
43 generation postponed deaths. Its simulated treatment responses agreed with propensity-matched observational

44 contrasts, and a null arm stayed at chance. A membership-inference audit of unconditional, prompted and
45 intervened sampling found a signal near chance that aligned with model fit. Attention rollout mirrored the
46 causal mask, whereas counterfactual editing and attention knockout reflected the trained model. Used as
47 prognostic covariates through the Didymos toolkit, Didymos risks reduced the sample sizes of simulated
48 trials by 14–37%. Finally, at two hospitals in Russia and China that contributed nothing to training, Didymos
49 transferred at the level of supervised models trained elsewhere, and brief training on each hospital’s unlabelled
50 stays brought it level with models trained on the hospital’s own labels.

51 **Results**

52 **Didymos models intensive care stays from seven international databases**

53 We harmonized 409,297 adult ICU stays from seven credentialed databases (MIMIC-IV, eICU-CRD, HiRID,
54 AmsterdamUMCdb, SICdb, NWICU and INSPIRE) in the USA, Switzerland, the Netherlands, Austria
55 and South Korea (Table 1).^{18–31} The databases differ in case mix and recording practice, with several-fold
56 differences in the rates of invasive ventilation and ICU death. Each stay, from 48 hours before ICU admission
57 to ICU discharge, was converted through the Common Longitudinal ICU data Format (CLIF)³² into one
58 token sequence of context, timed events with value deciles and discharge outcome (Supplementary Fig. 1). A
59 grammar fixes the order of these blocks and accepts all 409,297 real stays (Fig. 1a,c; Supplementary Note 1).

60 Didymos is a decoder-only transformer with 123.5 million parameters and a context of 32,768 tokens, trained
61 once by next-token prediction, without task labels, on the train split of all seven databases (Supplementary
62 Fig. 2; Supplementary Note 2). Conditioned on a stay’s record up to a prediction time, it samples simulated
63 futures under the grammar and the database’s support, the set of tokens that database records (Fig. 1c). A
64 risk is read from the futures in two ways. The sampled estimate is the share of futures in which the event
65 occurs. The hazard estimate sums the model’s step probabilities of the event along each future and averages
66 over futures, which gives the same expectation with lower variance.³³ Predictions were scored on 2,000
67 held-out stays per database, at rolling anchors (prediction times drawn along stays) and at the 24- and 48-hour
68 landmarks (Fig. 1b; Supplementary Fig. 3). Results are labelled pre-specified, follow-up or post hoc where the
69 distinction matters (Methods).

70 **Didymos forecasts intensive care outcomes without task-specific labels**

71 In the follow-up analysis, Didymos drew 128 simulated futures per anchor on 1,050 rolling anchors and on a
72 case-cohort sample of 700 stays at the 24-hour landmark.³⁴ Across 13 primary endpoints, its AUROC ranged
73 from 0.831 for hyperkalaemia within 24 hours to 0.982 for death within 24 hours of the landmark (Fig. 2a;
74 Supplementary Fig. 4; Supplementary Note 3). Its AUROC exceeded that of LightGBM³⁵ on 12 of the 13
75 endpoints and that of TabPFN-3.5³⁶ on 9 of 12, both trained with labels on the same features, and every
76 difference in a comparator’s favour had a 95% CI that included zero (Fig. 2b; Supplementary Table 1). After
77 Holm adjustment,³⁷ Didymos was ahead of LightGBM for new invasive mechanical ventilation (IMV) within
78 24 hours (+0.081; 95% CI 0.029 to 0.152; $P = 0.027$), acute kidney injury (AKI) of stage 2 or more within
79 48 hours (+0.053; $P = 0.0078$) and death within 24 hours of the landmark (+0.025; $P = 0.013$). Against
80 TabPFN-3.5, its advantage for a new vasopressor held after adjustment (+0.061; $P = 0.030$).

81 In a post hoc ranking of 20 predictors on the 12 primary endpoints on which all were fitted, Didymos had
82 the best mean rank (1.83; 95% interval 1.08 to 2.92) and the highest AUROC on eight (Table 2; Supplementary
83 Fig. 5). TabPFN-3.5 ranked first for ICU death, and the Didymos probe, a supervised readout of the same

model's state, had a mean rank of 7.50. Even with its sampled estimate, Didymos exceeded the best clinical score for each endpoint, chosen after the fact, by 0.040 to 0.247 AUROC (Holm $P \leq 0.0026$ on ten of 13 and 0.068 to 0.080 on the other three). The databases' own admission scores also fell below Didymos on their own data: the Acute Physiology and Chronic Health Evaluation (APACHE) IV³⁸ for in-hospital death at the landmark in eICU-CRD (0.850 against 0.941) and SAPS 3³⁹ in SICdb (0.812 against 0.869). A single bedside value approached Didymos only where the endpoint is a threshold on that value, as for platelets and thrombocytopenia.

The hazard estimate removed most of the noise of sampling. From a single simulated future, it gave AUROCs of 0.81–0.92 on the six endpoints where it applies, against 0.61–0.66 for the sampled estimate. For death within 24 hours on rolling anchors, eight hazard-read futures exceeded 128 sampled ones (0.937 against 0.918; Fig. 2c; Supplementary Fig. 6). With 32 sampled futures, the pre-specified analysis had trailed LightGBM on three mortality-related endpoints, most for death within 24 hours on rolling anchors (–0.062), while already leading for a new vasopressor and AKI. In the follow-up, Didymos was level with or ahead of LightGBM on all three endpoints and exceeded the pre-specified estimate on four endpoints after Holm adjustment (Fig. 2e). Fig. 2d and Supplementary Fig. 7 compare the follow-up's sampled estimate with LightGBM within each database. The 48-hour landmark, re-run with the follow-up's settings on 809 anchors, showed the same pattern. Didymos was level with LightGBM for ICU and in-hospital death and exceeded its pre-specified estimate on all four endpoints, most for an ICU stay of more than 7 days (+0.080; 95% CI 0.040 to 0.127; Holm $P < 0.001$; Supplementary Fig. 8).

Beyond the primary endpoints, an exploratory analysis of all 165 endpoints that the pre-specified analysis scored found Didymos ahead of LightGBM on 113 (68%). At a 5% false-discovery rate,⁴⁰ 49 comparisons favoured Didymos and 6 favoured LightGBM (Supplementary Fig. 9; Supplementary Data 1; Supplementary Note 4). Didymos trailed mainly on mortality and the remaining stay (Supplementary Fig. 10a), the families whose primary endpoints the follow-up's longer simulations brought to parity. In the pre-specified label-efficiency analysis, a LightGBM trained within one database needed more than 1,000 labelled stays to match Didymos in 50 of 78 endpoint–database cells, and in 38 it did not match Didymos even with all sampled training stays (Supplementary Fig. 10b). The pre-specified analysis also predicted the remaining ICU stay at the 24-hour landmark with a mean absolute error of 41.7 hours, against 58.6 hours for predicting the mean, and the time to death with a Harrell's concordance index of 0.85 (Supplementary Fig. 11).

The database token carried each database's recording schema. With the database and site set to unknown, AUROC changed by less than 0.04 on six of nine rolling endpoints, whereas another database's token lowered it, by up to 0.30 for new renal replacement therapy (RRT; Supplementary Fig. 12a; Supplementary Note 5). For new IMV, a new vasopressor and AKI, a LightGBM trained on the other six databases lost 0.04–0.11 AUROC against one trained within the database. Didymos, which had seen that database's stays without labels, exceeded the transferred LightGBM by 0.09–0.12 on average (Supplementary Fig. 12b). In pre-specified subgroups, discrimination was similar across sex, age, admission type and recorded race, whereas calibration varied more, for example with an observed-to-expected ratio (O/E) of 2.25 for in-hospital death in AmsterdamUMCdb (Supplementary Fig. 13).

Free generation delays deaths although the step hazards are calibrated

The pre-specified analysis under-predicted death, so we separated the model's step probabilities from the futures it generates (Fig. 3; Supplementary Note 6). Summed along the anchors' real futures, the step hazards of death were calibrated for death within 24 hours on rolling anchors (O/E 0.98; 95% CI 0.86 to 1.11) and for

ICU death at the landmark (0.95), but over-predicted death within 24 hours of the landmark (0.76). Simulated futures carried 0.61 of the real death hazard within 24 hours on rolling anchors and 0.57 at the landmark, but 0.91 of it until ICU discharge at the landmark. In the first 6 hours, the simulated death hazard was about half the real one, whereas the hazard of alive discharge stayed close to its real value (Fig. 3a). Whole synthetic stays lacked deaths mainly among short stays: in MIMIC-IV, ICU death among stays shorter than 24 hours was 3.9% against 11.5% in real stays (Fig. 3e). Didymos therefore postpones deaths in free generation rather than omitting them. At the landmark, the pre-specified token cap had also censored long simulations. With four times the cap, ICU death was calibrated (O/E 0.99 for the sampled and 1.01 for the hazard estimate) and AUROC was kept (Fig. 3d).

Treatment starts were over-predicted step by step along real futures (O/E 0.56–0.80), and mostly more so in generation. With 32 futures on rolling anchors, the raw O/E was 0.46–0.55 for new IMV, a new vasopressor and new RRT (Holm $P \leq 0.013$ against O/E = 1; Fig. 3b). Logistic maps fitted on the tuning split and adopted post hoc brought a new vasopressor to 1.02 and new RRT to 0.93 without changing AUROC. For new IMV they overshot (1.43), because its raw O/E on the tuning split had been 0.40, against 0.55 on the held-out anchors (Supplementary Fig. 14; Supplementary Note 7). Isotonic maps for four other endpoints overcorrected death within 24 hours (from 1.71 to 0.75) and hyperkalaemia (Fig. 3b,c). Before correction, O/E differed from 1 for five of the seven endpoints with maps (Holm $P \leq 0.013$), and after correction for two (Holm $P = 0.042$ each).

Simulated treatment responses reproduce observational contrasts

To test the simulator against data, we cut the records of patients who started a treatment 1 minute before the recorded start. Each was matched, on the propensity of starting, to a patient from the same database at the same ICU hour who did not start it within the next 6 hours (158–195 pairs per treatment; Fig. 4a; Supplementary Note 8).⁴¹ For eight of nine pre-specified responses, Didymos’s contrast between the pairs’ simulated factual futures lay within the 95% CI of the observed contrast (Fig. 4b), as for mean arterial pressure (MAP) after norepinephrine (−3.3 against −3.6 mmHg). Insulin was the exception. Real starters’ glucose stayed 38.0 mg/dL (6.4 to 72.2) above their matches’, against −0.3 mg/dL simulated, although within the simulation giving insulin lowered the starters’ glucose by 22.6 mg/dL against withholding it. No difference between the simulated and observed contrasts, insulin’s included, was significant after Holm adjustment ($P \geq 0.31$).

In a null arm, two simulations of the same withhold arm with different seeds differed in the expected direction in 0.40–0.55 of prefixes, close to the 0.5 of chance (Holm $P \geq 0.055$; Fig. 4c). Factual futures improved on carrying the last value forward, with a continuous ranked probability score (CRPS)⁴² 0.27–0.60 times that of persistence, although their 80% intervals covered only 0.66–0.84 of observed responses (Fig. 4d). The simulated effect of a treatment depended on where the record was cut. Cut 1 minute before the start, giving norepinephrine raised MAP against withholding it (+0.79 mmHg; 0.11 to 1.47). Cut 0.5–6 hours earlier, as pre-specified, it lowered MAP (−0.81 mmHg), because the fall in pressure that prompted the start dominated (Fig. 4e). The simulated contrasts therefore describe observed practice, including its confounding by indication, and are not causal treatment effects.

Synthetic patients are private, faithful and useful

We audited membership inference, which asks whether an attacker can tell from the model’s behaviour that a stay was in its training data (Fig. 5a; Supplementary Note 9).^{43,44} With a loss-based attack on whole stays, the area under the attack’s receiver operating characteristic curve (membership AUC) was 0.549–0.555 in every

database and the true-positive rate at a 1% false-positive rate 0.010–0.024, close to the 0.5 and 0.01 of an attack without information. Across nine checkpoints, the membership signal aligned with fit. In each database, membership AUC rose with the gap between the model’s loss on members and non-members (Spearman correlation 0.98–1.00; Holm $P < 0.001$; Fig. 5b), and the number of passes over each database was set from this relation. Because digital-twin applications condition on real records, we repeated the audit for prompted sampling, scoring a stay’s next 24 hours given its record up to a random time. Pooled AUC was 0.531 (95% CI 0.524 to 0.537) and the true-positive rate 0.013. INSPIRE, whose held-out split yields only 790 stays per side, had the widest interval (AUC 0.568). No simulated future of a member or a non-member reproduced the first 32 tokens of its real future, with or without an intervention.⁴⁵ The controls behaved as expected: a copying generator reproduced every member’s future, and two halves of the non-members, with no membership to detect, gave an AUC of 0.493.

Synthetic stays were faithful at the level of databases and hospitals. Their per-database profiles matched those of real held-out stays except for fewer deaths, in keeping with the delay described above (Fig. 5c; Supplementary Note 10). The structure between databases was preserved (correlation 0.73, against 0.95 between two real samples), as was the ranking of 17 eICU-CRD hospitals (Spearman correlation 0.45–0.72, against 0.52–0.85 between two real samples; Fig. 5d), and all 25 rare header strata were reached. Models trained on synthetic stays and tested on real ones reached 0.93–1.03 of the gain over chance of models trained on as many real stays (Fig. 5e). At a small site with 100 real training stays, adding as many synthetic stays raised AUROC by 0.038 at the landmark (gains in 19 of 19 cells; Holm $P < 0.001$), a gain that shrank to about 0.01 with 1,000 real stays (Fig. 5f).

Editing and knockout reflect the trained model, unlike attention rollout

Explanations of an individual’s risk are informative only if they depend on what the model learned. We ranked groups of a patient’s record, each a channel family within a time block, by their influence on the Didymos probe’s risk, and compared each method’s ranking between the trained model and a randomly initialized one (Fig. 6a; Supplementary Note 11).¹⁷ Counterfactual editing, which removes a group from the record, and attention knockout, which blocks attention to the group’s tokens in every layer,⁴⁶ reflected the trained model. Knockout’s ranking was unrelated to the random model’s (rank correlation -0.02) and agreed with editing (median 0.71).

Attention rollout⁴⁷ ranked groups almost as it did in the random model (0.96–0.97). In a causal decoder, multiplying attention across layers drains mass towards early tokens whatever the attention does. The rollout of uniform causal attention, with no model at all, ranked groups like the trained rollout (0.97) and exactly like the random one (1.00; Fig. 6b). In both models, 35–38% of the rollout’s mass landed on the first token, which received 0.2% of the trained model’s last-layer attention (Fig. 6c), and summing token scores over a group added a confound by group size. Knockout’s agreement with editing exceeded rollout’s by a median of 0.74 per anchor (152 anchors; Wilcoxon signed-rank test, $P < 0.001$).

Against the generator’s own effects of each edit, read from simulated futures, editing agreed at 0.12–0.20 on three endpoints, within reliability ceilings of 0.48–0.54, and at 0.07 for hypotension, within a ceiling of 0.13. Integrated gradients⁴⁸ and attention rollout did not exceed 0.08 (Fig. 6d). When the generator’s effects were drawn again for all 160 anchors, knockout agreed with them as well as editing did, with a median per-anchor difference of 0.00 (158 anchors; Wilcoxon signed-rank test, $P = 0.61$) and medians of 0.06–0.23 within ceilings of 0.34–0.58 (Supplementary Fig. 15).

Didymos risks reduce the sample sizes of simulated trials

The Didymos toolkit packages these capabilities as seven tools (Forecast, What-if, Explain, Similar patients, Cohort synthesizer, Trial designer and Teaching cases), each output carrying the evidence and caveats of the corresponding analysis (Fig. 7a). With new draws, its AUROC and O/E lay within simulation noise of one of the paper's two runs on all nine rolling primary endpoints, a criterion set after the pre-specified conformance rule had met 14 of its 15 checks (Supplementary Fig. 16; Methods; Supplementary Note 12). Neighbours in Didymos's state predicted outcomes about as well as neighbours by term frequency-inverse document frequency (TF-IDF) of the record or by the comparators' features, and better than TF-IDF on 8 of 12 endpoints, whereas an untrained model's state did clearly worse (Supplementary Fig. 17).

The Trial designer uses the Didymos risk as a prognostic covariate in the analysis of a randomized trial.⁴⁹ In 2,000 simulated trials per endpoint, drawn from landmark stays with a relative risk of 0.8 and 80% power, the adjusted analysis needed 0.63–0.86 of the unadjusted sample size, least for in-hospital death (0.63; Monte Carlo 95% interval 0.59 to 0.68; Fig. 7b). LightGBM needed 0.64–0.93, and SOFA, SAPS II, the Oxford Acute Severity of Illness Score (OASIS) and NEWS 0.80–1.00. On eICU-CRD, adjusting for APACHE IVa needed 0.81 of the unadjusted size for ICU death, against 0.67 with Didymos. Didymos's ratios were smaller than every clinical score's, with separated intervals, and smaller than or equal to LightGBM's on all six endpoints, with overlapping intervals for several (no test against LightGBM was specified). Enrolling the quarter of patients at highest Didymos risk needed 0.18–0.30 of an unenriched trial's size (Fig. 7c). At a relative risk of 1, the adjusted analyses' type I error was 0.0375–0.0715 (nominal 0.05), and it reached 0.079 in small eICU-CRD-only subsets sized for a relative risk of 0.7 (Supplementary Fig. 18). For teaching, the Teaching cases tool produced nine synthetic cases for three objectives (shock, respiratory failure and renal support), each meeting its automated checks, and Fig. 7d follows one shock case through Forecast, What-if and Explain.

Raters did not tell synthetic from real patients

A blinded rater study, fixed in a protocol amendment before any stimulus was built, assessed realism, simulated responses, explanations and decision support. Twenty raters, recruited as intensivists under the protocol, completed its packets, and their answers were analysed with the pre-specified analysis (Methods).

In 1,600 real-or-synthetic calls on first-day case sheets, raters called synthetic stays real in 65% of calls and real stays in 47%. Their accuracy at telling the two apart, 0.41 (90% CI 0.35 to 0.47), was below the 0.5 of chance, with every rater between 0.34 and 0.49 (Fig. 8a). Synthetic 6-hour treatment courses were rated close to real ones in plausibility (–0.23 points on a 1–5 scale; 95% CI –0.56 to 0.11; Fig. 8b). Factors chosen by editing, by knockout and at random were all rated relevant (mean 3.6 of 5 each), so relevance ratings did not separate the methods (Fig. 8c). The forecast card improved raters' risk estimates, with Brier score differences (with minus without the card) of –0.016 for ICU death (–0.031 to –0.001; $P = 0.037$) and –0.028 for AKI ($P = 0.0064$; Fig. 8d). The card alone scored 0.121 for ICU death and 0.126 for AKI, at least as well as the aided raters. Raters moved their estimates about a third of the way towards the card (weight of advice 0.36), including most estimates for which it pointed the wrong way. The gain specific to Didymos, against the same shift towards LightGBM's risks, was –0.016 (–0.038 to 0.001), estimated as pre-specified rather than tested.

Didymos transfers to unseen hospitals and adapts to them without labels

Two databases that contributed nothing to training served to test transfer to new hospitals: RICD, from a centre for intensive care and neurorehabilitation in Moscow, Russia,⁵⁰ and zigongICU, from ICU patients with

infection in Zigong, China.^{51,52} New adapters mapped both to the common format and the frozen vocabulary, and prompts carried the database and site as unknown. A gate fixed before any future was drawn kept the 15 endpoint–database pairs whose source was timed to the hour and that had at least 20 positive anchors (Methods). Without further training, Didymos reached AUROCs of 0.54–0.93 (median 0.72) on 2,398 anchors, highest for death within 24 hours in RICD (0.934; Supplementary Fig. 19a; Supplementary Note 13). It exceeded LightGBM trained on the seven databases on 10 of 15 pairs and, among the predictors that never saw the new databases, ranked second after TabPFN-3.5 (mean rank 3.80 against 3.27; Supplementary Fig. 19d). A LightGBM trained on each database’s own labelled stays was ahead on 13 of 15 pairs, and recalibration within each database brought O/E from 0.04–3.46 to 0.93–1.01 (Supplementary Fig. 19c).

We then adapted Didymos to each database without labels, under a protocol amendment fixed before any adapted model was trained. Each database’s patients were split once: half into a locked test set, 35% into unlabelled stays for continued training and 15% into a labelled slice (Fig. 9a). Mapping codes that the vocabulary lacked to their nearest kept codes removed most unknown events, for example 85.1% of RICD’s continuous infusions, but changed AUROC only by –0.04 to +0.03 (Supplementary Fig. 20a,b; Supplementary Note 14). Continuing Didymos’s next-token training on the unlabelled stays, mixed one to one with training windows from the seven databases, took 7.5 GPU-minutes. It lowered the next-token loss on held-out patients of the new databases from 2.10 to 1.13 nats per token, whereas the loss on the seven databases rose by 0.01 to 0.03 (Fig. 9b).

On the locked test anchors, adapted Didymos exceeded the zero-shot model on 8 of the 13 pairs with at least 10 events (Fig. 9c,d). The largest gains were for outcomes shaped by local practice, such as an ICU stay of more than 7 days in RICD (+0.119; 95% CI 0.014 to 0.214) and a new vasopressor in zigongICU (+0.125; 0.005 to 0.233). No single difference held after Holm adjustment (lowest $P = 0.21$). Recalibrated on the labelled slice, adapted Didymos was level with a local LightGBM trained on the labels of half the patients, being ahead on 6 of 13 pairs. It was ahead of TabPFN-3.5 on 8 of 13, most for an ICU stay of more than 7 days in RICD (+0.167; Holm $P = 0.021$). Stacking adapted Didymos with a LightGBM trained on the labelled slice gave the best mean rank among 15 predictors (3.15; 95% interval 2.15 to 4.23; Fig. 9e). A LightGBM trained on up to 877 labelled stays reached adapted Didymos on only 3 of 15 pairs (Fig. 9f).

Adaptation also improved calibration at the 24-hour landmark, where O/E was 0.71–1.39 against 0.32–3.60 before adaptation, whereas rare events within 24–48 hours remained under-predicted after recalibration on the labelled slice (Supplementary Fig. 21). Continued training gave the adapted stays a membership signal (AUC 0.564; 95% CI 0.548 to 0.579, against 0.502 before adaptation), as training does for any stay it uses (Supplementary Fig. 20d). On the seven training databases, the adapted model’s membership signal matched the original model’s, with a pooled true-positive rate at a 1% false-positive rate of 0.015 for both (Supplementary Fig. 22; Methods).

Discussion

Didymos is a single generative model trained without task labels on seven intensive care databases from five countries. It forecast outcomes level with or ahead of supervised models, simulated treatment responses that agreed with observed contrasts and generated synthetic patients whose membership signal stayed near chance, so one model answered many prognostic questions without per-task labels. At two hospitals it had never seen, a few minutes of training on their unlabelled stays brought it level with local supervised models. Where it departed from the data, in the timing of simulated deaths, the frequency of treatment starts and the explanations

of attention rollout, the analyses identified why, which tells a user which estimate to trust for which question.

These results corroborate, across recording schemas and countries, the finding that generative event models can match task-specific supervised models without fine-tuning, which CoMET showed with 118 million patients from 310 US health systems¹⁰ and ETHOS within one hospital database.⁷ Earlier generative models of intensive care data were each trained within one database or health system (Supplementary Table 2).^{7,11,13} The advantage of Didymos was concentrated in treatment starts, kidney injury and death within 24 hours of the landmark, and for ICU and in-hospital death it was level with the supervised models. The headline configuration, 128 futures read through the hazard estimate, came from a follow-up specified after the pre-specified results were known. That follow-up compared two estimators and more futures, fixed in writing before they ran, and the pre-specified results are reported beside it. The supervised comparators were trained on up to 5,000 labelled stays per database, with features summarizing the 64 most charted channels. More labels or richer features might narrow some gaps, although within one database LightGBM often did not reach Didymos even with all sampled training stays.

The Didymos probe, a supervised readout of the same model's state, ranked far below the simulated forecasts (mean rank 7.50 against 1.83). Rolling the model forward therefore recovered information about the future that a linear readout of its current state did not, which suggests that the advantage comes from simulation and not only from the representation. The hazard estimate, in turn, showed that the sampled estimate's early deficit was mostly Monte Carlo noise. Reading the model's step probabilities along each future removes the variance of counting rare events, applies to any autoregressive event model that exposes its next-event probabilities and reduces the number of futures that simulation-based prediction needs. A further, untested explanation for the advantage is that the event stream holds precursors of treatment starts, such as orders, fluid boluses and changes in charting, that summary features over fixed windows capture only coarsely.

Separating the step hazards from free generation located the error in simulated mortality, and we propose this decomposition as a general diagnostic for generative clinical models. The delay of simulated deaths is consistent with exposure bias, the drift of autoregressive models conditioned on their own samples,⁵³ acting on a terminal event, where small departures that keep a simulated patient alive compound over steps. We did not test this explanation, and the grammar and support masks could produce a similar delay by removing tokens that precede death in some schemas. Training with the model's own samples would distinguish the two. Either way, risks of death should be read through the hazard estimate or from simulations long enough to reach the outcome. The over-prediction of treatment starts, in contrast, is present step by step, before any drift, which points to under-fitted timing of rare events. Recalibration corrects it only as well as the tuning split resembles the test data, so larger tuning sets or recalibration at each site are needed, and calibration should be checked wherever such risks are used.⁵⁴

A simulator faithful to the data must inherit the data's confounding. The what-if contrasts therefore describe observed practice, and the observational contrasts they match may themselves carry residual confounding after matching. Estimating treatment effects with Didymos would need the design of observational causal inference, such as emulating a target trial with explicit identification assumptions.⁴¹ Insulin showed the limit of this fidelity. Giving insulin lowered glucose within the simulation, so the model learned the drug's direction, but it did not reproduce the observed course of the patients who received it, whose glucose stayed high, probably because of the hyperglycaemia that prompted the infusion. This fidelity is what trial design and teaching need, because they rely on realistic variation in observed practice, but it is not sufficient for choosing an individual patient's treatment.

The privacy results suggest that the membership signal of a generative model can be set by design. Across

nine checkpoints and two optimizers, it aligned almost perfectly in rank with the fit reached on each database, so it could be bounded by limiting the passes over each database rather than discovered after training. The audit also covered prompted and intervened sampling, which every digital-twin application uses, whereas synthetic-data evaluations usually audit unconditional samples.^{15,16} A membership AUC of 0.55 is a weak but measurable signal: at a 1% false-positive rate, the attack flagged about 1–2% of members, against the 1% of non-members it flags by construction. Such audits are empirical and specific to the attacks tested, unlike formal guarantees such as differential privacy, and synthetic patients derived from data held under data-use agreements need governance before release.

Rollout-based saliency in decoder-only clinical models should not be read as evidence of what drove a prediction, because in a causal decoder attention rollout reduces to the geometry of the causal mask. The structural null we used, the rollout of uniform causal attention, is a cheap test for any attention-based attribution method. Editing and knockout depended on the trained weights and agreed with each other, but their agreement is partly expected because both remove a group’s information. Against the generator’s own effects, drawn a second time, knockout agreed as well as editing, so either can serve for explanations. Both stayed well below their reliability ceilings at the compute we could afford.

The two hospitals outside the training data show how Didymos behaves at a new site. Without further training, it transferred at the level of supervised models trained on the seven databases, while a model trained on the hospital’s own labels was ahead, most in a neurorehabilitation centre with sparse monitoring. Mapping the codes that the vocabulary lacked changed little, which suggests that the gap lay in the hospital’s practice rather than in its vocabulary. Continued next-token training on a third of the hospital’s unlabelled stays took minutes and closed that gap. The gains were largest for the remaining stay, treatment starts and laboratory events, whose timing follows local practice and is already written in unlabelled trajectories. A hospital with an unlabelled export can therefore adapt Didymos locally. Stacking the adapted forecasts with a model trained on a small labelled sample gave the best predictor we tested. Adaptation makes the hospital’s stays members of the adapted model, as any training does, so the adapted weights belong under that hospital’s governance.

The evidence supports different users to different degrees. For a trial statistician planning a covariate-adjusted analysis, Didymos risks computed at baseline cut simulated sample sizes by 14–37%, with type I error close to nominal in most settings. This is in line with the 10–30% reductions expected when a prognostic model explains a clinically realistic share of outcome variance.⁴⁹ Prognostic adjustment with generative digital twins has been developed for trials in Alzheimer’s disease,^{49,55} and its benefit can be checked on completed trials before use. Enrichment cuts sizes further, at the cost of screening more patients and narrowing the population to which the result applies. For researchers, synthetic cohorts reproduced database- and hospital-level structure and added information at small sites, with gains that shrank to about 0.01 AUROC once a site had 1,000 real stays. For educators, the toolkit generates synthetic cases that raters did not tell from real records. For clinicians at the bedside, the evidence is weakest. The forecast card improved raters’ risk estimates, but the card alone was at least as accurate as the aided raters, and raters followed it even when it pointed the wrong way. None of these analyses shows that patients benefit, which requires prospective evaluation in the workflow where a decision is taken, reported to early-clinical-evaluation standards and attentive to over-reliance.⁵⁶

The seven databases each contributed training stays, so the main results describe held-out patients from known databases. Transfer and adaptation were tested at two hospitals, with fewer than 30 events for several endpoints, and no single endpoint’s gain from adaptation held after adjustment. The adapted model’s membership test on the seven databases missed its pre-specified bound by a narrow margin, so its futures were drawn with the study lead’s permission and no synthetic data were made from it. The held-out split was also

used before evaluation, in a held-out calibration screen (300 stays per database) and in synthetic-data audits that informed the choice of checkpoint, and the test store was drawn from the same split. These audits used no discrimination endpoint and are unlikely to have favoured the forecasting comparisons, but calibration estimates could be modestly optimistic. The headline simulations used one seed, and the paper's two runs on the same anchors differed beyond simulation noise for one endpoint. The landmark sample for the two death endpoints was case-cohort, which preserves discrimination and calibration through weighting but makes trial sizes there relative. Race was recorded in three databases only, which limits the assessment of fairness. Simulation is costly, at about 10 GPU-hours for 128 futures on 1,750 anchors, although the hazard estimate needs far fewer futures, at one extra forward pass each. In the rater study, time per case and rater seniority were not analysed, and the card's benefit specific to Didymos was estimated rather than tested.

Prospective validation at new sites, and adaptation from fewer unlabelled stays with less exposure of each stay, would extend the transfer results. Training with the model's own samples and recalibrating within each site would address the delay of simulated deaths and the over-prediction of treatment starts. Re-analysing completed trials with the Trial designer would test its sample-size claims against real outcomes.

Methods

Study design and reporting

This retrospective study developed and evaluated a generative model on seven de-identified intensive care databases and is reported following TRIPOD+AI (Supplementary Table 3).⁵⁷ The model was trained on the train split of every database and evaluated on held-out stays from the same databases, so none of the seven was withheld from training. Two further databases, never used for training, tuning or model selection, tested transfer and adaptation to new hospitals. A written protocol fixed the designs, tasks, comparators, metrics and expected directions of the evaluation before any evaluation score was read. Later analyses were added as dated amendments, each fixed before it was computed, and the pre-specified results are reported unchanged beside every follow-up. Pre-specified analyses are the protocol and six early amendments that corrected definitions. Follow-up analyses, fixed after the pre-specified results were known and before they ran, are more simulated futures, the hazard estimate, the mortality decomposition, generator-level attribution, what-if validation, repeated augmentation, site prompts, additional comparators, the toolkit's validation, the re-run of the 48-hour landmark, the redraw of the generator's effects, and transfer and adaptation to the two further databases. Exploratory or post hoc analyses are the attention diagnosis, three toolkit checks added after a first result, the breadth analysis and the ranks of the main and transfer analyses. Reporting statistics (*P* values and Monte Carlo intervals) were fixed before any was computed. Patients and the public were not involved in the design, conduct or reporting of this study.

Data sources and cohort

The cohort comprised every adult ICU stay (age at least 18 years at hospital admission) in seven databases, 409,297 stays in total (Table 1):^{19,21,23,25,27,28,30} MIMIC-IV version 3.1 (Beth Israel Deaconess Medical Center, Boston, USA; admissions 2008–2022), eICU-CRD version 2.0 (208 US hospitals; 2014–2015), HiRID version 1.1.1 (Bern University Hospital, Switzerland), AmsterdamUMCdb version 1.0.2 (Amsterdam UMC, the Netherlands; 2003–2016), SICdb version 1.0.8 (Salzburg University Hospital, Austria; 2013–2021), NWICU version 0.1.0 (Northwestern Medicine, Chicago, USA; mostly 2020–2022) and INSPIRE version 1.4.2 (Seoul National University Hospital, South Korea; post-operative ICU stays). HiRID and INSPIRE do not release

admission years for individual stays. The unit of analysis was the ICU stay, from 48 hours before ICU admission to ICU discharge, with the header and outcome events of its hospitalization. Stays were split within each database by patient, through a hash of the database and patient identifiers, into train (80%), tuning (5%) and held-out (15%) sets. A test store of 2,000 held-out stays per database (14,000) supplied every prediction time and label of the evaluation, and up to 5,000 training stays per database supplied the comparators' training rows. The tuning split served early stopping, hyperparameter selection and recalibration. The two databases of the transfer analysis are described below. All processing ran locally or on hosts approved under each database's data-use agreement, and only aggregate results left them.

Harmonization and preprocessing

Each database was converted to the tables of CLIF version 3.0,³² through the official pipelines for MIMIC-IV and AmsterdamUMCdb and, for the other five, through adapters driven by mapping tables that record the provenance of every row (from ricu concept dictionaries, the mimic-code and eicu-code definitions and the source dictionaries).^{58,59} The tables passed schema and vocabulary validation without hard errors, as well as cross-database distribution checks. They were reconciled event by event with community pipelines to the Medical Event Data Standard (MEDS; <https://medical-event-data-standard.github.io>), with a documented reason for every unpaired row except two minor defects of the training snapshot, since fixed (four INSPIRE records and the rounding of HiRID infusion rates). They were then harmonized into MEDS, with the ICU stay as subject and one rule per topic for every database. Monitor streams were reduced to one value per hour, continuous medications were kept as one row per rate change (doses converted only with weights charted earlier) and binned demographics were set to their bin midpoints. Billing diagnoses entered only at discharge, never in a prediction input, and header fields documented after admission (such as the diagnosis group in four databases) were hidden from prediction inputs until their documentation window had passed. Derived concepts served only as labels: Kidney Disease: Improving Global Outcomes (KDIGO) stages of AKI from creatinine and urine output (mimic-code logic, reproduced exactly),⁶⁰ ventilation, vasopressor and RRT episodes, the first-day SOFA score⁴ and the Adult Sepsis Event,⁶¹ the last only in MIMIC-IV and AmsterdamUMCdb, which record cultures and administered antimicrobials.

Representation and grammar

Each stay became one token sequence (Supplementary Fig. 1b). A prefix gives the database, the site (eICU-CRD hospitals with at least 10 training patients), sex, a 5-year age band, the admission era and recorded race and ethnicity. A header gives the unit type, admission source and type, surgical status, diagnosis group, prior ICU stays, time in hospital before the ICU and the hour of admission. An optional goal block can state outcome fields for conditioning and is never a prediction target. The timeline follows as groups of events within 1 minute of each other, separated by one of 12 interval tokens (1–5 minutes to more than 7 days), with a clock block after every ICU-day crossing that states the day and each organ-support treatment's status. An end block gives the hospital discharge, death and billing diagnoses. A numeric event is a token of its code and value decile, with deciles per code fitted on the training values of all databases, and diagnoses and procedures are a code-system token followed by sub-tokens.¹⁰ The vocabulary, fitted on the train split, has 25,601 tokens, 19,774 of them event tokens over 4,434 codes. A window holds up to 32,768 tokens, and the 0.08% of training stays that are longer continue in windows that repeat the prefix and header. Values decode from per-bin sketches of the training values, so a decoded stay re-encodes to the same tokens, and derived quantities such as norepinephrine equivalents are computed after decoding rather than generated. A prediction

input is the stay as known at a cutoff: the timeline up to it, without the end block and with later-documented header fields set to unknown.

A grammar, enforced by an automaton during sampling, fixes the order of the blocks, allows each header field once, places intervals only between time groups, orders events within a group, permits a rate change or stop of an infusion only while it runs and writes the clock blocks. It accepted all 409,297 real stays, every stay re-encoded to its own tokens after decoding, and no prediction input contained an outcome or a later-documented field.

Model, training and selection of the final checkpoint

Didymos is a decoder-only transformer with the Qwen3 architecture,⁶² trained from random initialization, with 12 layers of width 768, grouped-query attention with 12 query and 4 key-value heads, rotary position embeddings (base 1,000,000), SwiGLU feed-forward layers of width 3,072, root-mean-square pre-normalization and an embedding tied to the output head, for 123,491,328 parameters (Supplementary Fig. 2a). It was trained once by next-token cross-entropy on the train split of all seven databases, excluding the goal and clock blocks from the loss. Training used transforms that leave the targets unchanged: a goal block with a random subset of fields in 25% of windows, shuffled header fields, and the database and site set to unknown in 10% of windows and each header field in 5%. The number of passes over each database (MIMIC-IV 1.59, eICU-CRD 2.66, HiRID 2.42, AmsterdamUMCdb 3.05, SICdb 2.01, NWICU 1.96, INSPIRE 1.81) was set so that its membership-inference AUC, projected from slopes measured on earlier checkpoints, was 0.555. The optimizer was Muon on hidden weight matrices and AdamW on embeddings and norms,⁶³ with a learning rate of 6×10^{-4} , 100 warm-up steps, cosine decay to 10%, weight decay 0.1, gradient clipping at 1.0 and batches of 524,288 tokens. Training took 1,990 steps (997.5 million tokens) and 47 minutes on two NVIDIA RTX PRO 6000 Blackwell GPUs, without dropout, and reached a tuning loss of 1.5525 nats per token.

Didymos 1.0 was chosen among ten candidate checkpoints that varied size, context, clock blocks, input noise, optimizer and passes (Supplementary Table 4), each judged on the tuning loss, the privacy audit (below) and audits of synthetic stays, by rules written before it was measured. In the final round, a held-out calibration screen compared the observed-to-expected ratios of simulated death, new IMV, vasopressor and RRT and an ICU stay over 7 days on 300 held-out stays per database (16 futures each), and authenticity audits compared synthetic stays with 1,000 held-out stays per database. The adopted checkpoint met the privacy audit, had the lowest tuning loss among the candidates that met it, met more authenticity criteria than the reference checkpoint (30.5 against 25.0 of 35) and was no worse on calibration (21 against 22). One departure from the rules is recorded: the study lead sent this checkpoint, rather than the rule's tie-break winner, to the final decision run, because the tie-break rested on a difference within one standard error. The held-out split thus informed the choice of checkpoint through calibration and authenticity, but no discrimination endpoint, comparator or anchor set of this study was used, and no held-out stay entered training. The test store was drawn from the same held-out split, so it overlaps these audit samples.

Simulation and the two estimates

Simulated futures were drawn by ancestral sampling at temperature 1, without top- k , nucleus or repetition penalties, which would remove the tails of the distribution, with vLLM⁶⁴ and a reference sampler that gives identical greedy output (Supplementary Fig. 2b). At each step, next-token probabilities were restricted by the grammar, by the database's support (the tokens its training windows contain) and by a token budget that reserves room for the end block, then renormalized. A future ended at the end block or at a token cap. For the

pre-specified analysis, the cap was the database's 95th percentile of tokens over 72 hours (rolling anchors) or 7 days (landmarks), and for the follow-up at the landmark, four times that cap.

One label engine computed every label, on a stay's real future and on each decoded simulated future alike. The sampled estimate of a binary endpoint is the share of valid futures with the event, scored when at least half the futures were valid. For a remaining ICU stay over 7 days, it is the Kaplan–Meier survival at 168 hours, with capped futures censored. The hazard estimate passes each simulated future through Didymos once and, at every timeline position, reads the probability that the next token is the event (death, alive discharge, or the first IMV, vasopressor or RRT record), renormalized over the tokens the sampler could draw, which reproduced the exact masked probabilities within 0.975–1.006. The step probabilities are summed in hourly bins while the future is at risk, with death and alive discharge competing, and averaged over futures. The result is the compensator of the sampled events, a Rao–Blackwellized Aalen–Johansen estimator with the same expectation and lower variance.³³ Within the simulated futures, summed hazards matched sampled events for every cause (ratios 0.993–1.022). The hazard estimate was used for death, new IMV, vasopressor and RRT, and the sampled estimate for the other endpoints.

Prediction tasks and anchor designs

Endpoints came from a catalog of 72 programs (binary events within a horizon, times to event and regressions) grouped into families: mortality, organ failure, sepsis, intervention onsets, discharge, threshold crossings and the remaining stay. Twelve were pre-specified as primary. On rolling anchors, these were death, new IMV and a new vasopressor within 24 hours, new RRT and AKI of stage 2 or more within 48 hours, sepsis (Adult Sepsis Event) within 24 hours, hypotension within 4 hours, hyperkalaemia within 24 hours and thrombocytopenia within 48 hours. At the 24-hour landmark, they were ICU death, in-hospital death and a remaining ICU stay over 7 days. The follow-up added death within 24 hours of the landmark. A threshold on a value was placed at the decile edge of its tokens nearest the clinical cut-off, because a cut-off inside a token bin would make the label depend on the decoder's draw. Hyperkalaemia was therefore potassium above 4.9 mmol/L, the top decile edge, thrombocytopenia a platelet count below 71×10^3 per μL and hypotension a MAP below 67 mmHg (invasive) or 62 mmHg (non-invasive) sustained for 30 minutes (definitions in Supplementary Table 5). As an alignment control, each anchor's real stay, decoded on its own clock, reproduced the real store's labels.

The pre-specified analysis used three anchor sets with 32 futures per anchor (Supplementary Fig. 3). Each landmark set took up to 500 stays per database that were in the ICU and alive at 24 or 48 hours. The rolling set took 500 anchors per database at random ICU hours, spread over length-of-stay bands, with half of each band enriched for an organ-support start or ICU death in the next 48 hours and weighted by the inverse probability of inclusion. The follow-up drew 128 futures per anchor on two subsets, 150 rolling anchors per database (1,050) and a case-cohort sample of the 24-hour landmark (every in-hospital death, 346, and 354 other stays, weighted to the whole landmark set).³⁴ A later amendment, fixed before it ran, repeated the 48-hour landmark with the follow-up's settings on a case-cohort sample of its anchors (every in-hospital death, 455, and 354 weighted controls; 128 futures with four times the pre-specified token caps). Every comparator was scored on the same anchors at the same prediction time.

Comparators

Learned comparators were trained on labelled rows of real training stays drawn with the anchors' designs, with features decoded from the same tokens Didymos sees. For the 64 most frequently charted numeric channels, the features were the last value, mean, minimum, maximum and count over the 6 and 24 hours

before the prediction time and the hours since the last value, together with the active organ-support treatments, the statics documented at admission, the database and the hours since ICU admission. LightGBM³⁵ was fitted pooled (with the database as a feature) and per database (up to 1,000 trees, learning rate 0.05, 31 leaves, early stopping on the tuning split). TabPFN-3.5 (Prior Labs; weights run locally and offline)³⁶ used all features and a stratified sample of up to 50,000 rows, with positives kept up to half and the odds corrected back to the training prevalence. TabPFN v2 used up to 10,000 rows and the features its input limit allows. TabPFN was not fitted for death within 24 hours at the landmark. The other learners (logistic regression with L2, L1 and elastic-net penalties, a decision tree, random forest, extra trees, Gaussian naïve Bayes, k -nearest neighbours and a multilayer perceptron) used median imputation, missing-value indicators and standardization, with hyperparameters chosen on the tuning split from pre-specified grids. The Didymos probe is a ridge logistic regression on Didymos’s last-layer state at the end of the prompt, trained with labels. Every comparator’s settings are listed in Supplementary Table 6.

Clinical scores were computed without training from the harmonized raw values of the anchor’s stay, using only data recorded or resulted in the 24 hours before the prediction time, with missing components scored as normal. SOFA used the worst values,⁴ SAPS II (without its chronic-disease items) and OASIS were computed at the landmarks,^{5,65} and NEWS, the Modified Early Warning Score (MEWS), quick SOFA (qSOFA) and the systemic inflammatory response syndrome (SIRS) criteria used the latest values.^{6,66–68} The windowed SOFA equalled the pipeline’s first-day SOFA in 96–99% of 24-hour anchors in six databases. The databases’ own admission scores were APACHE IVa in eICU-CRD, SAPS 3 in SICdb and the American Society of Anesthesiologists (ASA) physical status class in INSPIRE.^{38,39} Single bedside values, such as the shock index, lactate, the ROX index (oxygen saturation over the inspired oxygen fraction, divided by the respiratory rate), MAP, creatinine, potassium and platelets, were scored at their latest value, with a missing value set to normal.

Calibration and recalibration

Calibration was summarized by the observed-to-expected ratio (O/E; above 1, under-prediction), the expected calibration error over 10 equal-mass bins and reliability curves.⁵⁴ Step calibration along real futures compared the model’s step hazards, summed along each anchor’s real future, with the observed events. Generation drift was the ratio of the simulated futures’ summed death hazard to the real futures’, over 24 hours, until ICU discharge and by hour. Censoring was examined by applying the pre-specified token cap to the landmark’s longer futures, and whole synthetic stays were compared with real held-out stays by length of stay. Recalibration maps were fitted on the tuning split (100 rolling anchors per database, 32 futures) and applied to new draws on the held-out rolling anchors. Isotonic maps, weighted by inclusion, were adopted where AUROC stayed within 0.005 at both 32 and 16 futures, which held for death within 24 hours, AKI, hypotension and hyperkalaemia. For the three treatment starts, logistic maps, $\text{logit } p' = a + b \text{ logit } p$, were fitted post hoc on the same anchors after the isotonic maps had tied risks. They were adopted because $b > 0$ and, at both 32 and 16 futures, O/E moved closer to 1 without a larger calibration error. Being strictly increasing, they cannot change AUROC.

Privacy audits

The membership-inference audit of training compared up to 3,000 training stays (members) with 3,000 held-out stays (non-members) per database, scoring each by Didymos’s mean negative log-likelihood per token (the loss attack).⁴³ It reported the AUC and the true-positive rate at a 1% false-positive rate with bootstrap intervals.⁴⁴ It also tested prefix-prompted extraction, exact copies and nearest-neighbour distances, a density-ratio attack and attribute inference, and compared membership AUC with the gap between members’ and non-members’

losses across nine checkpoints. For prompted and intervened sampling, stays of at least 30 hours were cut at a whole hour drawn uniformly from 4 hours after the start of the stay to 25 hours before its end, and the loss attack scored the real next 24 hours given the prompt (up to 3,000 stays per side). Extraction asked whether any of 16 simulated futures reproduced the first 32 timeline tokens of the real future, and how similar the best future's events were over 24 hours, for plain prompts (up to 150 per side) and for norepinephrine starts (the give arm, up to 100 per side).⁴⁵ A copying generator, which returns members' real futures, and a private reference, two disjoint halves of the non-members, served as controls.

The pre-specified rule judged an audit acceptable when, per database and pooled, the upper 95% bound of the AUC was at most 0.60 and that of the true-positive rate at most 0.03, unacceptable when a lower bound exceeded 0.60 or 0.05, and inconclusive otherwise. The audit of training met the rule in every database. For prompted sampling, an audit with 500 stays per side left three databases inconclusive. A repeat at full size (post hoc) met the rule pooled and in every database except INSPIRE, whose held-out split yields only 790 eligible stays per side and whose true-positive-rate interval (0.009 to 0.039) straddled 0.03, an inconclusive result. The study lead recorded the prompted audit as passed, with INSPIRE's sample-limited result stated. Extraction and both controls behaved as specified.

Fidelity and utility of synthetic stays

The Didymos synthetic cohort comprised 13,416 stays conditioned on each database. It was drawn after the privacy audit and resampled by iterative proportional fitting so that length-of-stay bands and organ-support onsets after the 24-hour landmark matched the training stays' margins, which corrects the cohort's composition but not the model's dynamics. Per-database profiles were compared with held-out stays and with a second sample of training stays (the real-against-real control). The structure between databases was the correlation between the synthetic and real matrices of between-database differences. eICU-CRD hospitals with at least 30 held-out and 30 synthetic stays (17 hospitals) were compared by the rank correlation of hospital-level rates, and rare header strata (at least 20 training stays and under 0.5% of a database) were counted. Utility was the ratio of gains over chance, $(TSTR - 0.5)/(TRTR - 0.5)$, of LightGBM trained on synthetic stays (TSTR) or on as many real stays (TRTR) and tested on real held-out anchors. For augmentation, LightGBM per database was trained on 100, 300 or 1,000 random real training stays, alone and with one or three times as many synthetic stays, in 20 repetitions, and tested on that database's held-out anchors. Summaries use endpoint-database cells with at least 20 test positives.

What-if simulation

An intervention writes a treatment's start record at the cutoff (give) or forbids the treatment's tokens throughout (withhold), with 16 futures per arm capped at the database's 24-hour token quantile. In the pre-specified design, held-out stays whose first record of a treatment in the ICU was a start at least 2.5 hours after admission were cut at a time drawn uniformly 0.5–6 hours before it (up to 200 per treatment) and given the database's most common start record at its median rate. Nine responses were pre-specified with their expected signs, each over the 6 hours after the cutoff: MAP after norepinephrine and after a fluid bolus, glucose after insulin, potassium after continuous RRT and creatinine over 24 hours after it, urine output after furosemide and after a fluid bolus, and inspired-oxygen charting and oxygen saturation after intubation. Single urine-output records above 2,000 mL, cumulative totals in one source, were dropped. In the follow-up, starters were cut 1 minute before their first recorded start and matched within the database to a held-out non-starter at the same ICU hour (within 6 hours), by nearest-neighbour matching on the propensity of starting within 6 hours

620 (LightGBM on the comparator features; caliper 0.2 standard deviations of the logit).⁴¹ Starters received give
621 (their own recorded start) and withhold, and non-starters likewise. The observational contrast was the starters'
622 real response minus their matches', Didymos's factual contrast the starters' give arm minus their matches'
623 withhold arm, and the within-model effect the starters' give minus withhold arm. The null arm compared two
624 withhold arms drawn with different seeds. Factual accuracy was the CRPS of the factual ensemble relative to
625 persistence and the coverage of its central 80% interval.⁴²

626 Attribution and sanity checks

627 A patient's record before the prediction time was divided into groups, each a channel family (vital signs,
628 laboratory panels, vasoactive drugs, sedation, fluids and outputs, respiratory support, RRT, other medications,
629 assessments) within a time block (0–6 hours, 6–24 hours, earlier), plus the admission details. The target was
630 the Didymos probe's risk for death within 24 hours, new IMV, AKI or hypotension. Counterfactual editing
631 removed a group's events from the decoded stay and re-encoded the prompt, which keeps it grammatical.
632 Integrated gradients used 50 steps from the padding embedding,⁴⁸ attention rollout multiplied the 12 layers'
633 head-averaged attention, each mixed with the identity,⁴⁷ and attention $\times$ gradient weighted attention by its
634 gradient, with token scores summed per group. In the follow-up, 40 anchors per endpoint were drawn from
635 the higher-risk half, and each group's edit was simulated with 48 futures to obtain the generator's own effect,
636 whose split-half reliability (Spearman–Brown corrected) bounds any agreement. A later amendment drew
637 these effects again for all 160 anchors, with 48 new futures per edit and a new seed, and scored attention
638 knockout against them beside editing. Sanity checks replaced the model's weights with a random initialization
639 of the same architecture, keeping the trained probe.¹⁷ An exploratory diagnosis, added after attention rollout
640 failed this check, compared rollout with a structural null, the rollout of uniform causal attention with no model,
641 and tested remedies on the 152 anchors with prompts of at most 4,096 tokens. The remedies were group
642 means instead of sums, norm-weighted rollout,⁶⁹ gradient-weighted rollout,⁷⁰ the last layer's attention and
643 attention knockout, which blocks attention from every other token to a group's tokens in all layers and records
644 the fall in the probe's logit.⁴⁶ The rule set before the diagnosis (a sanity correlation of about 0.3 or less and
645 positive agreement with editing) was met by knockout and by three gradient or last-layer variants with group
646 means, and not by the rollout variants (Supplementary Table 7).

647 Similarity and site prompts

648 Leave-one-out nearest neighbours ($k = 5\text{--}50$) predicted each anchor's outcome as the share of neighbours with
649 the event, with neighbours found by Didymos's last-layer state, by TF-IDF of the prompt's event tokens, by
650 the comparators' features or by the state of an untrained model of the same architecture (a negative control).
651 Site prompts were tested on the follow-up rolling anchors with 32 futures, with the database and site tokens
652 unknown, as in training, or with the database token replaced by another database's. For supervised transfer,
653 LightGBM and the probe were trained on six databases and tested on the seventh, against training within the
654 database.

655 The Didymos toolkit and trial simulation

656 The toolkit computes the paper's quantities as functions, with command-line and programming interfaces and a
657 self-contained report card for each output that states the evidence and caveats behind it. Recalibrated risks use
658 the maps in force. The toolkit was validated with new draws on a separate host. The pre-specified conformance

rule required the AUROC and O/E of the nine rolling primary endpoints to lie within the pre-specified run's 95% intervals, the hazard estimate's AUROC to lie within 0.02 of the follow-up's curve, and the norepinephrine and furosemide what-if contrasts to lie within the follow-up's intervals. Of its 15 checks, 14 were met, the exception being sepsis, with 8 positive anchors. A check added before it was computed, which allows for both runs' simulation noise ($1.96 \sqrt{2}$ times the standard deviation of the toolkit's estimate under resampling of futures within anchors), was met by 7 of 9 endpoints. A criterion set afterwards by the study lead, conformance within that limit of either of the paper's two runs, was met by all nine (Supplementary Table 7).

The Trial designer simulates two-arm trials. Each trial draws N anchors with replacement from an eligible population, randomizes them 1:1 and imposes a constant relative risk by keeping a treated patient's observed event with that probability. The unadjusted analysis is a two-proportion z test, and the adjusted analysis is a logistic regression on treatment and the prognostic score (the logit of a risk clipped at 0.5%, or a standardized raw score), tested by the treatment's Wald statistic.⁴⁹ The size for 80% power at a two-sided α of 0.05 follows from each analysis's simulated power at the analytic unadjusted size, with the test statistic taken as normal with a mean proportional to $\sqrt{N}$. Each cell was simulated with 2,000 trials at relative risks of 0.8 and 0.7. Populations were the 24-hour landmark's case-cohort sample for ICU and in-hospital death (128 futures) and its natural sample for the other endpoints (32 futures). Prognostic scores were Didymos's sampled estimate, LightGBM, SOFA, SAPS II, OASIS and NEWS, and APACHE IVa on eICU-CRD anchors, and enrichment enrolled the highest-risk 75%, 50% or 25%. The checks, type I error of 0.04–0.06 at relative risk 1 and an unadjusted simulated size within 5% of the analytic size, were each met in 92 of 108 cells. Teaching cases were built from the synthetic cohort for three pre-specified objectives (shock, respiratory failure and renal support), three cases each, and checked for meeting the objective and for being self-contained.

Rater study

The rater study was fixed in a protocol amendment, with sizes from simulation and its analysis in code, before any stimulus was built or answer collected. It planned 10–20 critical care physicians from two to four centres, credentialed for MIMIC-IV and eICU-CRD and not involved in model development. Each received a self-contained packet of real held-out stays from those databases and synthetic stays conditioned on them, rendered identically from tokens. Module A showed 80 first-day case sheets (40 real and 40 synthetic, matched on database and length of stay) to be called real or synthetic. Module B1 showed 24 treatment starts followed by either one simulated future of the give arm or the real 6-hour course, rated for plausibility on a 1–5 scale, with insulin as a separate stratum. Module B2 showed 20 rolling anchors with the top factors from editing, knockout or random groups, rated for relevance. In module C, raters estimated the risks of ICU death and AKI for 48 landmark cases (deaths enriched to about a quarter), first unaided and then with a forecast card from 128 futures (ICU death raw, AKI recalibrated). The pre-specified primary analyses were pooled accuracy equivalent to chance in A (a 90% CI within 0.40–0.60; mixed model with raters and items), non-inferiority of synthetic courses in B1 (margin –0.5 points, insulin excluded), editing and knockout each rated above random factors in B2 (Holm), and the ICU-death Brier score with minus without the card in C. Mixed models were fitted in R 4.3.3 with lme4 1.1-35.5 and cross-checked by bootstrap.

Answers for all 20 packets were analysed with the pre-specified code, and the study lead attested that they came from the study's raters. The protocol's per-packet export checks (checksum, consent and timing) were not run, and the two-person pilot did not precede the study. The timing, role and qualification fields of the received files were not analysed, because their dates preceded the creation of the study items, so time per case, seniority and eligibility were not assessed. Against the pre-specified rules, accuracy in module A was

0.408 (90% CI 0.351 to 0.468), below the equivalence range, so equivalence to chance was not shown because accuracy fell below chance (a post hoc non-detection analysis gave an upper one-sided 95% bound of 0.47). In B1 the difference was -0.23 (-0.56 to 0.11), so non-inferiority was not established. In B2 neither method was rated above random factors (Holm $P = 1.00$ each), and in C the Brier score improved with the card (-0.016 ; -0.031 to -0.001 ; $P = 0.037$), meeting the primary criterion.

Transfer to unseen hospitals and adaptation

Two databases were added after the main analyses, under protocol amendments fixed before any of their futures was drawn. RICD version 3.0 (Federal Research and Clinical Center of Intensive Care Medicine and Rehabilitology, Moscow, Russia; intensive care and neurorehabilitation; admissions 2017–2025; 5,382 ICU stays) is distributed by its provider (<https://fnkcurr-dataset.ru/>), and zigongICU version 1.1 (Zigong Fourth People's Hospital, China; ICU patients with infection; 2,899 stays) through PhysioNet.^{50–52} Neither was used for training, tuning or model selection of Didymos. New adapters with versioned mapping tables brought them into the common format. zigongICU's nursing chart records dates without hours, so its rows were spread in order over each day, and RICD's daily medication orders became infusion episodes. Both were encoded with the frozen vocabulary, without grammar rejections, and every stay was a test stay.

A pre-run gate recorded the share of events without a token, value distributions against the training databases and an alignment control (each anchor's real stay reproduced its labels, agreement 1.0). It fixed the scored endpoints by a rule set beforehand: the endpoint's defining source is timed to the hour, and at least 20 anchors of the design are positive. The designs were the rolling anchors (500 per database; 998 kept after the clock check) and a case-cohort sample of the 24-hour landmark (up to 350 in-hospital deaths and 350 other stays per database; 1,400). Each used 128 futures, prompts with the database and site unknown, and the follow-up's estimates. The comparators were the learned models trained on the seven databases, TabPFN-3.5 and the clinical scores. A local LightGBM, trained within each database with the same recipe and five-fold cross-fitted by patient, served as the reference with local labels. Logistic recalibration within each database was cross-fitted by the same folds.

Adaptation followed a further amendment. Each database's patients were assigned by a hash of database and patient identifier (seed 17) to a locked test set (50%), unlabelled stays for continued training (35%) or a labelled slice (15%). The test anchors were the transfer analysis's anchors from test-set patients, with the case-cohort weights recomputed within the test set, and every arm used the transfer analysis's labels. The code mapping, used for these two databases only, replaced a code the vocabulary lacked with the nearest kept code of the same concept. An infusion's unit became an unspecified or kept unit, a dose's unit and route a kept variant with unit conversion, stool other gastrointestinal output, and an unmapped location its nearest kept location (Supplementary Table 8). Continued training started from the final weights with a fresh optimizer, kept the next-token objective and set the database and site tokens to unknown. The unlabelled stays were mixed one to one, by training window, with the seven databases' training windows. A grid fixed in advance crossed learning rates of 3×10^{-5} and 10^{-4} with one or three passes over the unlabelled stays (warm-up 2%, cosine decay to 10%, 131,072 tokens per step). The cell with the lowest next-token loss on a held-out 10% of the unlabelled patients was chosen, and the loss on the seven databases' tuning split measured forgetting.

The labelled slice supplied anchors drawn with the transfer analysis's rules (250 rolling anchors per database and a landmark case-cohort sample of up to $125 + 125$). Recalibration fitted the intercept and slope of the logit risk, or the intercept alone below 10 positives. Stacking fitted a logistic regression on the logits of adapted Didymos and of a LightGBM trained on the labelled slice, cross-fitted by patient. Local

LightGBMs were trained on the labelled slice and on the unlabelled and labelled stays together. For label efficiency, LightGBM was trained on 25, 50, 100, 200 or all labelled stays, 20 random draws each, and the median AUROC was compared with adapted Didymos. Membership of the continued-training stays was tested against the test stays with the loss attack of the privacy audit, and the adapted model's audit on the seven databases repeated the original's. In that audit, one group's upper bound of the true-positive rate at a 1% false-positive rate exceeded the pre-specified 0.03 (eICU-CRD, 0.0303). A larger test, fixed before it ran with up to 10,000 stays per side, placed AmsterdamUMCdb's bound at 0.0313 (2,747 per side, all its held-out stays), with every other test within its limit (Supplementary Fig. 22; Supplementary Table 7). The study lead permitted the adapted model's futures for this analysis only. No synthetic data were produced from it, and its futures were deleted after scoring.

Statistics and reproducibility

Unless stated, estimates are on held-out anchors with 95% intervals from 1,000 bootstrap resamples of stays, stratified by database and weighted by the designs' inclusion or case-cohort weights. Differences between predictors on the same anchors use the paired bootstrap. Resampling was used because AUROC differences and ratios have no convenient closed-form variance under these designs. P values were added under a plan fixed before any was computed, without changing any estimate or interval (all reporting statistics in Supplementary Data 2). Where both predictors' per-anchor values were kept, the stored computation was first rerun and had to reproduce every stored interval within 10^{-12} (62 reproduced). The paired bootstrap was then repeated with 10,000 resamples and a new seed, giving the two-sided $P = \min(1, 2(1 + k)/10,001)$, where k counts resamples on the other side of zero, so the smallest attainable value is 0.0002. P values below 0.001 are reported as $P < 0.001$, and Holm-adjusted values that inherit this bound as upper bounds ($P \leq$). For the hazard estimate, whose per-anchor values were not kept, P came from the stored interval by a normal approximation, with the standard error from the half-width on the side of zero.⁷¹ Applied to the 36 comparisons with exact P values, the approximation differed by a median of 0.06 in $\log_{10}P$ (maximum 0.72) and gave the same decision at 0.05 in all 36 (Supplementary Table 9).

Tests were two-sided at $\alpha = 0.05$, with Holm's adjustment within each family.³⁷ Four families compared predictors: Didymos against LightGBM on the 13 primary endpoints (the primary family), against TabPFN-3.5 (12) and against the best clinical score per endpoint (13), and the follow-up against the pre-specified estimate (13). The other families tested O/E against 1, and its change after recalibration (7 endpoints); simulated against observational contrasts (9 responses, with a standard error that combines the two stored intervals as if independent, which is conservative when the estimates are positively correlated); the null arm's expected-sign share against 0.5 (exact binomial tests); membership AUC against the loss gap across 9 checkpoints (Spearman correlation with exact permutation tests over all 9! orderings; 7 databases); augmentation gains (sign tests and Wilcoxon signed-rank tests; 12 combinations); and knockout's against rollout's agreement with editing over 152 anchors (Wilcoxon signed-rank tests, pooled and per endpoint; 4 endpoints).

Later amendments added families under the same rules, each fixed before its results were computed. For the 48-hour landmark re-run, Didymos was compared with LightGBM and with its pre-specified estimate (4 endpoints each), and for the redraw, knockout with editing in their agreement with the redrawn effects (4 endpoints). At the new databases, Didymos was compared with the transferred and the local LightGBM, and its sampled estimate with the best clinical score (15 pairs each). For adaptation, each arm was compared with the zero-shot model (Holm within each arm), and the recalibrated adapted model with the local LightGBM on all adaptation labels and with TabPFN-3.5 (13 pairs each, endpoints with at least 10 test positives), with mean

ranks over the same pairs. Their stored intervals were reproduced before any P value (118 reproduced).

The breadth analysis, defined after the per-endpoint intervals were known, used Benjamini–Hochberg false-discovery rates across 165 comparisons,⁴⁰ which are valid under the positive dependence that shared anchors make plausible, with Holm as a sensitivity analysis and summaries from one joint bootstrap that resampled stays once across all comparisons. Ranks across endpoints (post hoc) used a Friedman test and two-sided Wilcoxon signed-rank tests of Didymos against each predictor on the paired AUROC differences, Holm-adjusted, with intervals for mean ranks from 10,000 bootstrap resamples of the endpoints. Because endpoints share anchors, these tests describe consistency across tasks. For the trial simulations, type I errors carry Clopper–Pearson intervals and sample-size ratios carry Monte Carlo intervals from a parametric bootstrap of the two simulated powers (10,000 draws), which overstates their variance because both powers come from the same trials. Pre-specified decision rules are reported beside their observed estimates (Supplementary Table 7). “Significant” is used only with a P value. Numbers of anchors, events and pairs are given in the figures and legends.

Analyses used Python 3.12 with PyTorch 2.13, Transformers 5.18, vLLM 0.31, LightGBM 4.7, scikit-learn 1.9, SciPy 1.18 and statsmodels 0.15, and R 4.3.3. Seeds were fixed for every design, simulation and resampling: 0 for the pre-specified analysis, 7 for the follow-up, 16 for the transfer analysis, the 48-hour re-run and the redraw, and 17 for adaptation. Every figure has source data listing its inputs’ sha256 and the code version (Supplementary Data 3). Simulated futures were deleted after their analyses under the data-use agreements, and the aggregates behind every number are kept.

Ethics

This study used de-identified data released for research under credentialed access and data-use agreements, through PhysioNet for MIMIC-IV, eICU-CRD, HiRID, SICdb, NWICU, INSPIRE and zigongICU,³¹ through Amsterdam Medical Data Science for AmsterdamUMCdb, and under its provider’s terms of use for RICD. Each database was collected and released under the approval of its institutions’ review boards, with informed consent obtained or waived as described in its release. Informed consent was obtained from all raters, whose identities are not disclosed.

Use of AI tools

An AI coding assistant (Codex, OpenAI) was used under H.W.’s direction for code editing and for language editing of the manuscript. H.W. designed and performed the experiments. The authors take responsibility for the content. The schematic artwork in Figs. 1a, 1c, 4a and 7a and Supplementary Figs. 1–3 was generated with GPT-image (OpenAI) from prompts that specified every label and number, and is credited “(GPT-image)” wherever it appears (native files and their content credentials in Supplementary Data 4). The authors confirm the accuracy and scientific validity of this artwork and assume full responsibility for it. All data plots were drawn by code from the results. TabPFN-3.5 ran locally and offline and sent no data to external services.

Code availability

The code for harmonization, tokenization, training, simulation, evaluation, statistics, figures and the Didymos toolkit is available at <https://github.com/Hanziwww/Didymos/releases/tag/v0.1.0>. In the code, Didymos 1.0 is the checkpoint MUC32_main, the package is icutwin and the toolkit’s command is icutwin_toolkit; Supplementary Table 10 maps the paper’s names to the code’s and to the protocol’s.

Data availability

The seven databases are available to credentialed researchers under their data-use agreements: MIMIC-IV, eICU-CRD, HiRID, SICdb, NWICU and INSPIRE through PhysioNet, and AmsterdamUMCdb through Amsterdam Medical Data Science.^{19,21,23,25,27,28,30} The two databases of the transfer and adaptation analyses are available from their providers: zigongICU through PhysioNet under credentialed access, and RICD from its provider (<https://fnkcr-dataset.ru/>) under the provider's terms of use.^{50,52} The aggregate results behind every figure and table are provided with this article: the comparison of every scored endpoint (Supplementary Data 1), the reporting statistics (Supplementary Data 2) and the source data of every figure (Supplementary Data 3). Individual-level data, simulated futures and synthetic stays derived from the databases cannot be shared by the authors under these agreements. The model weights are available on application at <https://huggingface.co/HanZiwww/Didymos-1.0>. Access is reviewed manually under the applicable data-use agreements.

Acknowledgements

We thank the teams that developed and maintain MIMIC-IV, the eICU Collaborative Research Database (eICU-CRD), HiRID, AmsterdamUMCdb, the Salzburg Intensive Care database (SICdb), the Northwestern ICU (NWICU) database, INSPIRE, zigongICU and the Russian Intensive Care Dataset (RICD) for making these data available for research.

We also thank the critical care physicians at the Affiliated Hospital of Guangdong Medical University; Guangdong Provincial People's Hospital, Southern Medical University; and the First Affiliated Hospital of Jinan University for their support with the clinical evaluations.

This work was supported by funding to C.T. from the National Natural Science Foundation of China (grant nos. 82160215 and 81660174); the Guangdong Provincial Key Laboratory of Autophagy and Major Chronic Non-communicable Diseases (grant no. 2022B1212030003); the Research Startup Funds for High-level Talents in the Affiliated Hospital of Guangdong Medical University (grant no. GCC20220013); the College Students' Innovative Entrepreneurial Training Plan Program (grant no. S202510571025); and the 2025 Artificial Intelligence-Enabled Medical Device Innovation Task (Open Competition Mechanism, project no. 2025-AISJK-02007). The funders had no role in the study design, data collection, analysis and interpretation, or the writing of the manuscript.

Author contributions

H.W. designed the experiments, collected the data, wrote the code, performed all experiments, prepared the figures and drafted the manuscript. K.Y., H.Y., W.C., Y.C., Y.T., T.L., Y.X., X.L., L.P., H.L., Y.M., Z.Y., X.Z., J.H. and C.T. provided feedback and reviewed the manuscript.

Competing interests

The authors declare no competing interests.

References

- [1] Katsoulakis, E. *et al.* Digital twins for health: a scoping review. *npj Digit. Med.* **7**, 77 (2024).
- [2] Lal, A. *et al.* Development and verification of a digital twin patient model to predict specific treatment response during the first 24 hours of sepsis. *Crit. Care Explor.* **2**, e0249 (2020).
- [3] Rockenschaub, P. *et al.* The impact of multi-institution datasets on the generalizability of machine learning prediction models in the ICU. *Crit. Care Med.* **52**, 1710–1721 (2024).
- [4] Vincent, J. L. *et al.* The SOFA (Sepsis-related Organ Failure Assessment) score to describe organ dysfunction/failure. *Intensive Care Med.* **22**, 707–710 (1996).
- [5] Le Gall, J. R., Lemeshow, S. & Saulnier, F. A new Simplified Acute Physiology Score (SAPS II) based on a European/North American multicenter study. *JAMA* **270**, 2957–2963 (1993).
- [6] Smith, G. B., Prytherch, D. R., Meredith, P., Schmidt, P. E. & Featherstone, P. I. The ability of the National Early Warning Score (NEWS) to discriminate patients at risk of early cardiac arrest, unanticipated intensive care unit admission, and death. *Resuscitation* **84**, 465–470 (2013).
- [7] Renc, P. *et al.* Zero shot health trajectory prediction using transformer. *npj Digit. Med.* **7**, 256 (2024).
- [8] Kraljevic, Z. *et al.* Foresight—a generative pretrained transformer for modelling of patient timelines using electronic health records: a retrospective modelling study. *Lancet Digit. Health* **6**, e281–e290 (2024).
- [9] Shmatko, A. *et al.* Learning the natural history of human disease with generative transformers. *Nature* **647**, 248–256 (2025).
- [10] Waxler, S. *et al.* Generative medical event models improve with scale. Preprint at <https://arxiv.org/abs/2508.12104> (2025).
- [11] Makarov, N. *et al.* Large language models forecast patient health trajectories enabling digital twins. *npj Digit. Med.* **8**, 588 (2025).
- [12] Guo, W. DT-ICU: towards explainable digital twins for ICU patient monitoring via multi-modal and multi-task iterative inference. Preprint at <https://arxiv.org/abs/2601.07778> (2026).
- [13] Pickard, J. *et al.* EHR-MPC: inference-time control for sepsis treatment with generative patient digital twins. Preprint at <https://arxiv.org/abs/2607.08793> (2026).
- [14] Wornow, M. *et al.* The shaky foundations of large language models and foundation models for electronic health records. *npj Digit. Med.* **6**, 135 (2023).
- [15] Stadler, T., Oprisanu, B. & Troncoso, C. Synthetic data—anonymisation groundhog day. Preprint at <https://arxiv.org/abs/2011.07018> (2020).
- [16] Yan, C. *et al.* A multifaceted benchmarking of synthetic electronic health record generation models. *Nat. Commun.* **13**, 7609 (2022).
- [17] Adebayo, J. *et al.* Sanity checks for saliency maps. Preprint at <https://arxiv.org/abs/1810.03292> (2018).
- [18] Johnson, A. E. W. *et al.* MIMIC-IV, a freely accessible electronic health record dataset. *Sci. Data* **10**, 1 (2023).
- [19] Johnson, A. *et al.* MIMIC-IV (version 3.1). PhysioNet (2024). URL <https://doi.org/10.13026/kpb9-mt58>.
- [20] Pollard, T. J. *et al.* The eICU Collaborative Research Database, a freely available multi-center database for critical care research. *Sci. Data* **5**, 180178 (2018).
- [21] Pollard, T. *et al.* eICU Collaborative Research Database (version 2.0). PhysioNet (2019). URL <https://doi.org/10.13026/C2WM1R>.
- [22] Hyland, S. L. *et al.* Early prediction of circulatory failure in the intensive care unit using machine learning. *Nat. Med.* **26**, 364–373 (2020).
- [23] Faltys, M. *et al.* HiRID, a high time-resolution ICU dataset (version 1.1.1). PhysioNet (2021). URL <https://doi.org/10.13026/nkwc-js72>.
- [24] Thorat, P. J. *et al.* Sharing ICU patient data responsibly under the Society of Critical Care Medicine/European Society of Intensive Care Medicine Joint Data Science Collaboration: the Amsterdam University Medical Centers

- Database (AmsterdamUMCdb) example. *Crit. Care Med.* **49**, e563–e577 (2021).
- [25] Elbers, P. W. G. AmsterdamUMCdb v1.0.2. DANS Data Station Life Sciences (2019). URL <https://doi.org/10.17026/dans-22u-f8vd>.
 - [26] Rodemund, N., Wernly, B., Jung, C., Cozowicz, C. & Koköfer, A. Harnessing big data in critical care: exploring a new European dataset. *Sci. Data* **11**, 320 (2024).
 - [27] Rodemund, N. *et al.* Salzburg Intensive Care database (SICdb) (version 1.0.8). PhysioNet (2024). URL <https://doi.org/10.13026/8m72-6j83>.
 - [28] Moukheiber, D. *et al.* Northwestern ICU (NWICU) database (version 0.1.0). PhysioNet (2024). URL <https://doi.org/10.13026/s84w-1829>.
 - [29] Lim, L. *et al.* INSPIRE, a publicly available research dataset for perioperative medicine. *Sci. Data* **11**, 655 (2024).
 - [30] Lim, L. & Lee, H. INSPIRE (version 1.4.2). PhysioNet (2026). URL <https://doi.org/10.13026/1eay-yc85>.
 - [31] Pollard, T. *et al.* PhysioNet as a global platform for biomedical research. *Nat. Health* **1**, 792–795 (2026).
 - [32] Rojas, J. C. *et al.* A common longitudinal intensive care unit data format (CLIF) for critical illness research. *Intensive Care Med.* **51**, 556–569 (2025).
 - [33] Casella, G. & Robert, C. P. Rao-Blackwellisation of sampling schemes. *Biometrika* **83**, 81–94 (1996).
 - [34] Prentice, R. L. A case-cohort design for epidemiologic cohort studies and disease prevention trials. *Biometrika* **73**, 1–11 (1986).
 - [35] Ke, G. *et al.* LightGBM: a highly efficient gradient boosting decision tree. In *Advances in Neural Information Processing Systems 30* (2017).
 - [36] Hollmann, N. *et al.* Accurate predictions on small data with a tabular foundation model. *Nature* **637**, 319–326 (2025).
 - [37] Holm, S. A simple sequentially rejective multiple test procedure. *Scand. J. Stat.* **6**, 65–70 (1979).
 - [38] Zimmerman, J. E., Kramer, A. A., McNair, D. S. & Malila, F. M. Acute Physiology and Chronic Health Evaluation (APACHE) IV: hospital mortality assessment for today’s critically ill patients. *Crit. Care Med.* **34**, 1297–1310 (2006).
 - [39] Moreno, R. P. *et al.* SAPS 3—from evaluation of the patient to evaluation of the intensive care unit. Part 2: development of a prognostic model for hospital mortality at ICU admission. *Intensive Care Med.* **31**, 1345–1355 (2005).
 - [40] Benjamini, Y. & Hochberg, Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Stat. Soc. B* **57**, 289–300 (1995).
 - [41] Rosenbaum, P. R. & Rubin, D. B. The central role of the propensity score in observational studies for causal effects. *Biometrika* **70**, 41–55 (1983).
 - [42] Gneiting, T. & Raftery, A. E. Strictly proper scoring rules, prediction, and estimation. *J. Am. Stat. Assoc.* **102**, 359–378 (2007).
 - [43] Yeom, S., Giacomelli, I., Fredrikson, M. & Jha, S. Privacy risk in machine learning: analyzing the connection to overfitting. In *2018 IEEE 31st Computer Security Foundations Symposium*, 268–282 (2018).
 - [44] Carlini, N. *et al.* Membership inference attacks from first principles. Preprint at <https://arxiv.org/abs/2112.03570> (2021).
 - [45] Carlini, N. *et al.* Extracting training data from large language models. Preprint at <https://arxiv.org/abs/2012.07805> (2020).
 - [46] Geva, M., Bastings, J., Filippova, K. & Globerson, A. Dissecting recall of factual associations in auto-regressive language models. Preprint at <https://arxiv.org/abs/2304.14767> (2023).
 - [47] Abnar, S. & Zuidema, W. Quantifying attention flow in transformers. Preprint at <https://arxiv.org/abs/2005.00928> (2020).
 - [48] Sundararajan, M., Taly, A. & Yan, Q. Axiomatic attribution for deep networks. Preprint at <https://arxiv.org/abs/1703.01365> (2017).

- [49] Schuler, A., Walsh, D., Hall, D., Walsh, J. & Fisher, C. Increasing the efficiency of randomized trial estimates via linear adjustment for a prognostic score. *Int. J. Biostat.* **18**, 329–356 (2021).
- [50] Grechko, A. V. *et al.* RICD: Russian Intensive Care Dataset. *Gen. Reanimatol.* **20**, 22–31 (2024).
- [51] Xu, P. *et al.* Critical care database comprising patients with infection. *Front. Public Health* **10**, 852410 (2022).
- [52] Xu, P., Chen, L. & Zhang, Z. Critical care database comprising patients with infection at Zigong Fourth People's Hospital (version 1.1). PhysioNet (2022). URL <https://doi.org/10.13026/xpt9-z726>.
- [53] Bengio, S., Vinyals, O., Jaitly, N. & Shazeer, N. Scheduled sampling for sequence prediction with recurrent neural networks. Preprint at <https://arxiv.org/abs/1506.03099> (2015).
- [54] Van Calster, B., McLernon, D. J., van Smeden, M., Wynants, L. & Steyerberg, E. W. Calibration: the Achilles heel of predictive analytics. *BMC Med.* **17**, 230 (2019).
- [55] Fisher, C. K., Smith, A. M. & Walsh, J. R. Machine learning for comprehensive forecasting of Alzheimer's Disease progression. *Sci. Rep.* **9**, 13622 (2019).
- [56] Vasey, B. *et al.* Reporting guideline for the early-stage clinical evaluation of decision support systems driven by artificial intelligence: DECIDE-AI. *Nat. Med.* **28**, 924–933 (2022).
- [57] Collins, G. S. *et al.* TRIPOD+AI statement: updated guidance for reporting clinical prediction models that use regression or machine learning methods. *BMJ* **385**, e078378 (2024).
- [58] Bennett, N., Plečko, D., Ukor, I.-F., Meinshausen, N. & Bühlmann, P. ricu: R's interface to intensive care data. *GigaScience* **12**, giad041 (2023).
- [59] Johnson, A. E. W., Stone, D. J., Celi, L. A. & Pollard, T. J. The MIMIC Code Repository: enabling reproducibility in critical care research. *J. Am. Med. Inform. Assoc.* **25**, 32–39 (2018).
- [60] Khwaja, A. KDIGO clinical practice guidelines for acute kidney injury. *Nephron Clin. Pract.* **120**, c179–c184 (2012).
- [61] Rhee, C. *et al.* Incidence and trends of sepsis in US hospitals using clinical vs claims data, 2009–2014. *JAMA* **318**, 1241–1249 (2017).
- [62] Yang, A. *et al.* Qwen3 technical report. Preprint at <https://arxiv.org/abs/2505.09388> (2025).
- [63] Liu, J. *et al.* Muon is scalable for LLM training. Preprint at <https://arxiv.org/abs/2502.16982> (2025).
- [64] Kwon, W. *et al.* Efficient memory management for large language model serving with PagedAttention. Preprint at <https://arxiv.org/abs/2309.06180> (2023).
- [65] Johnson, A. E. W., Kramer, A. A. & Clifford, G. D. A new severity of illness scale using a subset of Acute Physiology And Chronic Health Evaluation data elements shows comparable predictive accuracy. *Crit. Care Med.* **41**, 1711–1718 (2013).
- [66] Subbe, C. P., Kruger, M., Rutherford, P. & Gemmel, L. Validation of a modified Early Warning Score in medical admissions. *QJM* **94**, 521–526 (2001).
- [67] Seymour, C. W. *et al.* Assessment of clinical criteria for sepsis: for the Third International Consensus Definitions for Sepsis and Septic Shock (Sepsis-3). *JAMA* **315**, 762–774 (2016).
- [68] Bone, R. C. *et al.* Definitions for sepsis and organ failure and guidelines for the use of innovative therapies in sepsis. *Chest* **101**, 1644–1655 (1992).
- [69] Kobayashi, G., Kuribayashi, T., Yokoi, S. & Inui, K. Attention is not only a weight: analyzing transformers with vector norms. Preprint at <https://arxiv.org/abs/2004.10102> (2020).
- [70] Chefer, H., Gur, S. & Wolf, L. Transformer interpretability beyond attention visualization. Preprint at <https://arxiv.org/abs/2012.09838> (2020).
- [71] Altman, D. G. & Bland, J. M. How to obtain the P value from a confidence interval. *BMJ* **343**, d2304 (2011).

Figures and legends

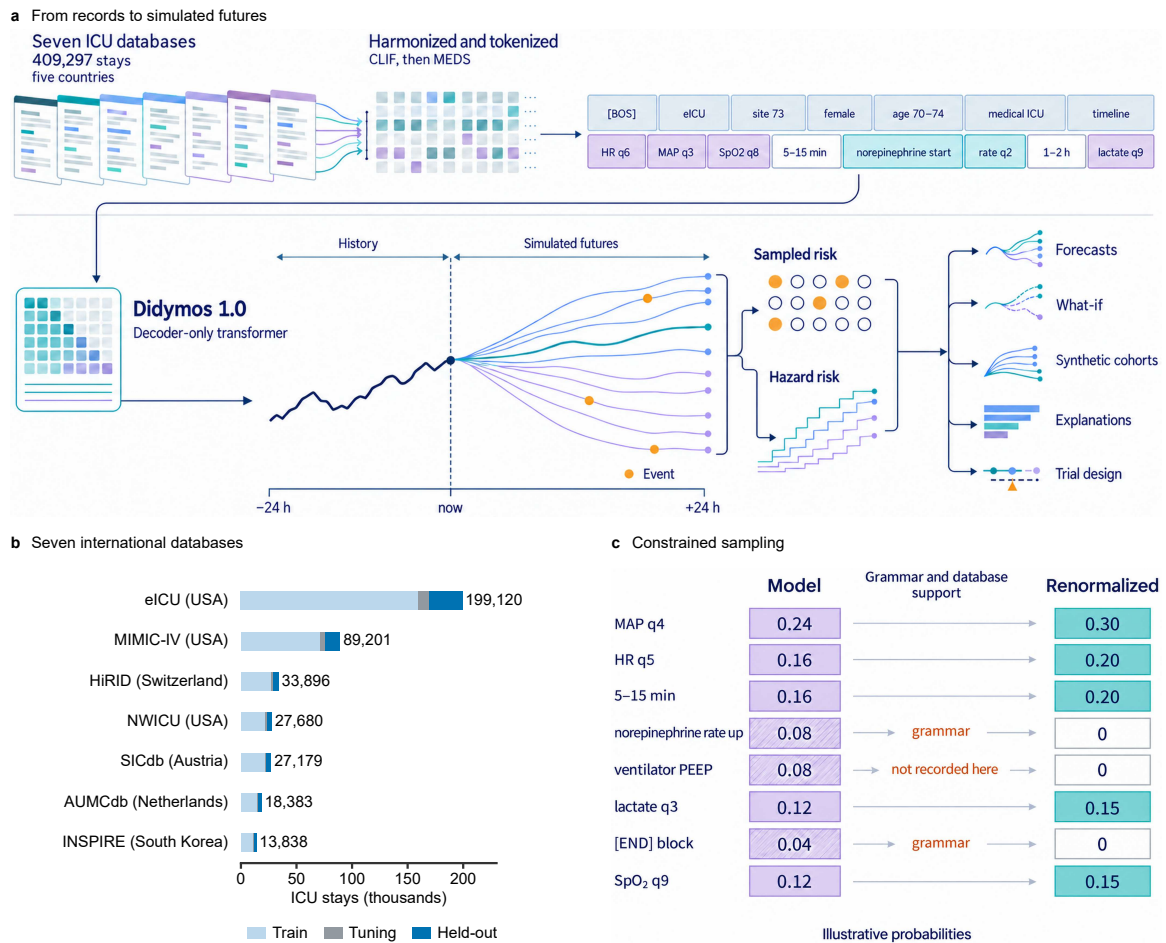


Fig. 1 | Didymos, a generative model of intensive care trained on seven international databases. **a**, From records to simulated futures (GPT-image). Seven intensive care databases were harmonized (CLIF, then MEDS) and tokenized into one sequence per stay, and Didymos, a 12-layer decoder with 123 million parameters, was trained once on them, with the passes over each database bounded by a privacy audit. Conditioned on a record up to a prediction time, Didymos draws simulated futures. The sampled estimate (labelled “sampled risk”) is the share of futures in which an event occurs; the hazard estimate (“hazard risk”) sums the model’s step probability of the event along each future. The token strip shows one stay as one sequence (the database, the patient, then events with value deciles and time intervals, in order); tokens and trajectories are illustrative. **b**, The seven databases, with ICU stays by split (bars: train, tuning and held-out). The held-out sets (2,000 stays per database) gave every prediction time and label reported; they were scored at rolling anchors and at the 24- and 48-hour landmarks, on a catalog of 72 endpoints, against 13 learned models, 7 clinical scores and the databases’ own scores. Data processing, the network and the evaluation sets are shown in Supplementary Figs. 1–3. **c**, Constrained sampling (GPT-image). Before each draw, the grammar removes tokens that would make the sequence invalid (for example, a rate change of an infusion that is not running) and the database’s support removes tokens that database never records; the remaining probabilities are renormalized. Probabilities are illustrative. The grammar accepts all 409,297 real stays. Artwork in a and c is credited (GPT-image); the authors confirm its accuracy and scientific validity and take full responsibility for it. CLIF, Common Longitudinal ICU data Format; MEDS, Medical Event Data Standard; MAP, mean arterial pressure; HR, heart rate; q, value decile.

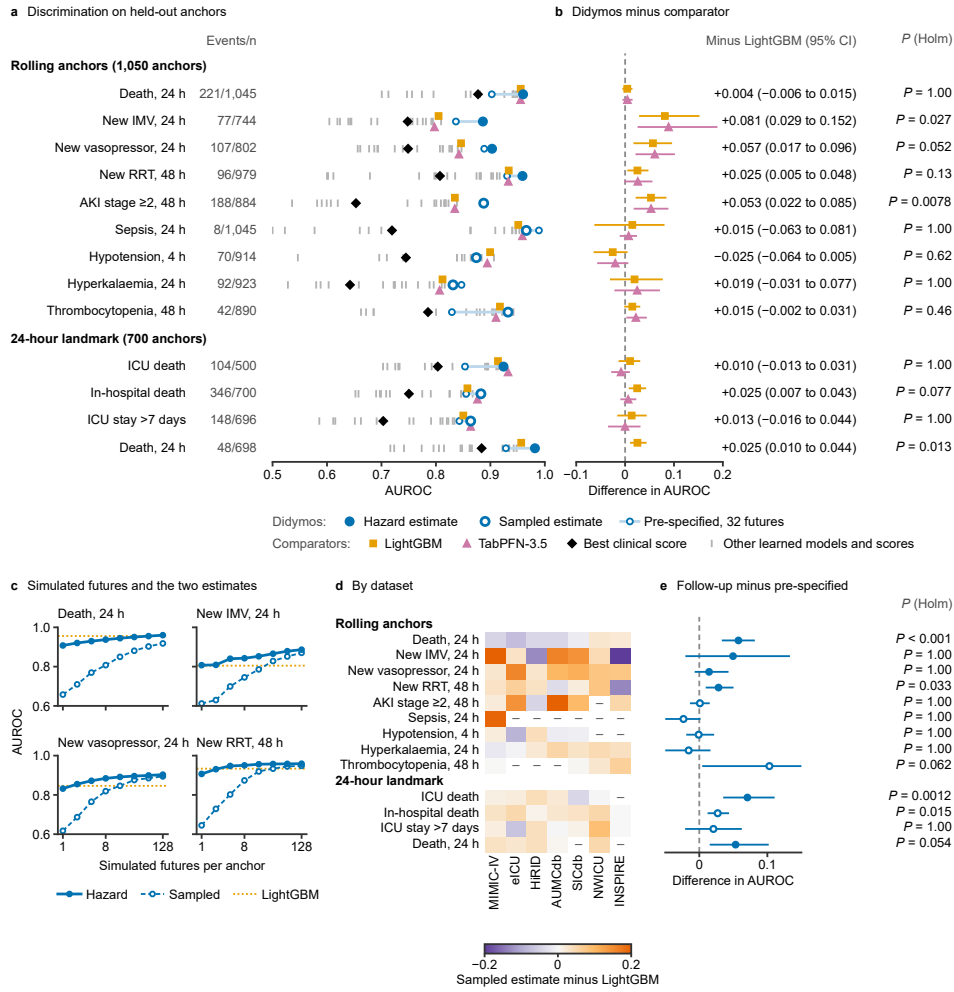


Fig. 2 | Didymos forecasts intensive care outcomes without labels, level with or ahead of supervised models. **a**, AUROC on held-out anchors for 13 primary endpoints: rolling anchors (1,050 prediction times on 1,050 stays, 150 per database) and the 24-hour landmark (a case-cohort sample of 700 stays: 346 in-hospital deaths and 354 weighted controls). Didymos's estimate from 128 simulated futures is shown as filled blue circles (hazard estimate, where it applies) or open blue circles (sampled estimate); its pre-specified value (32 futures, sampled estimate) as small open circles joined to it. LightGBM (orange squares) and TabPFN-3.5 (purple triangles) were trained with labels; black diamonds mark the best clinical score per endpoint, chosen after the fact; grey ticks mark every other learned model and score. Events/n: positive anchors and anchors. **b**, Paired differences, Didymos minus LightGBM (orange) and minus TabPFN-3.5 (purple), with 95% CIs from a paired bootstrap over stays (1,000 resamples). Two-sided P values, from 10,000 paired resamples for the sampled estimate and from the stored interval for the hazard estimate (normal approximation), Holm-adjusted across the 13 endpoints, are shown for LightGBM. TabPFN-3.5 was not fitted for death within 24 h at the landmark. **c**, AUROC against the number of simulated futures per anchor for the hazard estimate (solid) and the sampled estimate (dashed), each the mean over 20 random subsets of each size; LightGBM dotted. **d**, The sampled estimate minus LightGBM per database; dashes mark databases with fewer than 5 events or non-events. **e**, Follow-up minus pre-specified AUROC on the same anchors, with 95% CIs and Holm-adjusted P . Every predictor's AUROC, ranks across endpoints, panel c for every endpoint and AUROC by database are shown in Supplementary Figs. 4–7; every endpoint of the pre-specified analysis and label efficiency in Supplementary Figs. 9 and 10. AUROC, area under the receiver operating characteristic curve; IMV, invasive mechanical ventilation; RRT, renal replacement therapy; AKI, acute kidney injury.

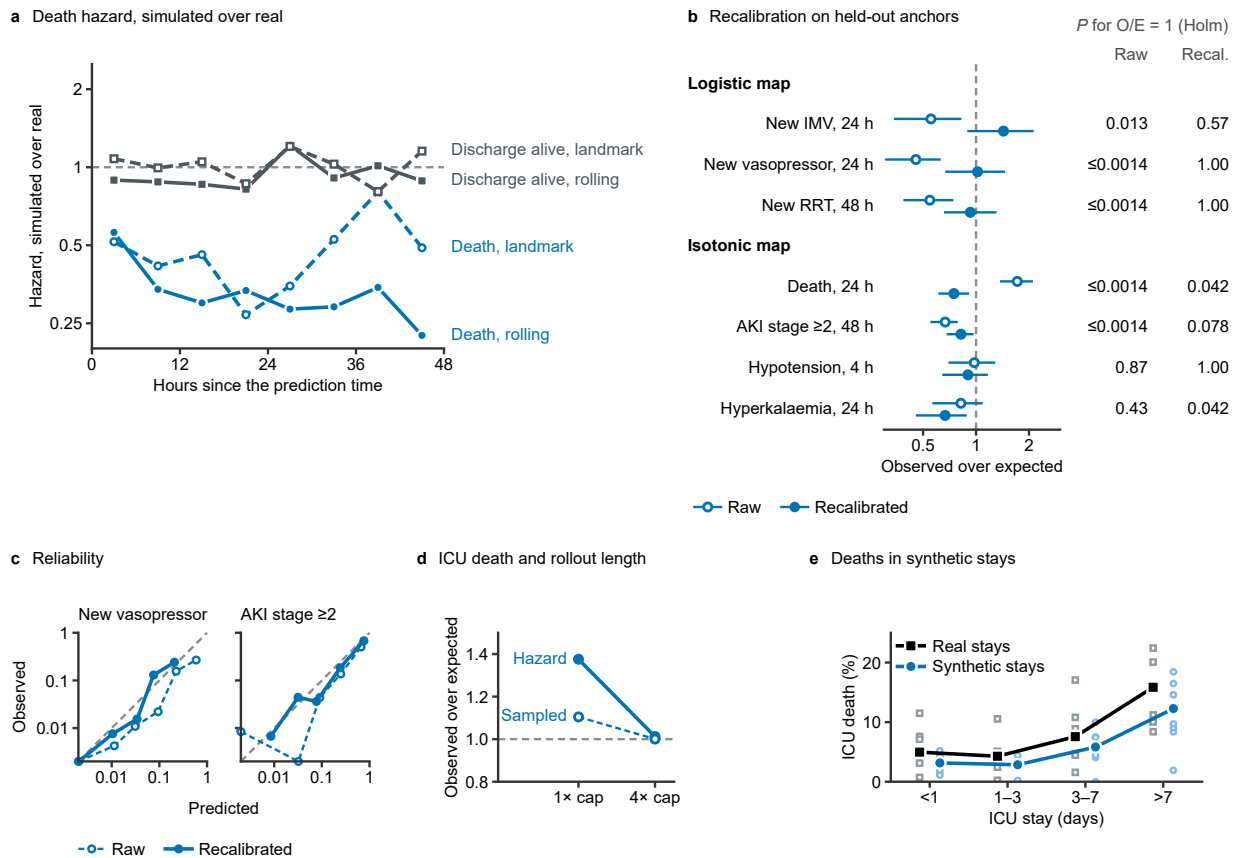
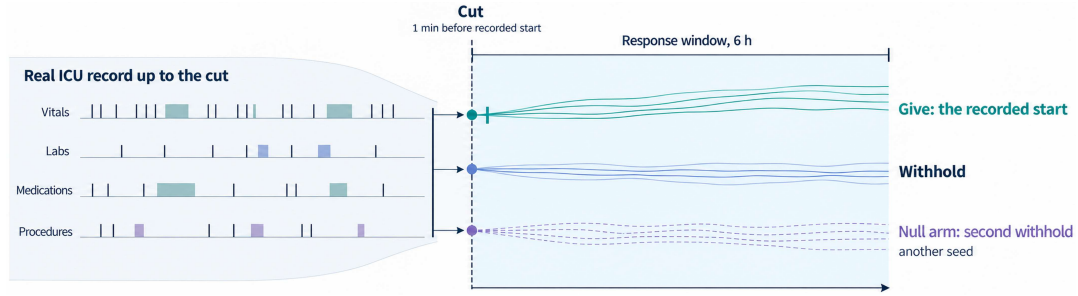
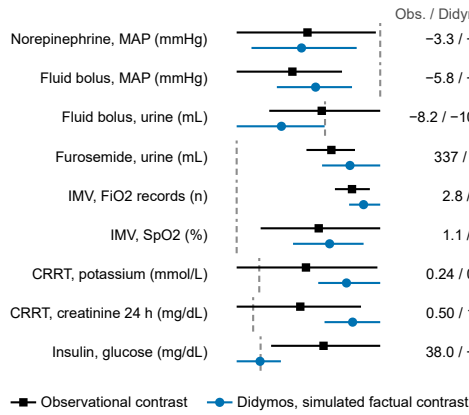


Fig. 3 | Didymos's step hazards of death are calibrated; free generation delays deaths, and correction largely restores calibration. **a**, The hazard of death (blue) and of alive discharge (grey) in the simulated futures, divided by the real futures', in 6-hour blocks after the prediction time (rolling anchors, solid lines and filled markers; 24-hour landmark, dashed lines and open markers). Along the real futures, the model's step hazards are calibrated (O/E 0.98, 95% CI 0.86–1.11, for death within 24 h on rolling anchors; 0.95, 0.82–1.12, for ICU death at the landmark); until ICU discharge, the simulated death hazard is 0.91 of the real one at the landmark. **b**, O/E before (open circles) and after (filled circles) the recalibration maps in force, on 1,050 rolling anchors with 32 futures per anchor. The maps were fitted on the tuning split (logistic or isotonic). The 95% CIs come from a bootstrap over stays (1,000 resamples); *P* tests O/E = 1 with 10,000 resamples, Holm-adjusted across the seven endpoints; ≤ marks a bound (no resample beyond 1). **c**, Reliability of raw (open circles, dashed) and recalibrated (filled circles, solid) risks, in bins of predicted risk with at least 20 anchors. **d**, ICU death at the 24-hour landmark: O/E of both estimates at the pre-specified token cap and at four times it. The longer cap reduced capped futures from 21.6% to 8.5% and kept discrimination (AUROC of the hazard estimate 0.920 to 0.924; of the sampled estimate 0.902 to 0.910). **e**, ICU death by length of stay in the Didymos synthetic cohort (blue circles) and in real held-out stays (black squares); large markers pooled, small markers per database. O/E, observed over expected (above 1, under-prediction).

a Design of the what-if validation

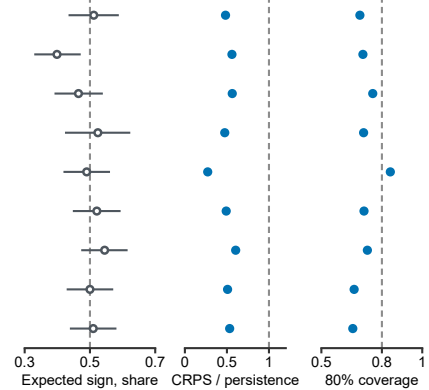


b Contrast, observational and simulated



c Null arm

d Factual accuracy



e Timing of the cut, MAP

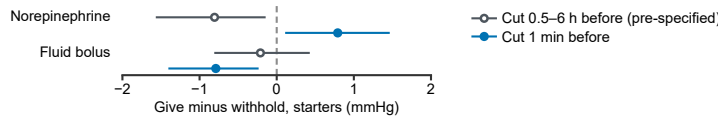


Fig. 4 | Simulated treatment responses reproduce observational contrasts, and a null arm stays at chance. **a**, The design (GPT-image). Starters were cut 1 minute before their first recorded start of a treatment, each matched with a non-starter from the same database at the same ICU hour on the propensity of starting. Didymos drew 16 futures per arm: give (the recorded start) and withhold; a second withhold seed forms the null arm. The response window is 6 hours (creatinine, 24 hours). **b**, For each pre-specified response, the observational contrast, starters' real response minus their matches' (black squares), and Didymos's version, the same pairs' simulated factual futures (blue circles), with 95% CIs from a bootstrap over pairs; each row has its own scale, with 0 dashed, and the numbers are the two means. P tests the difference, combining the two intervals as if independent (conservative), Holm-adjusted across the nine responses; a large P alone does not show equivalence. For insulin, real starters' glucose stayed 38.0 mg/dL above their matches', against -0.3 mg/dL simulated; within Didymos, giving insulin lowered the starters' glucose by 22.6 mg/dL (95% CI 13.4–32.8) against withholding it. **c**, The null arm: the share of prefixes whose two withhold arms differ in the expected direction (ties count against), with exact 95% CIs; every share lies between 0.40 and 0.55, close to the 0.5 of chance (exact binomial tests, Holm $P \geq 0.055$). **d**, Factual accuracy for starters: CRPS relative to persistence (the last value), 0.27–0.60, and coverage of the 80% interval, 0.66–0.84; dashed lines mark 1 and the nominal coverage. **e**, Give minus withhold MAP among starters, with the pre-specified cut (0.5–6 hours before the start; open grey circles) and a cut 1 minute before it (filled blue circles), with 95% CIs. Cut just before the start, giving norepinephrine raises MAP (+0.79 mmHg); cut hours earlier, the fall in MAP that prompted the start dominates (confounding by indication). Artwork in a is credited (GPT-image); the authors confirm its accuracy and scientific validity and take full responsibility for it. CRPS, continuous ranked probability score; MAP, mean arterial pressure; FiO2, fraction of inspired oxygen; SpO2, peripheral oxygen saturation.

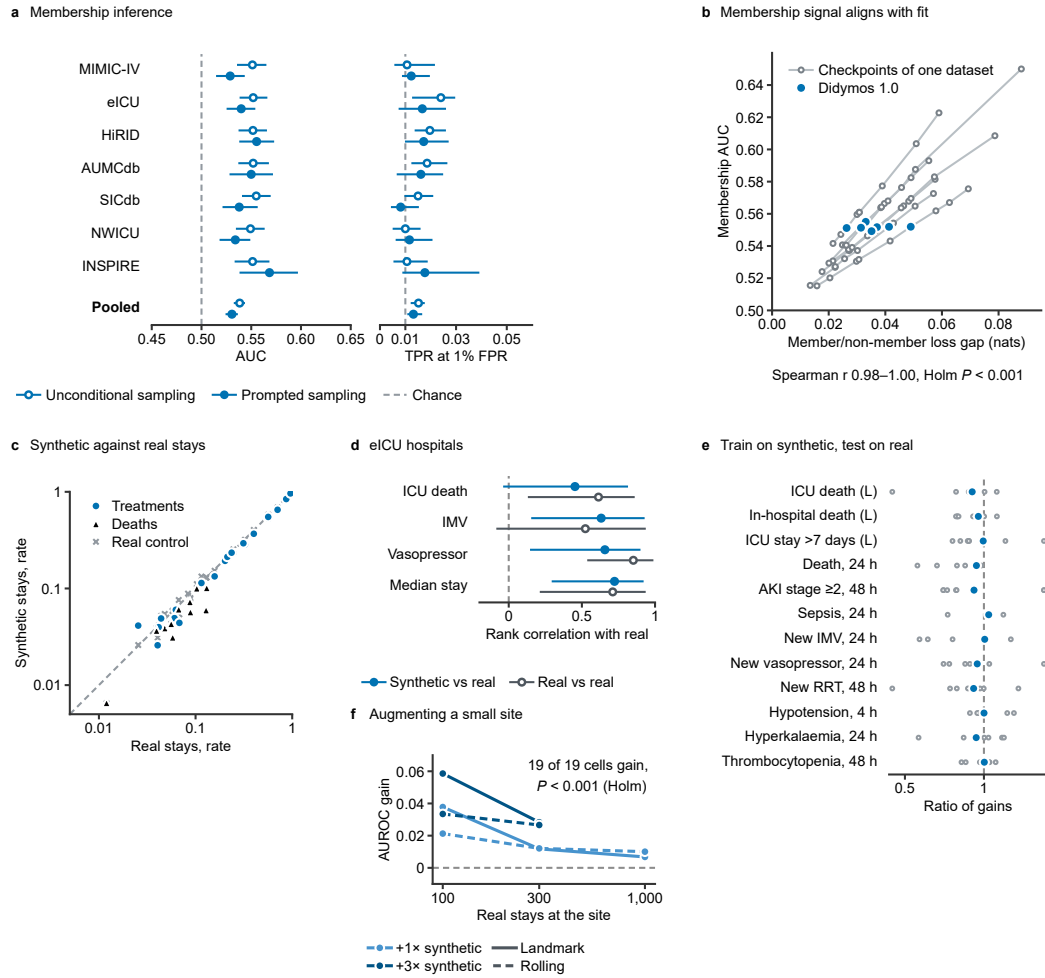


Fig. 5 | Synthetic patients from Didymos are private, faithful and useful. **a**, Membership inference (loss attack) by database and pooled: AUC and true-positive rate (TPR) at a 1% false-positive rate (FPR), with 95% CIs, from the model's loss on whole stays (unconditional sampling; open circles) and on a stay's next 24 hours given its record (prompted sampling; filled circles); dashed lines mark chance (AUC 0.5; TPR 0.01, an attack with no information). In every database, membership AUC is 0.53–0.57 (pooled 0.538 unconditional and 0.531 prompted) and TPR 0.008–0.024. INSPIRE, the smallest database, has the widest intervals. No simulated future of a member or non-member reproduced its real first 32 tokens; a copying generator, which returns members' real futures, reproduces all of them; two halves of the non-members, with no membership to detect, give an AUC of 0.493 (0.483–0.502). **b**, Membership AUC against the member/non-member loss gap across nine checkpoints, per database (grey lines), with Didymos 1.0 in blue; Spearman correlations per database, with exact permutation P values, Holm-adjusted across the seven databases. **c**, Per-database rates in synthetic against real stays (treatments, blue circles; deaths, black triangles); grey crosses are a second real sample (real-against-real control). Between databases, synthetic rates correlate with real ones at 0.73 (95% CI 0.61–0.83; real control 0.95). Death rates sit below the identity line, in keeping with the later simulated deaths of Fig. 3. **d**, Rank correlation of hospital-level rates in 17 eICU-CRD hospitals: synthetic against real (filled circles) and real against real (open circles), with bootstrap 95% CIs. **e**, Train on synthetic, test on real: the ratio of gains (TSTR - 0.5)/(TRTR - 0.5), pooled (blue circles) and per database (small grey circles); (L), 24-hour landmark, otherwise rolling anchors. **f**, AUROC gain from adding synthetic stays to a small site's real stays (mean over cells with at least 20 test positives; 20 repetitions per cell); solid lines, landmark; dashed lines, rolling anchors. With 100 real stays at the landmark, 19 of 19 cells gain (two-sided Wilcoxon signed-rank test, Holm-adjusted). AUC, area under the curve; TSTR, train on synthetic, test on real; TRTR, train on real, test on real.

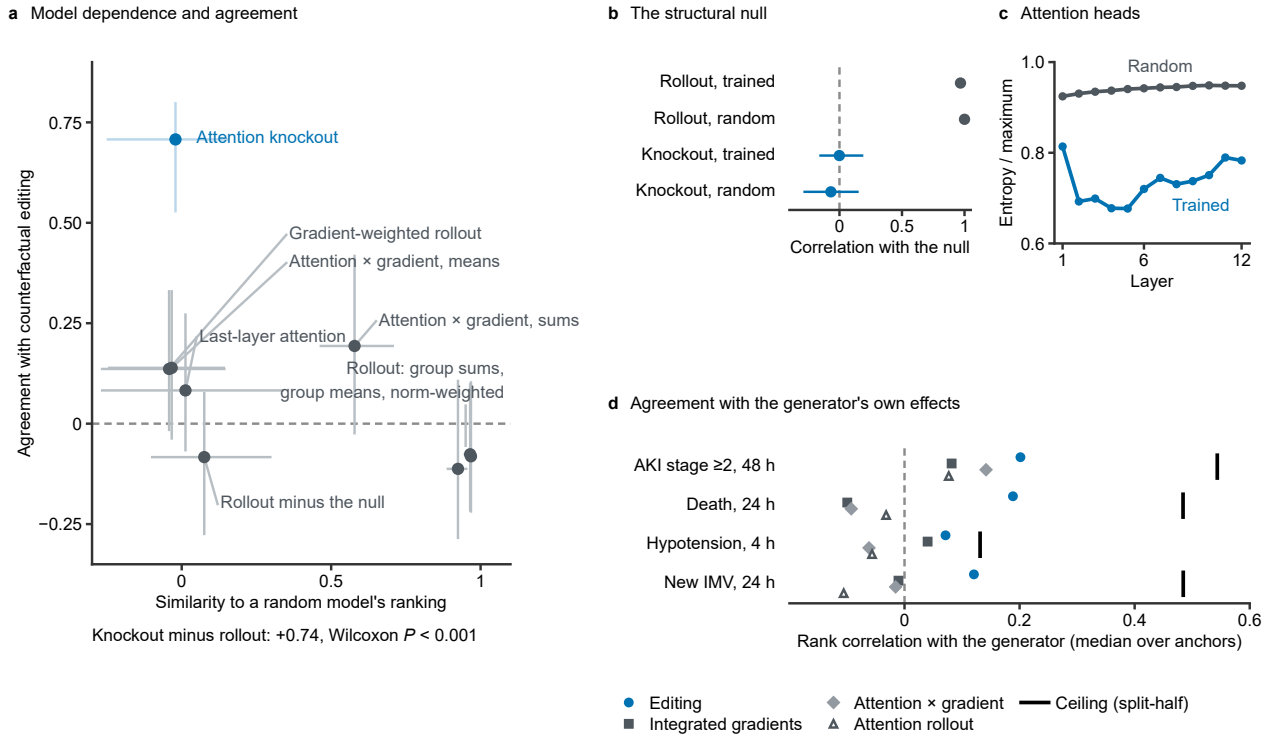


Fig. 6 | Counterfactual editing and attention knockout reflect the trained model; attention rollout mirrors the causal mask. **a**, For each attribution method on 152 anchors: on the horizontal axis, the rank correlation of its group ranking between the trained and a randomly initialized model (near 0, the ranking depends on what the model learned; near 1, training does not change it); on the vertical axis, its rank correlation with counterfactual editing. Points are medians with interquartile ranges over anchors; attention knockout, which the toolkit uses beside editing, is blue. Knockout's ranking is unrelated to the random model's (-0.02) and agrees with editing (0.71); the rollout variants rank groups almost as a random model does (0.96 – 0.97). Knockout's agreement exceeds rollout's (two-sided Wilcoxon signed-rank test over anchors). The pre-specified rule is reported in the Methods. **b**, Rank correlation of rollout and knockout with the structural null, the rollout of uniform causal attention with no model, for the trained and the random model (medians with interquartile ranges). **c**, Normalized entropy of the attention heads by layer, trained and random. In the trained model, the first token holds 35% of the rollout's mass but 0.2% of the last layer's attention. **d**, Agreement with the generator's own effects of editing each group (median over 40 anchors per endpoint): editing (blue circles), integrated gradients (dark grey squares), attention $\times$ gradient (grey diamonds) and attention rollout (open triangles); the split-half reliability ceiling is a black line.

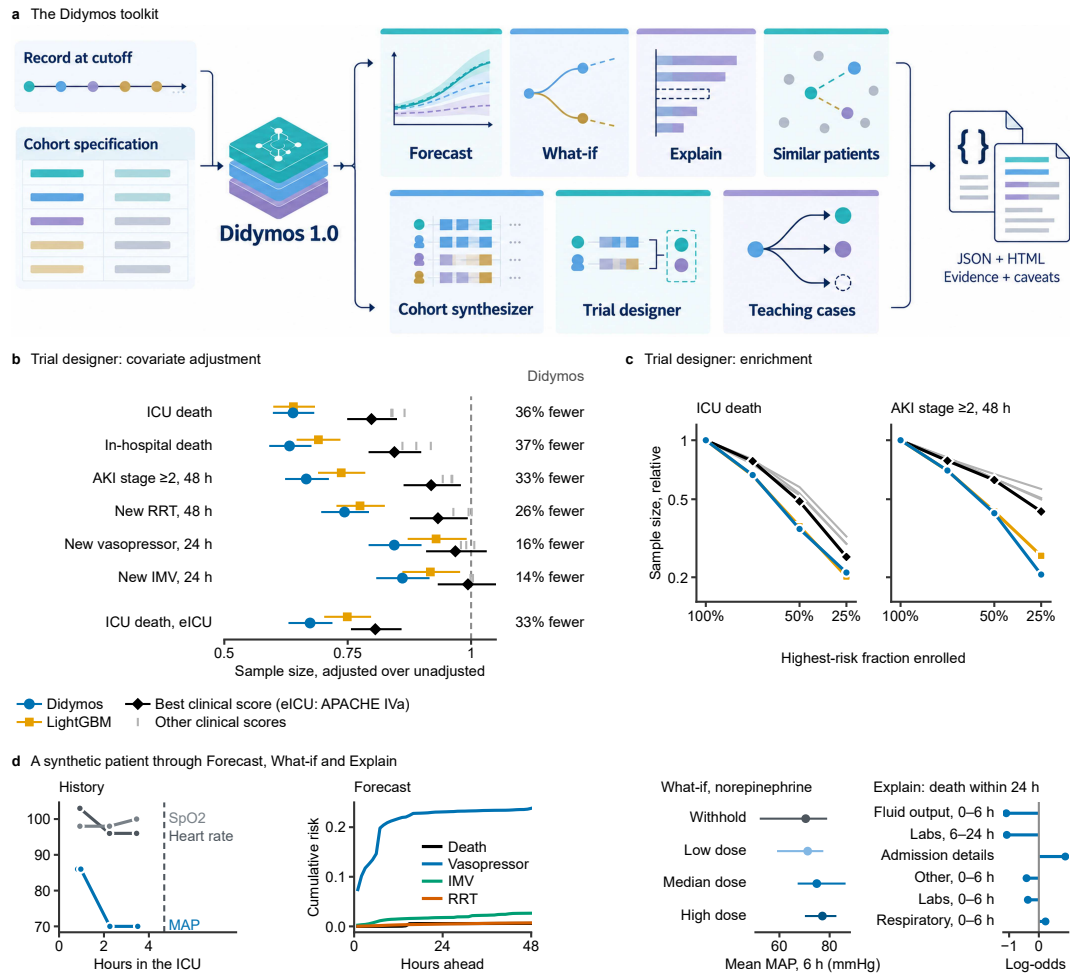


Fig. 7 | The Didymos toolkit, and smaller trials with Didymos prognoses. **a**, The toolkit (GPT-image): seven tools on Didymos's simulated futures, each output carrying the paper's evidence and its caveats. The toolkit's prompts were identical to the paper's (1,050 of 1,050), its risks lay within simulation noise of a paper run (Supplementary Fig. 16), seven recalibration maps are in force, and under prompted sampling membership AUC is 0.53–0.57 in every database (Fig. 5a). **b**, Trial designer, prognostic covariate adjustment: the sample size an adjusted analysis needs relative to an unadjusted one (relative risk 0.8, 80% power, two-sided $\alpha = 0.05$), for Didymos (blue circles), LightGBM (orange squares), the best clinical score (black diamonds; on eICU-CRD, APACHE IVa) and the other scores (grey ticks), with Monte Carlo 95% intervals from 2,000 simulated trials per score. In these trials the adjusted analyses' type I error was 0.0375–0.0715 at a nominal 0.05 (Supplementary Fig. 18). The landmark sample for the two death endpoints is case-cohort, so their sizes are not population values. **c**, Enrichment: the sample size, relative to an unenriched trial, when only the highest-risk fraction by each score is enrolled. **d**, A synthetic patient (a teaching case on shock, prompted from AmsterdamUMCdb), 1 minute before its first norepinephrine start, 4.7 hours into the ICU stay; each tool used 16 simulated futures. History: MAP, heart rate and SpO₂ up to the decision time (dashed line). Forecast: cumulative risk of death, a new vasopressor, IMV and RRT over the next 48 hours (hazard estimate). What-if: mean MAP over the next 6 hours with norepinephrine withheld or started at a low, median or high dose (medians with 10th–90th percentiles of the futures). Explain: the six groups of the record (a channel family in a time block before the decision time, or the admission details) whose removal most changes the Didymos probe's log-odds of death within 24 hours (counterfactual editing); positive groups raise the risk. Artwork in a is credited (GPT-image); the authors confirm its accuracy and scientific validity and take full responsibility for it. MAP, mean arterial pressure; IMV, invasive mechanical ventilation; RRT, renal replacement therapy; APACHE, Acute Physiology and Chronic Health Evaluation.

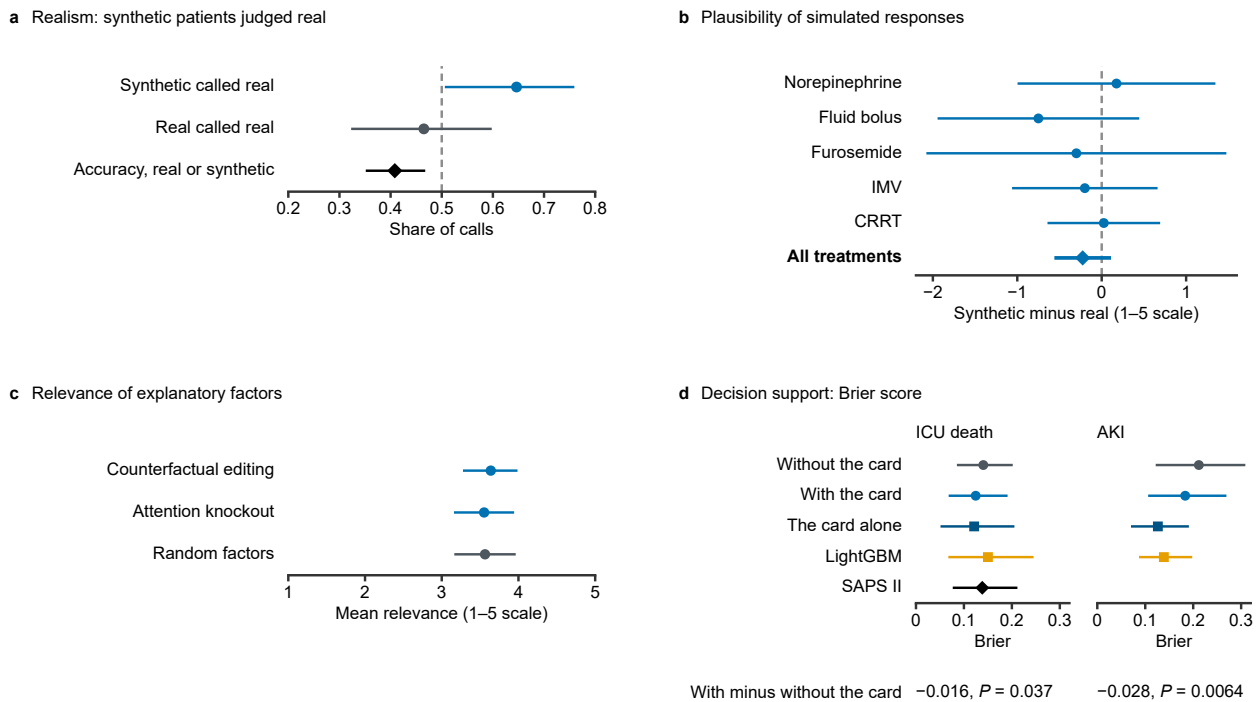


Fig. 8 | Raters did not tell synthetic from real patients, and the forecast card improved their risk estimates. Twenty raters, recruited as intensivists under the study protocol, rated packets of real and synthetic material, and their answers were analysed by the pre-specified analysis, whose decision rules and outcomes are reported in the Methods. **a**, Realism, 80 items per rater (1,600 calls): the share of synthetic and of real stays called real (95% CIs), and the accuracy of telling them apart (mixed model; 90% CI, as pre-specified), against chance (dashed line). **b**, Simulated responses (20 items): plausibility of the synthetic minus the real course on a 1–5 scale, by treatment and for all treatments (mixed model), with 95% CIs; insulin, a separate stratum, is excluded from this analysis, as pre-specified. **c**, Explanations (20 items): mean relevance, on a 1–5 scale, of the factors chosen by counterfactual editing, by attention knockout and at random, with 95% CIs. **d**, Decision support: Brier scores (lower is better) of the raters' risk estimates without and with the forecast card, of the card alone, of LightGBM and of SAPS II, with 95% CIs, for ICU death (48 cases) and AKI (39 cases); below, with minus without the card (mixed model; two-sided P). AKI, acute kidney injury; SAPS, Simplified Acute Physiology Score.

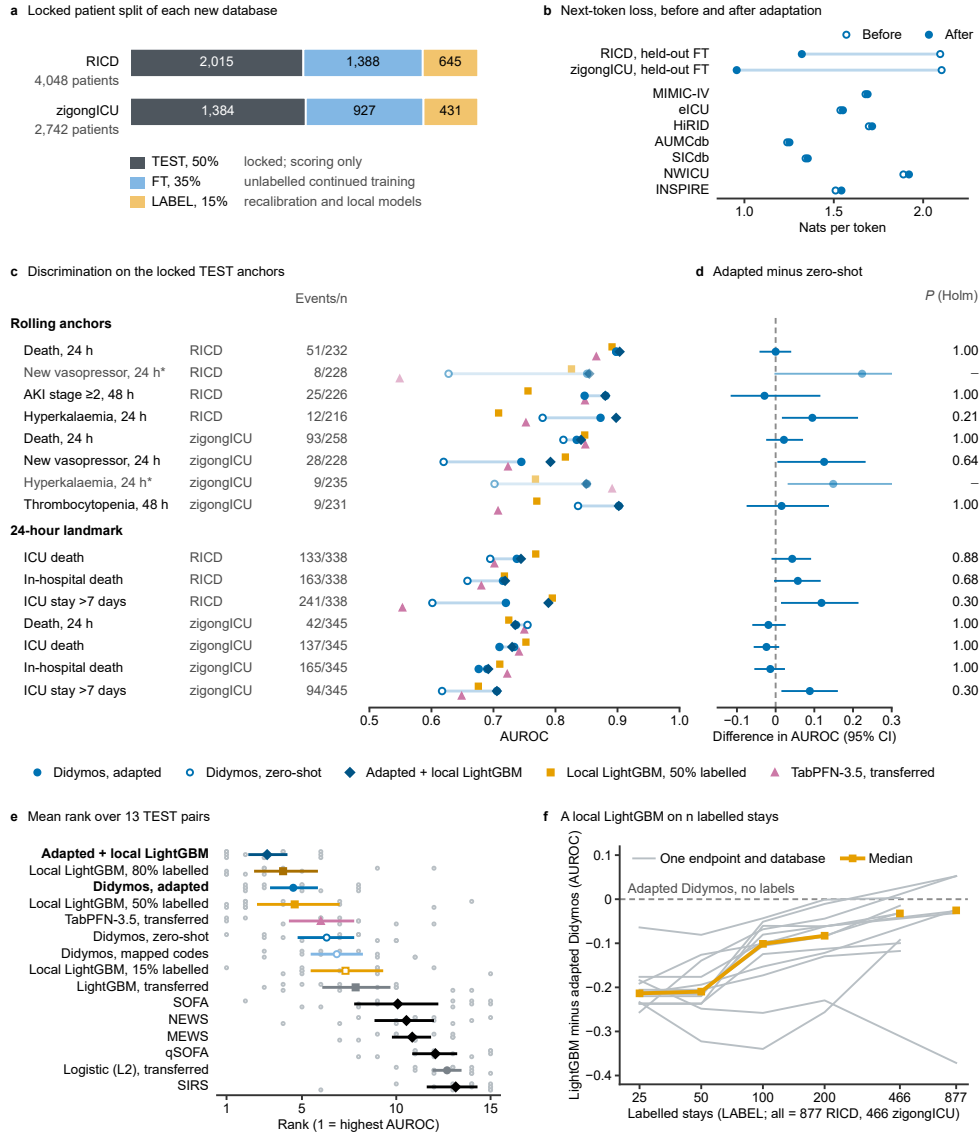


Fig. 9 | Didymos adapts to two hospitals it never saw, without labels. RICD (Moscow, Russia) and zigongICU (Zigong, China) contributed nothing to Didymos’s training. **a**, Each database’s patients were split once, before any adapted model was trained, into a locked test set (TEST, 50%), unlabelled stays for continued training (FT, 35%) and a labelled slice (LABEL, 15%); numbers are patients. **b**, Next-token loss before (open circles) and after (filled circles) adaptation, on held-out FT patients of the new databases and on the tuning split of the seven training databases. **c**, AUROC on the locked test anchors (rolling anchors, 490; a case-cohort sample of the 24-hour landmark, 683; 128 simulated futures): Didymos zero-shot (open blue circles) and adapted (filled blue circles, joined to the zero-shot value), adapted Didymos stacked with a LightGBM trained on the labelled slice (dark blue diamonds), a local LightGBM trained on the labels of the FT and LABEL patients (orange squares) and TabPFN-3.5 trained on the seven databases (purple triangles). Events/n: positive anchors and anchors. Asterisks mark endpoints with fewer than 10 test positives, reported descriptively. **d**, Adapted minus zero-shot AUROC with 95% CIs from a paired bootstrap over stays (1,000 resamples), and two-sided P values from 10,000 paired resamples, Holm-adjusted across the 13 scored pairs. **e**, Mean rank by AUROC over the 13 pairs (1, highest), with 95% intervals from 10,000 bootstrap resamples of the pairs; small circles show the ranks in single pairs. The local LightGBM labelled 80% was trained within each database on all stays and five-fold cross-fitted by patient. **f**, A local LightGBM trained on n labelled stays (median of 20 random draws) minus adapted Didymos recalibrated on the labelled slice, whose recalibration keeps its AUROC; grey lines are single pairs, and the points at the right use every labelled stay (466 in zigongICU, 877 in RICD). Metrics are weighted by the rolling design’s inclusion weights and by case-cohort weights recomputed within the test set. Per-pair calibration and label efficiency are in Supplementary Fig. 21. AUROC, area under the receiver operating characteristic curve; AKI, acute kidney injury.

Tables

Table 1 | The cohort by database. All splits. Ages are 5-year bands decoded at their midpoints, because the databases bin ages differently. HiRID and AmsterdamUMCdb record mortality at ICU discharge only, so in-hospital death equals ICU death there. Treatments are counted from derived episodes in the ICU window: any invasive ventilation record, any vasoactive infusion and any continuous or intermittent RRT record. NWICU's admissions fall mostly (97.9%) in 2020–2022. SD, standard deviation; IQR, interquartile range; IMV, invasive mechanical ventilation; RRT, renal replacement therapy; ED, emergency department.

	MIMIC-IV	eICU-CRD	HiRID	Amsterdam-UMCdb	SICdb	NWICU	INSPIRE
Country	USA	USA	Switzerland	Netherlands	Austria	USA	South Korea
Setting	1 centre	208 hospitals	1 centre	1 centre	1 centre	1 centre	1 centre, post-operative
Admission years	2008–2022	2014–2015	not released	2003–2016	2013–2021	2020–2022	not released
ICU stays	89,201	199,120	33,896	18,383	27,179	27,680	13,838
Train	71,434	159,293	27,194	14,748	21,708	22,178	11,107
Tuning	4,544	9,896	1,691	888	1,373	1,402	659
Held-out	13,223	29,931	5,011	2,747	4,098	4,100	2,072
Age, mean (SD), years	65.2 (16.6)	63.7 (17.0)	66.0 (15.1)	60.8 (15.1)	68.5 (15.2)	63.4 (16.8)	61.8 (15.3)
Female, %	43.5	46.0	35.8	32.3	37.6	43.8	42.6
ICU stay, h, median (IQR)	48.2 (27.1–94.8)	37.6 (19.7–71.8)	24.2 (19.3–53.0)	31.1 (22.3–116.1)	35.2 (19.8–75.8)	46.7 (25.4–93.1)	25.8 (20.6–50.6)
ICU death, %	8.2	5.5	6.1	12.0	3.7	8.0	1.2
In-hospital death, %	12.6	9.0	6.1	12.0	7.8	11.6	5.2
IMV in the ICU, %	41.1	25.7	56.5	86.7	19.9	19.8	78.2
Vasoactive infusion, %	32.0	11.1	35.5	69.8	33.5	22.4	9.4
RRT in the ICU, %	6.5	3.9	5.4	5.8	3.8	4.9	3.0
Timeline events per stay, median	693	471	539	796	583	270	513
Largest admission type	ED, 87.5%	ED, 41.2%	not recorded	elective, 50.4%	elective, 34.3%	ED, 69.7%	elective, 77.3%
Race recorded	yes	yes	no	no	no	yes	no

Table 2 | Didymos ranks first among 20 predictors on the primary endpoints. AUROC on the 12 primary endpoints on which every learned comparator was fitted, on the follow-up anchors: rolling anchors (1,050) and, marked (L), the 24-hour landmark (700, case-cohort weighted); death within 24 hours at the landmark is omitted because TabPFN-3.5 was not fitted there. Didymos is its current estimate from 128 simulated futures: the hazard estimate for death within 24 hours, new IMV, a new vasopressor, new RRT and ICU death, and the sampled estimate otherwise. Learned models were trained with labels on real stays; clinical scores were computed without training. The highest AUROC for each endpoint is in bold. Mean rank: on each endpoint, rank 1 is the highest AUROC and ties take mid-ranks (post hoc analysis; 95% intervals and tests in Supplementary Fig. 5). IMV, invasive mechanical ventilation; RRT, renal replacement therapy; AKI, acute kidney injury.

Predictor	Death, 24 h	New IMV, 24 h	New vasopressor, 24 h	New RRT, 48 h	AKI stage ≥ 2 , 48 h	Sepsis, 24 h	Hypotension, 4 h	Hyperkalaemia, 24 h	Thrombocytopenia, 48 h	ICU death (L)	In-hospital death (L)	ICU stay > 7 days (L)	Mean rank
Didymos	0.960	0.886	0.903	0.959	0.888	0.966	0.874	0.831	0.932	0.924	0.883	0.864	1.83
TabPFN-3.5	0.955	0.797	0.842	0.932	0.835	0.959	0.894	0.807	0.910	0.932	0.876	0.864	3.42
LightGBM	0.956	0.805	0.846	0.932	0.835	0.951	0.900	0.812	0.918	0.914	0.858	0.850	3.67
TabPFN v2	0.957	0.757	0.847	0.912	0.818	0.953	0.884	0.796	0.922	0.911	0.867	0.834	5.00
Random forest	0.949	0.776	0.812	0.917	0.818	0.926	0.881	0.811	0.889	0.914	0.857	0.844	6.75
Didymos probe	0.954	0.788	0.738	0.928	0.816	0.952	0.861	0.746	0.865	0.903	0.867	0.852	7.50
Extra trees	0.947	0.795	0.807	0.882	0.809	0.947	0.907	0.767	0.881	0.894	0.847	0.831	7.83
LightGBM per database	0.942	0.792	0.828	0.905	0.839	0.700	0.853	0.816	0.906	0.896	0.846	0.820	7.92
Logistic (L1)	0.943	0.782	0.840	0.875	0.816	0.946	0.864	0.738	0.942	0.895	0.853	0.791	8.08
Logistic (elastic net)	0.942	0.766	0.830	0.901	0.815	0.940	0.848	0.745	0.942	0.894	0.852	0.823	8.58
Logistic (L2)	0.941	0.754	0.825	0.880	0.823	0.940	0.866	0.722	0.926	0.857	0.825	0.759	9.83
Multilayer perceptron	0.926	0.810	0.813	0.797	0.773	0.963	0.870	0.703	0.875	0.890	0.826	0.808	10.08
Decision tree	0.864	0.643	0.766	0.800	0.797	0.864	0.875	0.725	0.903	0.799	0.775	0.751	12.67
<i>k</i> -nearest neighbours	0.881	0.624	0.739	0.769	0.801	0.909	0.839	0.724	0.856	0.830	0.791	0.812	13.42
Naive Bayes	0.774	0.692	0.724	0.835	0.748	0.863	0.745	0.658	0.800	0.733	0.699	0.701	15.33
NEWS	0.877	0.748	0.749	0.679	0.607	0.660	0.745	0.603	0.685	0.803	0.715	0.651	15.67
SOFA	0.747	0.681	0.661	0.807	0.653	0.523	0.696	0.642	0.785	0.703	0.690	0.703	17.17
MEWS	0.855	0.619	0.687	0.612	0.599	0.719	0.710	0.588	0.662	0.792	0.697	0.615	17.58
qSOFA	0.796	0.604	0.693	0.602	0.582	0.577	0.740	0.528	0.672	0.725	0.652	0.611	18.67
SIRS	0.713	0.639	0.656	0.600	0.536	0.677	0.546	0.580	0.686	0.711	0.658	0.586	19.00

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- [SupplementaryData2statistics.zip](#)
- [SupplementaryData1breadth.csv](#)
- [SupplementaryData3sourcedata.zip](#)
- [SupplementaryData4artwork.zip](#)
- [DidymosSupplementaryInformation.pdf](#)