

Supplementary Information

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).

Contents

Supplementary Notes	4
Supplementary Note 1	Seven databases in one representation 4
Supplementary Note 2	Model development and the choice of Didymos 1.0 5
Supplementary Note 3	Forecasting: estimators, simulated futures and every comparator 6
Supplementary Note 4	Breadth, label efficiency, remaining stay and time to event 7
Supplementary Note 5	Site prompts, transfer between the seven databases, novelty and subgroups 8
Supplementary Note 6	Where the deficit in simulated deaths arises 9
Supplementary Note 7	Calibration of treatment starts and the recalibration maps 10
Supplementary Note 8	What-if simulation: responses, matching, the null arm and factual accuracy 11
Supplementary Note 9	Privacy audits of training and of prompted sampling 12
Supplementary Note 10	Fidelity and utility of the synthetic cohort 13
Supplementary Note 11	Attribution: sanity checks, the failure of attention rollout and agreement with the generator 14
Supplementary Note 12	The toolkit's validation, similar patients and trial simulation 15
Supplementary Note 13	Transfer to two hospitals never seen in training 16
Supplementary Note 14	Adaptation to the new hospitals without labels 18
Supplementary Figures	20
Supplementary Fig. 1	From seven source databases to model-ready token sequences 20
Supplementary Fig. 2	Didymos's network and how it simulates futures 21
Supplementary Fig. 3	Data and evaluation sets 22
Supplementary Fig. 4	Every predictor's AUROC on the same anchors 23
Supplementary Fig. 5	Didymos ranks first across endpoints, among 20 predictors on the primary endpoints and among those scored on all 165 24
Supplementary Fig. 6	Discrimination grows with the number of simulated futures, fastest for the sampled estimate 25
Supplementary Fig. 7	Discrimination by database: Didymos's sampled estimate against LightGBM 26
Supplementary Fig. 8	The 48-hour landmark with the follow-up's settings 27
Supplementary Fig. 9	Didymos against LightGBM on every endpoint the pre-specified analysis scored 28
Supplementary Fig. 10	Didymos's advantage by clinical family, and the labelled stays a single-database model needs to match it 29
Supplementary Fig. 11	Remaining ICU stay and time to event in the pre-specified analysis . . . 30
Supplementary Fig. 12	The site prompt, and supervised transfer to a database left out of training 31
Supplementary Fig. 13	Subgroups in the pre-specified analysis at the 24-hour landmark 32
Supplementary Fig. 14	The recalibration maps in force 33
Supplementary Fig. 15	Attention knockout agrees with the generator's own effects as well as counterfactual editing 34
Supplementary Fig. 16	The toolkit reproduces the paper's forecasts within simulation noise of a paper run 35
Supplementary Fig. 17	Nearest neighbours as a predictor, by the representation used to find them 36
Supplementary Fig. 18	Trial simulations: type I error and sample size 37
Supplementary Fig. 19	Zero-shot transfer to two databases never seen in training 38
Supplementary Fig. 20	Adaptation: mapped codes, the training grid and the membership signal of the adapted stays 39
Supplementary Fig. 21	Calibration after adaptation, and the labelled stays a local model needs . 40
Supplementary Fig. 22	The adapted model's membership test on the seven training databases . 41

Supplementary Tables	42
Supplementary Table 1	Headline results beside their pre-specified values 42
Supplementary Table 2	Didymos beside the closest generative models of health records and intensive care twins 43
Supplementary Table 3	TRIPOD+AI checklist 44
Supplementary Table 4	Candidate checkpoints and the selection decision 46
Supplementary Table 5	The primary endpoints and their definitions 47
Supplementary Table 6	The comparators and their settings 48
Supplementary Table 7	Pre-specified decision rules beside the observed estimates 49
Supplementary Table 8	The code mapping for the two new databases 51
Supplementary Table 9	The <i>P</i> -value approximation check 52
Supplementary Table 10	Terminology and traceability 53
Supplementary Data	54
Supplementary Data 1	Didymos against LightGBM on every scored endpoint 54
Supplementary Data 2	Reporting statistics 54
Supplementary Data 3	Source data for every figure 54
Supplementary Data 4	Schematic artwork and diagram sources 54
Supplementary References	55

Conventions. The Supplementary Notes follow the order of the Results, and the Supplementary Figures follow the main figures they accompany: Supplementary Figs. 1–3 accompany Fig. 1, 4–13 Fig. 2, 14 Fig. 3, 15 Fig. 6, 16–18 Fig. 7 and 19–22 Fig. 9. Every procedure is defined in the main manuscript’s Methods, to which the Supplementary Notes refer by subsection. Unless stated, intervals are 95% intervals from a bootstrap over stays (1,000 resamples) stratified by database and weighted by the designs’ inclusion or case-cohort weights. Panels marked (GPT-image) carry schematic artwork generated with GPT-image, credited “(GPT-image)” wherever it appears; the authors confirm the accuracy and scientific validity of this artwork and assume full responsibility for it.

Supplementary Notes

Supplementary Note 1 | Seven databases in one representation

Question. Can stays recorded under seven different schemas be written in one token language without losing or inventing information, so that one model can learn from all of them and one label engine can read real and simulated futures alike?

Analysis. One pipeline processed every database (Methods, ‘Harmonization and preprocessing’ and ‘Representation and grammar’; Supplementary Fig. 1). Archives were mirrored at row-count parity and converted to CLIF tables,¹ through the official pipelines for MIMIC-IV and AmsterdamUMCdb and through mapping-table adapters for the other five, in which every row records the definition it came from. The tables were checked against the CLIF schema and vocabularies, compared across databases and reconciled event by event with each database’s community MEDS pipeline, with a documented reason required for every unpaired row. Harmonization then applied one rule per topic to all databases. Monitor streams kept one value per stay, category and clock hour: a single measurement kept its time, and several became their median, stamped at the hour’s last measurement. Continuous medications kept one row per rate change, and doses were converted only with weights charted at or before them. Header fields documented after admission were hidden from prediction inputs until their documentation window had passed: 24 hours for the diagnosis group and surgical status in eICU-CRD, HiRID, AmsterdamUMCdb and SICdb, and 1 hour in INSPIRE.

Results. The databases differ by design (Table 1). AmsterdamUMCdb and INSPIRE are dominated by post-operative ventilated patients (invasive ventilation in 86.7% and 78.2% of stays), eICU-CRD by general units with low treatment intensity (25.7% ventilated, 11.1% with a vasoactive infusion) and NWICU by the COVID-19 era (97.9% of admissions in 2020–2022). ICU death ranged from 1.2% (INSPIRE) to 12.0% (AmsterdamUMCdb), and the median number of timeline events per stay from 270 (NWICU) to 796 (AmsterdamUMCdb).

Every check of the pipeline came out clean at full scale. The CLIF tables had no hard validation error in any database, and the conversion to MEDS reproduced the reference converter with no differing event in any database. No dose used a weight charted after it. KDIGO stages agreed with the mimic-code logic² on every column, and ventilation and norepinephrine-equivalent episodes were identical to mimic-code’s. Every event that our pipeline and the community pipelines did not pair had a documented reason, except two minor defects of the training snapshot that have since been fixed (four INSPIRE records, and HiRID infusion rates rounded to six decimals).

The vocabulary, fitted on the train split, has 25,601 tokens. Of its 19,774 event tokens over 4,434 codes, 12,950 are value tokens (deciles of each code’s training values across all databases, or one bin per value for codes with at most 10 distinct values), 2,390 mark codes recorded without a value, 2,390 mark non-numeric results and 2,011 are flags; 29 stand for codes seen in fewer than 10 training patients, one per domain, and 4 are code-system tokens. A further 3,057 tokens are sub-tokens of diagnosis and procedure codes, and 1,792 summarize the state of a stay that continues into a new window (running infusions with their rate bin, the last respiratory device and mode, and continuous RRT or mechanical circulatory support within 24 hours). The remaining 978 form the prefix, header, goal, clock, interval, end and special blocks. Medication orders (1,339 codes), intermittent (979) and continuous medications (765) and microbiology (769) account for most event codes, and vital signs need only 14. Values decode from five-point sketches (minimum, quartiles and maximum) of the training values within each bin, per code and database. Quantities derived from several values, such as the total Glasgow Coma Scale, the ratio of arterial to inspired oxygen, SOFA and norepinephrine

equivalents, are computed after decoding rather than generated, so they cannot contradict their components.

The grammar's automaton and an independent parser accepted all 409,297 stays. Every stay survived the round trip from MEDS to tokens, to a decoded stay and back to identical tokens. No prediction input cut at 0, 1, 6 or 24 hours contained an outcome or a later-documented header field, and none changed when a later event changed. In each pre-specified anchor set, every anchor's real stay, decoded on its own clock as if it were a simulated future, reproduced the real store's label for every endpoint defined on a decile edge (agreement 1.000). Endpoints whose clinical cut-off falls inside a value bin were therefore defined at the nearest decile edge (Supplementary Table 5).

Conclusion. One token language held the seven schemas without rejections or losses, and labels read from it were exact for real and simulated futures alike. The heterogeneity that remains is genuine practice and case mix, which the database token conditions on (Supplementary Note 5).

Supplementary Note 2 | Model development and the choice of Didymos 1.0

Question. How was the final model chosen, and what did the candidates trained on the way show about model size, optimizer and the membership signal?

Analysis. Ten candidate checkpoints were trained during development (Supplementary Table 4). Each was judged on the tuning loss, the membership-inference audit of training (Methods, 'Privacy audits'; Supplementary Note 9) and, in the final round, audits of synthetic stays, by rules written before the candidate was measured (Methods, 'Model, training and selection of the final checkpoint'). Eight checkpoints were then compared on their representations, by the Didymos probe's AUROC on three endpoints, and on step calibration, by the teacher-forced observed-to-expected ratio of treatment starts after the first ICU day (the model's next-event probabilities summed over every eligible position of real held-out stays, against the observed onsets).

Results. Didymos 1.0 has 123,491,328 parameters, 19.7 million of them in the token embedding, which is shared with the output head. Its context of 32,768 tokens holds 99.92% of training stays in a single window and leaves a margin of 1,024 positions for the goal and clock blocks. Training packed windows without padding, each attending causally only to itself, and ran in bf16 with float32 master weights at 350,800 tokens per second. The final tuning loss of 1.5525 nats per token was lowest in the database with the densest recording (AmsterdamUMCdb, 1.239) and highest in the sparsest (NWICU, 1.889).

The candidates showed three regularities. First, the membership signal followed the fit that each database reached, whatever the optimizer. Across nine audited checkpoints trained with AdamW or Muon, membership AUC in every database rose with the gap between members' and non-members' loss (Spearman correlation 0.98–1.00) and rose as the held-out loss fell (correlation with the held-out loss –0.85 to –0.97), on one curve for both optimizers (Fig. 5b). Muon reached a given loss in fewer steps and so reached the bound on the membership signal sooner: at learning rates of 6×10^{-4} and 1.2×10^{-3} , the tuning loss fell to 1.527 and 1.509, while MIMIC-IV's membership AUC rose to 0.604 and 0.623. The adopted checkpoint therefore used Muon, with the passes over each database set from slopes measured on earlier checkpoints to a projected AUC of 0.555; the audit then measured 0.549–0.555. A candidate projected to 0.565 reached a lower tuning loss (1.5388), but the upper bound of its true-positive rate at a 1% false-positive rate was 0.031, against the pre-specified 0.03. Second, capacity did not help at a fixed privacy budget. A model 2.2 times larger, trained for the same passes, matched the reference's tuning loss (1.563 against 1.561), whereas its membership AUC rose by 0.010–0.016, gains that accrued to members only. Third, design changes aimed at timing barely moved the over-prediction of treatment starts. With ICU-day clock blocks stating each treatment's status, the

teacher-forced ratio after the first ICU day was 0.72 for IMV and 0.41 for RRT, against 0.68 and 0.39 without them, and input-noise augmentation raised the tuning loss to 1.607.

The representations were robust to these choices. Across the eight checkpoints, the probe's AUROC ranged from 0.897 to 0.913 for ICU death at the 24-hour landmark, from 0.807 to 0.837 for new IMV and from 0.921 to 0.938 for new RRT. Didymos 1.0's teacher-forced ratios were 0.97 for ICU death, 0.76 for IMV, 0.59 for vasopressors and 0.44 for RRT. In the final round, it met 30.5 of 35 authenticity criteria against 25.0 for the reference (the mean of two audits, each comparing 1,000 synthetic stays per database with 1,000 held-out stays), and 21 calibration cells against the reference's 22, within the one cell the adoption rule allowed. A discriminator on the first day of records separated its synthetic stays from real ones with AUCs below 0.70 in every database except INSPIRE (0.75).

Conclusion. Didymos 1.0 sits on a frontier between fit and membership signal that was measured rather than assumed: limiting the passes over each database bounded its membership signal before training, and neither a larger model nor a different optimizer moved the frontier. Its representations were stable across candidates, and the over-prediction of treatment starts was a property of all of them (Supplementary Note 7).

Supplementary Note 3 | Forecasting: estimators, simulated futures and every comparator

Question. How do the number of simulated futures, the estimator and the length of each future shape Didymos's forecasts, and how do the forecasts compare with every supervised model, clinical score and bedside value on the same anchors?

Analysis. The pre-specified analysis drew 32 futures for each of 3,500 anchors per design and read the sampled estimate (Methods, 'Simulation and the two estimates'). The follow-up drew 128 futures on 1,050 rolling anchors and on 700 stays of the 24-hour landmark sampled as a case cohort,³ with four times the token cap at the landmark, and compared three estimators on the same futures. The first is the sampled estimate (E0). The second is an Aalen–Johansen estimate over the sampled events, with death and alive discharge competing and the cap censoring (E1). The third is the hazard estimate (E2), which replaces each future's sampled events with the model's step probabilities.⁴ Curves of AUROC against the number of futures averaged 20 random subsets of each size. A further amendment repeated the 48-hour landmark with the follow-up's settings on 809 anchors. Comparators (Supplementary Table 6) were scored on the same anchors.

Results. In the pre-specified analysis, Didymos's AUROC on the rolling anchors ranged from 0.804 for hyperkalaemia to 0.941 for sepsis (28 events). It was ahead of LightGBM for a new vasopressor (+0.056; 95% CI 0.028 to 0.086), AKI (+0.037; 0.020 to 0.055) and sepsis (+0.117; 0.008 to 0.219), and behind for death within 24 hours (−0.062; −0.079 to −0.048). At the landmarks it trailed LightGBM on ICU death (24 hours: 0.858 against 0.915; 48 hours: 0.882 against 0.903), in-hospital death and a remaining stay of more than 7 days. Across all binary endpoints of the catalog, its median AUROC was 0.888 at the 24-hour landmark (57 endpoints), 0.885 at the 48-hour landmark (56) and 0.882 on rolling anchors (53). The pre-specified token cap stopped 21.6%, 29.1% and 36.1% of futures in these three designs.

With 128 futures, AUROC was still rising at 128 for death within 24 hours (0.902 at 64 and 0.917 at 128), thrombocytopenia, sepsis and hyperkalaemia, and had levelled off for AKI and hypotension (Supplementary Fig. 6). The Aalen–Johansen estimate matched the sampled estimate's AUROC to the third decimal on most endpoints, because both count sampled events. The hazard estimate improved on both for death within 24 hours on rolling anchors (+0.042 over the sampled estimate; 0.025 to 0.062), new IMV (+0.016; 0.000 to 0.033), a new vasopressor (+0.009; 0.002 to 0.019), new RRT (+0.011; 0.001 to 0.023) and ICU death at the landmark (+0.014; 0.005 to 0.025). Its advantage was largest with few futures: from one future, AUROC was

0.81–0.92 on the six endpoints where it applies, against 0.61–0.66 for the sampled estimate, and for death within 24 hours eight hazard-read futures (0.937) exceeded 128 sampled ones (0.918). The hazards were exact: renormalized over the tokens the sampler could draw, they reproduced the sampler’s masked probabilities within 0.975–1.006 in every database, and summed along the futures they matched the sampled events (ratios 0.993–1.022). A first scoring, corrected while the follow-up ran, had omitted the renormalization. It changed no AUROC by more than 0.002, but in AmsterdamUMCdb the exact death probabilities were 1.24 times those first extracted, because the model places mass on infusion changes that the grammar forbids at the end of life.

Every comparator was scored on the same anchors (Supplementary Fig. 4). Among the learned models, TabPFN-3.5, LightGBM, random forest and extra trees usually led, followed by L1 and elastic-net logistic regression and the multilayer perceptron, then L2 logistic regression, k -nearest neighbours, the decision tree and naive Bayes. Where the hazard estimate applies, Didymos exceeded every other learned model and clinical score, except TabPFN-3.5 for ICU death at the landmark (0.932 against 0.924). On the other endpoints, no comparator chosen after the fact as the best of its kind was ahead of the sampled estimate with an interval excluding zero; the closest were extra trees for hypotension (0.907 against 0.874) and elastic-net logistic regression for thrombocytopenia (0.942 against 0.932). The Didymos probe was among the best models for death (0.974 for death within 24 hours at the landmark) but trailed on treatment starts (0.738 for a new vasopressor on rolling anchors).

Clinical scores trailed Didymos’s sampled estimate on every primary endpoint, by 0.040 for death within 24 hours on rolling anchors (NEWS) to 0.247 for sepsis (MEWS; Holm $P \leq 0.0026$ on ten of 13). Their inputs were missing at some sites: SICdb and NWICU record no Glasgow Coma Scale, NWICU records no urine output or blood gases, and only 71% of NWICU’s landmark anchors had vital signs in the 24 hours before the landmark. The published SAPS II and OASIS equations over-predicted death in these cohorts (O/E 0.24 and 0.39 for ICU death at the 24-hour landmark). The databases’ own admission scores fell below Didymos on their own data: APACHE IVa⁵ for in-hospital death in eICU-CRD (24 hours: 0.850 against 0.941, +0.091, 0.032 to 0.163; 48 hours: 0.766 against 0.901) and SAPS 3⁶ in SICdb (0.812 against 0.869, and 0.815 against 0.882). The one exception was APACHE IVa’s predicted ICU stay for a remaining stay of more than 7 days (0.763 against 0.704; −0.059, −0.165 to 0.048). A single bedside value came close only where the endpoint is a threshold on that value. On the pre-specified anchors, platelets gave 0.919 for thrombocytopenia against Didymos’s 0.939, potassium 0.793 for hyperkalaemia against 0.804, and MAP 0.863 for hypotension against 0.906.

At the 48-hour landmark with the follow-up’s settings, 11.2% of futures reached the cap, against 29.1% before (Supplementary Fig. 8). Didymos was level with LightGBM and TabPFN-3.5 for ICU death (hazard estimate 0.910; LightGBM 0.915; TabPFN-3.5 0.925) and in-hospital death (0.861; 0.861; 0.864), and ahead for a remaining stay of more than 7 days (0.860; 0.837; 0.841) and death within 24 hours (0.948; 0.915; 0.943). For the treatment starts at this landmark, a secondary analysis, its AUROC was 0.847 for new IMV, 0.938 for a new vasopressor and 0.946 for new RRT, against 0.683, 0.837 and 0.954 for LightGBM.

Conclusion. The forecasts are limited by Monte Carlo noise and by simulation length, not by what the model knows. Reading the step hazards along each future removes most of the noise, longer futures remove the censoring of long stays, and with both Didymos was level with or ahead of the strongest supervised models on every primary endpoint. Clinical scores and the databases’ own admission scores fell well below it, and the probe’s gap shows that the simulation, not only the representation, carries the advantage.

Supplementary Note 4 | Breadth, label efficiency, remaining stay and time to event

Question. Does the advantage of the primary endpoints extend to every endpoint the catalog defines, how

many labelled stays does a single-database model need to match Didymos, and how good are its distributional forecasts?

Analysis. The breadth analysis compared Didymos's pre-specified sampled estimate (32 futures) with LightGBM on all 165 endpoints that the pre-specified analysis scored: 53 on rolling anchors and 56 at each landmark (Methods, 'Statistics and reproducibility'). The label-efficiency analysis, pre-specified, trained LightGBM within each database on 100, 300, 1,000, 3,000 and all sampled training stays (up to 5,000) and found the number at which it first reached Didymos's AUROC in that database, by interpolation on $\log n$. Regression and time-to-event endpoints were read from the ensemble of 32 futures.

Results. Didymos was ahead of LightGBM on 113 of 165 comparisons (68%; joint-bootstrap 95% CI 62–73%; Supplementary Fig. 9), with a median difference of +0.014 AUROC (0.010 to 0.019). The advantage held in most families (Supplementary Fig. 10a): organ failure (ahead on 37 of 51; 22 at a 5% false-discovery rate), sepsis (8 of 9; 4), intervention onsets (29 of 33; 9), discharge (6 of 6; 5) and threshold crossings (33 of 57; 9). It was absent for mortality (0 of 7) and the remaining stay (0 of 2), the families whose primary endpoints the follow-up's longer simulations brought to parity (Supplementary Note 3). Of the six comparisons favouring LightGBM at a 5% false-discovery rate, five were mortality or remaining-stay endpoints and one was acidaemia within 12 hours (−0.055). Among the four predictors scored on every endpoint, Didymos had a mean rank of 1.58 (1.45 to 1.72) and ranked first on 102, against 2.00 for LightGBM (Supplementary Fig. 5b).

Label efficiency depended on the endpoint (Supplementary Fig. 10b). For AKI, a LightGBM trained within one database did not reach Didymos in six of seven databases, even with all sampled stays, and the same held for new RRT and a new vasopressor in five of seven and for new IMV in four. For death within 24 hours, LightGBM caught up with at most 1,000 stays in five of seven databases, and for hypotension with 100 in two. Over the 78 endpoint–database cells, more than 1,000 labelled stays were needed in 12 and Didymos was never reached in 38.

The remaining ICU stay at the 24-hour landmark had a mean absolute error of 41.7 hours for the median of the simulated remaining stays, against 58.6 hours for predicting the mean (48-hour landmark: 54.8 against 73.4 hours; Supplementary Fig. 11). The simulated distribution was too narrow: its central 80% interval covered 70% of observed stays, and the top decile of the probability integral transform held 538 of 2,942 stays where about 294 were expected, so the longest observed stays most often exceeded their simulated range. Futures stopped by the pre-specified token cap, the longest ones, were left out of these distributions, which accounts for at least part of this shortfall (Supplementary Note 6). Short-horizon values were forecast with 35–58% less error than predicting the mean: creatinine at 24 hours (0.43 against 0.92 mg/dL), the highest lactate over 12 hours (1.00 against 1.85 mmol/L), the lowest MAP over 6 hours (8.5 against 13.0 mmHg) and the heart rate in the next hour (6.5 against 15.4 beats per minute). Harrell's concordance index⁷ for the time to death was 0.854 at the 24-hour landmark, 0.790 at the 48-hour landmark and 0.834 on rolling anchors, and for the time to a new vasopressor 0.827, 0.809 and 0.746.

Conclusion. The label-free advantage is broad: it holds in most clinical families, and a model trained on one database's labels rarely replaces it for organ support and kidney injury. Where it was absent in the pre-specified configuration, longer simulations closed the gap. Distributional forecasts are informative but too narrow in their upper tail, where the pre-specified cap removed the longest simulations.

Supplementary Note 5 | Site prompts, transfer between the seven databases, novelty and subgroups

Question. How much do the forecasts depend on the database and site tokens, how do supervised models fare

when moved between databases, and do the forecasts hold for atypical patients and across subgroups?

Analysis. On the follow-up's rolling anchors, 32 futures per prompt were drawn with the recorded database and site, with both set to unknown, as in 10% of training windows, and with the database token swapped for another database's (Methods, 'Similarity and site prompts'). For supervised transfer, LightGBM and the probe were trained on six databases and tested on the seventh, against training within it and on all seven. An anchor's novelty was the cosine distance from its state to the nearest training anchor's state, grouped in fifths. Pre-specified subgroups at the 24-hour landmark were scored where they had at least 30 events and 30 non-events.

Results. With the database and site unknown, AUROC changed by less than 0.04 on six of nine rolling endpoints (Supplementary Fig. 12a). It rose for hypotension (+0.056; 0.008 to 0.112) and fell for thrombocytopenia (-0.084; -0.204 to 0.003) and sepsis (8 events). In the pre-specified analysis at the 24-hour landmark, prompts with the site unknown gave AUROCs of 0.85–0.88 against 0.81–0.85 with the site recorded. Swapping in another database's token lowered AUROC for new RRT (-0.298; -0.392 to -0.180), a new vasopressor (-0.123; -0.250 to -0.021) and AKI (-0.052; -0.106 to -0.005), and moved the death O/E from 1.87 to 0.56. The predicted risks did not move towards the other database's prevalence (correlations of 0.25 or less), so the swap acted through the support mask rather than through the database's case mix: a prompt labelled as another database can only generate the events that database records. Of the 60 RRT tokens, for example, MIMIC-IV's support allows 44, whereas that of HiRID, INSPIRE and NWICU allows one.

Supervised models lost most on endpoints shaped by practice when moved between databases (Supplementary Fig. 12b). Averaged over the seven databases, LightGBM trained on the other six lost 0.108 AUROC for new IMV, 0.070 for AKI and 0.035 for a new vasopressor against LightGBM trained within the database, and lost nothing for hypotension or death within 24 hours. On those three endpoints, Didymos exceeded the transferred LightGBM by 0.087, 0.124 and 0.100. The probe also lost when its database was left out, for example from 0.572 to 0.466 for new IMV in eICU-CRD.

Didymos and LightGBM degraded alike with novelty. From the most familiar to the most atypical fifth of anchors, AUROC fell for AKI (Didymos 0.937 to 0.830; LightGBM 0.915 to 0.791), hypotension (0.936 to 0.817; 0.958 to 0.867) and a remaining stay of more than 7 days (0.857 to 0.758; 0.896 to 0.807), and changed little for the treatment starts. Among eICU-CRD hospitals, every one of which has its own site token, AUROC showed no consistent trend with the hospital's number of training stays.

In the subgroups, discrimination varied less than calibration (Supplementary Fig. 13). For ICU death, AUROC was 0.878 in women and 0.844 in men and 0.79–0.92 across age bands; for in-hospital death it ranged from 0.850 to 0.913 across databases, while O/E ranged from 0.99 in NWICU to 2.25 in AmsterdamUMCdb, and was 1.58 in men against 1.18 in women. Where race was recorded, AUROC for white patients (ICU death 0.877) was close to that for stays without a recorded race (0.880), and too few patients of other groups met the threshold for scoring.

Conclusion. The database token carries each database's recording schema through the support mask, and an unknown token preserves discrimination, which is how Didymos is prompted at a hospital it has not seen (Supplementary Note 13). The endpoints on which supervised models transfer worst between databases are those where Didymos keeps its largest advantage. Calibration, not discrimination, is what varies across subgroups and databases.

Supplementary Note 6 | Where the deficit in simulated deaths arises

Question. The pre-specified analysis under-predicted death (O/E 1.14 for ICU death at the 24-hour landmark

and 1.83 for death within 24 hours on rolling anchors). Is the model miscalibrated about death, does generation lose deaths, or did the evaluation censor them?

Analysis. A decomposition fixed before it ran separated three causes, each with its expected signature (Methods, ‘Calibration and recalibration’). Step calibration summed the model’s step hazards along every anchor’s real future and compared them with the observed events: near 1, the model is calibrated given a real history. Generation drift compared the simulated futures’ summed death hazard with that of the same anchors’ real futures: below 1, generation loses deaths. Censoring compared the landmark’s futures at four times the pre-specified cap with the same futures cut back to that cap. Whole synthetic stays were compared with real held-out stays by length of stay.

Results. Given real histories, 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; 0.56–1.52 across databases) and for ICU death at the landmark (0.95; 0.82 to 1.12), but over-predicted death within 24 hours of the landmark (0.76; 0.61 to 0.91). Treatment starts were over-predicted at every step (O/E 0.56–0.80; Supplementary Note 7).

Free generation carried too little death hazard early. Within 24 hours, simulated futures carried 0.61 of the real futures’ death hazard on rolling anchors and 0.57 at the landmark, but 0.91 of it until ICU discharge. Hour by hour, among paths still at risk, the simulated death hazard was 0.56 of the real one in the first 6 hours on rolling anchors and 0.28–0.34 over the next 42 hours, and 0.52 and 0.34–0.49 at the landmark, whereas the hazard of alive discharge stayed at 0.84–1.08 of its real value. Simulated futures were also denser than real ones: a median token spanned 3.75 minutes in simulated and 4.23 minutes in real futures on rolling anchors, so a given token cap covers about a tenth less time in simulation. Treatment starts drifted the other way, with generation adding hazard (ratios 1.12–1.37 on rolling anchors).

At the landmark, the pre-specified cap had stopped 21.6% of futures, and the cap stops the futures of patients who die more often: even at four times the cap, 11.3% of futures were stopped for in-hospital deaths against 3.0% for survivors (ICU death 5.3% against 1.8%). With four times the cap, 8.5% of futures were stopped and ICU death was calibrated (O/E 0.99 for the sampled and 1.01 for the hazard estimate). Cutting the same futures back to the pre-specified cap restored the deficit (O/E 1.10 sampled, 1.38 hazard) while lowering AUROC only slightly (from 0.910 to 0.902 sampled, and from 0.924 to 0.920 for the hazard estimate). The re-run of the 48-hour landmark showed the same pattern: within 24 hours, simulated futures carried 0.76 of the real death hazard, and with four times the cap the hazard estimate’s O/E for ICU death was 1.13, against 1.55 at the pre-specified cap.

Whole synthetic stays lacked deaths mainly among short stays. In MIMIC-IV, ICU death among stays shorter than 24 hours was 3.9% in synthetic against 11.5% in real stays, and 3.5% against 5.0% among stays of 24–72 hours, whereas the longest stays were closer (16.5% against 20.1%). The synthetic stays were complete, with fewer than 1% censored in every database except SICdb (7.9%), so their missing deaths were postponed or replaced by survival rather than cut off.

Conclusion. Didymos is not miscalibrated about death given a real history. Its pre-specified deficit had two causes: free generation postpones deaths, most in the first hours, and the landmark’s token cap censored long simulations of patients who died. Risks of death should therefore be read through the hazard estimate, or from futures long enough to reach the outcome. The postponement is consistent with exposure bias acting on a terminal event; training with the model’s own samples, which would test that explanation, was not attempted here.

Supplementary Note 7 | Calibration of treatment starts and the recalibration maps

Question. How large is the miscalibration of each endpoint, where does it arise, and can a map fitted on a small tuning set correct it without harming discrimination?

Analysis. The pre-specified calibration analysis reported O/E, the calibration slope and the expected calibration error per endpoint and by ICU day of the anchor, and compared the futures' O/E with the teacher-forced O/E along real stays. Recalibration maps were fitted on 100 rolling anchors per database of the tuning split (32 futures each) and applied to new draws on the held-out rolling anchors, at 32 and 16 futures (Methods, 'Calibration and recalibration'). Isotonic maps were adopted where AUROC stayed within 0.005; logistic maps, $\text{logit } p' = a + b \text{ logit } p$, were fitted post hoc for the three treatment starts after the isotonic maps had tied risks.

Results. In the pre-specified analysis, Didymos over-predicted the treatment starts on rolling anchors (O/E 0.69 for new IMV, 0.49 for a new vasopressor and 0.66 for new RRT), as well as AKI (0.73), hyperkalaemia (0.78) and thrombocytopenia (0.81), whereas hypotension (1.04) and sepsis (1.01) were close to calibrated and death within 24 hours was under-predicted (1.83; Supplementary Note 6). The over-prediction of vasopressor starts grew with the day of the stay (O/E 0.59 on day 1 and 0.32 from day 4). Teacher-forced along real stays after the first ICU day, the ratios were already 0.76 for IMV, 0.59 for vasopressors and 0.44 for RRT, so the over-prediction is present at every step before any generation drift.

The isotonic maps corrected the treatment starts' O/E (new IMV from 0.55 to 1.16, a new vasopressor from 0.46 to 1.00 and new RRT from 0.55 to 0.84), but their flat steps tied risks and cost discrimination on the rarer endpoints (a new vasopressor 0.894 to 0.885; new RRT 0.958 to 0.940; thrombocytopenia 0.892 to 0.839). They were adopted for death within 24 hours, AKI, hypotension and hyperkalaemia (Supplementary Fig. 14). A map corrects O/E only as far as the tuning split's miscalibration matches the target's: for death within 24 hours, the raw O/E was 1.98 on the tuning anchors and 1.71 on the held-out anchors, so the map overcorrected to 0.75. The logistic maps ($a = -1.74, -1.41$ and 0.02 ; $b = 0.72, 0.80$ and 1.46 for new IMV, a new vasopressor and new RRT) kept every AUROC. They brought a new vasopressor to O/E 1.02 (1.04 at 16 futures) and new RRT to 0.93 (0.90), with expected calibration errors falling from 0.030 to 0.009 and from 0.016 to 0.002, and Brier scores falling too. New IMV overshoot (O/E 1.43), because its raw O/E on the tuning split had been 0.40 against 0.55 on the held-out anchors, although its calibration error still fell from 0.031 to 0.013. Before correction, O/E differed from 1 for five of the seven endpoints with maps (Holm $P < 0.05$), and after correction for two: death within 24 hours (0.75) and hyperkalaemia (0.67).

Conclusion. The over-prediction of treatment starts is a per-step miscalibration of rare-event timing, not a product of generation. A two-parameter logistic map fitted on 700 tuning anchors corrects most of it without changing ranks, but its success depends on the tuning set resembling the target, so larger tuning sets or recalibration at each site are needed. The adaptation analysis shows the same limit at new hospitals (Supplementary Note 14).

Supplementary Note 8 | What-if simulation: responses, matching, the null arm and factual accuracy

Question. Do simulated treatment responses move in the physiological direction, do they reproduce what is observed in matched patients, and how do they depend on where the record is cut?

Analysis. The pre-specified design cut held-out stays 0.5–6 hours before a recorded treatment start and compared a give arm, which writes the database's most common start record at its median rate, with a withhold arm, 16 futures per arm, for nine responses with expected signs (Methods, 'What-if simulation'). The follow-up cut starters 1 minute before their first recorded start and matched each, within the database and at the

same ICU hour, to a non-starter on the propensity of starting within 6 hours.⁸ It then compared the observed contrast between the pairs with Didymos's factual contrast, and added a null arm of two withhold arms drawn with different seeds. Factual accuracy compared each factual arm's 16 futures with the real response by the continuous ranked probability score⁹ and the coverage of central 80% intervals.

Results. Cut 0.5–6 hours early, as pre-specified, the give arm moved five responses in the expected direction with intervals excluding zero: glucose after insulin (−14.2 mg/dL; −22.6 to −6.4), potassium after continuous RRT (−0.095 mmol/L; −0.173 to −0.024), urine output after furosemide (+723 mL; 627 to 834), and inspired-oxygen charting (+2.46; 2.17 to 2.77) and oxygen saturation (+0.38; 0.23 to 0.54) after intubation. Creatinine after continuous RRT and both responses to a fluid bolus did not move, and MAP after norepinephrine moved the wrong way (−0.81 mmHg; −1.57 to −0.14).

The null arm behaved as chance: two withhold arms differed in the expected direction in 0.40–0.55 of prefixes, and every mean difference's interval covered zero (for example 0.03 mmHg for MAP after norepinephrine, −0.66 to 0.75, against a standard deviation of 4.7 mmHg across prefixes).

Matching balanced the propensity of starting closely: of 285 (continuous RRT) to 2,084 (insulin) eligible starters, 158–195 were matched per treatment, with mean propensity logits within 0.04 of their matches'. Didymos's factual contrast lay within the observed contrast's 95% CI for eight of nine responses (Fig. 4b). It reproduced the selection of patients as well as the response: continuous RRT starters tended to have higher potassium and creatinine than their matches in reality (+0.24 mmol/L and +0.50 mg/dL, with intervals including zero) and did so in simulation (+0.44; 0.30 to 0.61, and +1.04; 0.75 to 1.34). Urine output after furosemide was +337 mL observed against +403 mL simulated, inspired-oxygen charting after intubation +2.82 against +3.11, and MAP after a fluid bolus −5.84 against −4.30 mmHg. Insulin was the exception: real starters' glucose stayed 38.0 mg/dL above their matches' (6.4 to 72.2), against −0.3 mg/dL simulated, although within the simulation giving insulin lowered glucose by 22.6 mg/dL in starters and 22.7 mg/dL in non-starters. For urine output after a fluid bolus, the observed contrast was too imprecise to judge (−8 mL; −139 to 137, against −108 mL simulated).

Cut at the decision time, the within-model effect of giving against withholding a treatment had the expected sign with an interval excluding zero for six of nine responses: MAP after norepinephrine (+0.79 mmHg; 0.11 to 1.47), glucose after insulin, creatinine after continuous RRT (−0.28 mg/dL), urine output after furosemide (+439 mL), and inspired-oxygen charting (+2.39) and oxygen saturation (+0.49) after intubation. It was null for potassium after continuous RRT and urine output after a fluid bolus, and opposite for MAP after a fluid bolus (−0.79 mmHg), where a falling pressure prompts the bolus. The factual futures improved on carrying the last value forward for every response, with scores 0.27–0.60 times that of persistence among starters, and their 80% intervals covered 0.66–0.84 of observed responses, so their spread was somewhat narrower than the observed one.

Conclusion. The simulator reproduces observed practice, including the selection of patients for a treatment and its confounding by indication, and when cut at the decision time its within-model effects mostly have the physiological direction. Its contrasts describe what is observed, not what a treatment causes, and the insulin result marks where the two diverge.

Supplementary Note 9 | Privacy audits of training and of prompted sampling

Question. Can an attacker tell from Didymos's behaviour, or from what it generates, that a stay was in its training data, with or without a prompt from a real record?

Analysis. The audit of training (Methods, 'Privacy audits') ran five tests per database and pooled: a loss-based

membership attack on up to 3,000 training and 3,000 held-out stays per database;^{10,11} prefix-prompted extraction, in which greedy continuations of 1,000 members' and 1,000 non-members' first windows per database were compared with the stored continuation of 32 tokens after a 32-token prompt and of 64 tokens after a 128-token prompt;¹² exact copies and nearest-neighbour distances of 5,000 audit samples to every training window; a density-ratio attack in a 32-dimensional projection of the samples; and attribute inference of a recorded death from the nearest sample. The audit sampled without the grammar, which only removes invalid tokens and cannot add memorization. A copying generator and a private reference of unseen stays served as controls. The audit of prompted sampling scored a stay's real next 24 hours given its record up to a random whole hour, and asked whether 16 futures reproduced the real future, for plain prompts and for norepinephrine starts.

Results. Membership AUC was 0.549–0.555 in every database (pooled 0.538; 95% CI 0.532 to 0.544), and the true-positive rate at a 1% false-positive rate 0.010–0.024 (pooled 0.015). Greedy continuations reproduced no stored continuation, for members or non-members, in either setting, and members' continuations agreed with the stored one for barely more tokens than non-members' (0.75 against 0.72 on average). No audit sample copied a training window. The share of samples closer to a training window than the held-out stays' 5th percentile was 0.033 (0.028 to 0.038), and the same for the distance ratio, against a limit of 0.07. The density-ratio attack had an AUC of 0.495 (0.482 to 0.505). Attribute inference was as accurate for members as for non-members (difference in balanced accuracy 0.006; –0.004 to 0.018; 20,000 per side). The controls behaved as designed: the copying generator was flagged by every sample-based test (density-ratio AUC 0.961; attribute-inference difference 0.384), and the private reference by none.

Across nine checkpoints, the pooled membership AUC ranged from 0.521 to 0.564 and the worst database's from 0.547 to 0.650, rising with the fit reached (Supplementary Note 2; Fig. 5b). For prompted sampling, an audit with 500 stays per side gave a pooled AUC of 0.525 (0.512 to 0.539) and left three databases' true-positive-rate intervals too wide to judge. At up to 3,000 per side, the pooled AUC was 0.531 (0.524 to 0.537) and the true-positive rate 0.013 (0.011 to 0.017), with database AUCs of 0.529–0.568; INSPIRE, with 790 eligible stays per side, had the widest interval (Methods). No future of a member or a non-member reproduced the first 32 tokens of its real future, with or without an intervention, and the similarity of the best future's events over 24 hours did not separate members from non-members (AUC 0.516; 0.493 to 0.540; with a norepinephrine start, 0.482; 0.400 to 0.553). The copying generator reproduced every member's future, and two halves of the non-members gave an AUC of 0.493.

Conclusion. The membership signal of Didymos is weak, of the size expected from the fit reached, and similar for whole stays and for prompted continuations, the mode that every digital-twin application uses. No test found copying or extraction. These audits are empirical and specific to the attacks tested; attacks calibrated with reference models, which need models trained without each target, were not run, and formal guarantees would need differential privacy.

Supplementary Note 10 | Fidelity and utility of the synthetic cohort

Question. Do whole synthetic stays reproduce the databases and hospitals they are conditioned on, and do they add information when real data are scarce?

Analysis. The synthetic cohort of 13,416 stays (eICU-CRD 3,975, MIMIC-IV 2,608, HiRID 1,601, NWICU 1,492, SICdb 1,471, AmsterdamUMCdb 1,203 and INSPIRE 1,066) was compared with 2,000 held-out stays per database and with 2,000 training stays as a real-against-real control (Methods, 'Fidelity and utility of synthetic stays'). Utility compared LightGBM trained on synthetic stays with LightGBM trained on as many

real stays, both tested on real held-out anchors. Augmentation added one or three times as many synthetic stays to 100, 300 or 1,000 real stays of one database, in 20 repetitions per cell, with two intervals: over the repetitions, and over the test anchors.

Results. Per database, synthetic stays matched real ones in age, sex, treatments, rates of events and laboratory tests, and length of stay, as closely as the real control did (Fig. 5c). In eICU-CRD, for example, 23.3% of synthetic stays were ventilated against 24.6% of real and 25.4% of control stays, with 11.9 against 11.8 events per hour; the median stay in MIMIC-IV was 45.6 hours against 47.1. Deaths were the exception: ICU death was 5.7% against 8.8% in MIMIC-IV, 5.8% against 12.9% in AmsterdamUMCdb and 3.1% against 5.8% in HiRID, but close in NWICU (6.1% against 6.7%), SICdb (3.7% against 4.0%) and eICU-CRD (3.9% against 4.8%), in keeping with the postponement of deaths (Supplementary Note 6). The structure between databases correlated at 0.73 (0.61 to 0.83) with the real one, against 0.95 for the control. The total variation of eICU-CRD hospitals' composition from the real one was 0.146, against 0.145 for the control, and across the 17 hospitals with enough stays, the rank correlations of hospital rates were 0.45 for ICU death, 0.63 for invasive ventilation, 0.66 for vasopressors and 0.72 for the median stay, against 0.61, 0.52, 0.85 and 0.71 between two real samples, with wide intervals (Fig. 5d). All 25 rare header strata were reached, at 2.2% of synthetic against 1.5% of real stays. In the selection audits, the diversity of synthetic stays was 0.88–0.89 of the real stays' by the Vendi score, and they recalled 0.91–0.92 of rare codes.

Models trained on synthetic stays reached 0.93–1.03 of the gain over chance of models trained on as many real stays (Fig. 5e), for example 0.870 against 0.900 AUROC for ICU death at the 24-hour landmark. With 100 real stays of a database, adding as many synthetic stays raised AUROC by 0.038 at the landmark (gains in 19 of 19 cells) and by 0.021 on rolling anchors (25 of 34), and three times as many by 0.059 and 0.033; 100 real and 100 synthetic stays were worth a median of 179 real stays at the landmark. With 300 real stays, adding as many synthetic stays gained 0.012 in both designs, and with 1,000 it gained 0.010 or less. The pre-specified small-site arms, with one draw each, had varied in both directions; the repetitions resolved them into these consistent gains.

Conclusion. Synthetic stays reproduce each database's profile, the structure between databases and much of the structure between hospitals, and carry most of the information supervised models use. Their limit is fewer early deaths. Their value as added training data is largest where real data are scarcest and fades by about 1,000 real stays.

Supplementary Note 11 | Attribution: sanity checks, the failure of attention rollout and agreement with the generator

Question. Which methods of explaining an individual's predicted risk reflect what the model learned, why does attention rollout not, and how well does any method agree with the generator's own response to editing the record?

Analysis. Groups of a patient's record, each a channel family within a time block, were scored by counterfactual editing, integrated gradients,¹³ attention rollout¹⁴ and attention $\times$ gradient for the Didymos probe's risk (Methods, 'Attribution and sanity checks'). The pre-specified analysis measured faithfulness as the fall of the probe's logit when a method's top one to five groups were removed, on 200 anchors per endpoint. The follow-up simulated each group's edit with 48 futures on 40 higher-risk anchors per endpoint, giving the generator's own effect and its split-half reliability. Sanity checks repeated each method with randomized model weights or permuted probe labels.¹⁵ An exploratory diagnosis compared rollout with a structural null and tested remedies, and a later amendment drew the generator's effects again for all 160 anchors.

Results. Editing was the most faithful method in the pre-specified analysis: removing its top groups lowered the probe's logit by 1.88–2.66, against 0.50–0.82 for integrated gradients, 0.18–0.67 for rollout and 0.08–0.53 for a random order. The groups it ranked highest were clinically coherent: the admission details for ICU death (among the top three in 62% of anchors) and the vital signs of the last 6 hours for hypotension (80%) and AKI (51%).

Against the generator's own effects, editing agreed at 0.12–0.20 (median Spearman correlation per anchor) on three endpoints, within reliability ceilings of 0.48–0.54, and at 0.07 for hypotension, within a ceiling of 0.13; it agreed in sign on 61–66% of groups with clear effects. Integrated gradients (–0.10 to 0.08), rollout (–0.11 to 0.08) and attention $\times$ gradient (–0.09 to 0.14) did not agree. Editing and integrated gradients were unrelated to themselves after randomizing the weights (–0.08 to 0.07), whereas rollout ranked groups almost as it did in the trained model (0.96–0.97) and attention $\times$ gradient partly so (0.53–0.65). With permuted labels rollout was unchanged (1.00), because it never uses the target.

The diagnosis traced the failure to the causal mask (152 anchors with prompts of at most 4,096 tokens). Rollout of uniform causal attention, with no model, ranked groups like the trained rollout (0.97) and exactly like the random one (1.00). Multiplying attention over 12 layers drains mass towards early tokens whatever the attention does: 35–38% of the rollout landed on the first token in both models, which received 0.2% of the trained model's last-layer attention. The trained model's heads were sharper than the random model's (entropy 0.68–0.81 against 0.92–0.95 of the maximum), but averaging heads and multiplying layers brought its rollout back within 0.19 total variation of the null. Summing token scores over a group added a confound by size: the largest group held a median of 84 times the tokens of the smallest, and raw last-layer attention, unrelated between the two models token by token (0.03), correlated once summed per group (0.78). Remedies that removed the geometry or the size confound left no signal about the endpoint. Three variants aggregated by group means met the rule set before the diagnosis (a sanity correlation of about 0.3 or less and positive agreement with editing): attention $\times$ gradient (–0.03 and 0.14), gradient-weighted rollout¹⁶ (–0.04 and 0.14) and the last layer's attention (0.01 and 0.08). Attention knockout,¹⁷ which blocks attention to a group in every layer, met it most clearly (–0.02 and 0.71; 0.60–0.74 by endpoint), and was unrelated to group size (0.03) and to the null (0.00). Rollout variants, including norm-weighted rollout,¹⁸ did not (sanity 0.92–0.97).

Drawn again with 2,698 prompts and 48 new futures each, the generator's effects agreed with the first draw: editing's agreement was 0.09–0.22, against 0.07–0.20 in the first, within ceilings of 0.34–0.58 (Supplementary Fig. 15). Knockout agreed with the redrawn effects at 0.06–0.23, and in sign on 54–68% of groups with clear effects, against 62–73% for editing. Per anchor, knockout's agreement equalled editing's (median difference 0.00; Wilcoxon $P = 0.61$; 158 anchors), except for new IMV, where editing was ahead (median –0.10; Holm $P = 0.039$).

Conclusion. Interventions on the record or on attention paths, editing and knockout, reflect the trained model and agree with each other and, within the limits of simulation noise, with the generator. Attention rollout in a decoder-only model describes the causal mask, so it should not be read as evidence of what drove a prediction, and token scores should be aggregated by group means. Individual-level attribution remains noisy at affordable compute, because the generator's own effects are only moderately reproducible.

Supplementary Note 12 | The toolkit's validation, similar patients and trial simulation

Question. Does the Didymos toolkit reproduce the paper's estimates with new draws, do neighbours in Didymos's state share outcomes, and how much do Didymos risks shrink the size of a randomized trial?

Analysis. The toolkit was run with new draws on a separate host, with prompts identical to the paper's (1,050

of 1,050), and compared with both paper runs on the same anchors (Methods, ‘The Didymos toolkit and trial simulation’). Similarity was tested by leave-one-out nearest neighbours in four representations (Methods, ‘Similarity and site prompts’). The Trial designer simulated 2,000 trials per endpoint, prognostic score and relative risk.

Results. The toolkit’s AUROC and O/E lay within simulation noise of one of the paper’s two runs on all nine rolling primary endpoints (Supplementary Fig. 16). The paper’s runs themselves differed: for death within 24 hours, the pre-specified run (0.903) lay 3.7 noise standard deviations above the follow-up run’s expectation at 32 futures (0.881), and the toolkit gave 0.880. The pre-specified rule, which compared against that run’s interval alone, was met by 14 of 15 checks; the exception, sepsis with 8 events, gave 0.963 against 0.989 (0.974 to 0.999). The hazard estimate at 32 futures lay within 0.011 of the follow-up’s curve, and the what-if contrasts were reproduced: MAP after norepinephrine –3.85 mmHg against –3.58, and urine output after furosemide +405 mL against +403, with both null-arm intervals covering zero.

Neighbours in Didymos’s state shared outcomes. In the pre-specified analysis, a positive patient’s ten nearest neighbours were positive 2.47–4.42 times as often as the prevalence, against 2.03–3.29 for TF-IDF of the record (new RRT 4.42 against 3.29). Used as a predictor with ten neighbours, Didymos’s state gave AUROCs of 0.861 for death within 24 hours, against 0.845 for TF-IDF, 0.841 for the comparators’ features and 0.628 for an untrained model’s state, and 0.845 for new RRT, against 0.776, 0.760 and 0.666 (Supplementary Fig. 17). TF-IDF was slightly ahead for new IMV (0.668 against 0.640) and hyperkalaemia. AUROC rose with the number of neighbours (death within 24 hours 0.891 with 50), and neighbours drawn from the same database gave the same results as neighbours from all.

In simulated trials with a relative risk of 0.8, adjusting for Didymos risk needed 0.63–0.86 of the unadjusted sample size: 0.63 for in-hospital death, 0.64 for ICU death, 0.67 for AKI, 0.74 for new RRT, 0.84 for a new vasopressor and 0.86 for new IMV (Fig. 7b; Supplementary Fig. 18). LightGBM needed 0.64–0.93, and SOFA, SAPS II, OASIS and NEWS 0.80–1.00. At a relative risk of 0.7, the ratios were 0.64–0.84 for Didymos. Enrolling the quarter at highest Didymos risk needed 0.18–0.30 of an unenriched trial’s size, against 0.25–0.88 for the clinical scores. In eICU-CRD alone, adjusting for APACHE IVa needed 0.81–1.01 of the unadjusted size across the six endpoints, against 0.57–0.72 with Didymos. The design checks were met in 92 of 108 cells each. The adjusted analyses’ type I error was 0.0375–0.0715 at the sizes for a relative risk of 0.8, and the highest values, up to 0.079 at 0.7, were Didymos’s in small eICU-CRD-only subsets for a new vasopressor and new RRT, where a near-perfect score strains the Wald test. For teaching, each of three objectives yielded three synthetic cases from 40 candidates, all meeting their objective.

Conclusion. The toolkit reproduces the paper’s estimates within the run-to-run spread of 32-future simulations, and that spread is itself a finding: one 32-future run is not a precise estimate of discrimination for sparse or near-ceiling endpoints. Neighbours in Didymos’s state carry outcome-relevant similarity, with a modest margin over simple representations. As a prognostic covariate, Didymos risk shrinks simulated trials more than clinical scores do and at least as much as LightGBM, with type I error near nominal except in the smallest single-database subsets.

Supplementary Note 13 | Transfer to two hospitals never seen in training

Question. How well does Didymos forecast at hospitals whose patients it has never seen, recorded under schemas it was not built for, and which endpoints can such hospitals support?

Analysis. RICD¹⁹ and zigongICU²⁰ were harmonized with new adapters and encoded with the frozen vocabulary, and prompts carried the database and site as unknown (Methods, ‘Transfer to unseen hospitals

and adaptation'). A gate recorded before any future was drawn fixed the anchors, checked the encoding and decided which endpoints each database supports, by a rule set beforehand: the endpoint's source is timed to the hour, and at least 20 anchors of the design are positive. Each anchor received 128 futures. Comparators trained on the seven databases, a LightGBM trained within each database and cross-fitted by patient, and clinical scores were scored on the same anchors.

Results. The harmonized cohorts comprised 5,382 ICU stays from RICD (9.6 million events; 579 ICU deaths; in-hospital death 13.3%) and 2,899 from zigongICU (7.8 million events; 716 ICU deaths; 29.5%). Their recording differs from the training databases'. zigongICU's nursing chart records dates without hours, and RICD charts urine output as totals about every 4 hours (median 500 mL per entry), against about every 20 minutes in zigongICU (100 mL). Encoding gave no grammar or parser rejections and no round-trip failures, but sent many medication events to unknown tokens: 85.1% of RICD's continuous-infusion events, 31.6% of its intermittent doses, 14.5% of its output records and 12.7% of its location records, and 55.5% and 6.5% of zigongICU's infusions and doses, against 0.0–3.7% in any domain of the training databases' held-out stays. In RICD, 16 of 109 numeric codes had more than half their values in an end bin of the training deciles, mostly standard doses of intermittent drugs, and in zigongICU none of 76. Each anchor's real stay reproduced its labels (agreement 1.000 on 998 rolling and 1,400 landmark anchors).

The gate kept 15 of 26 endpoint–database pairs. RICD records ventilation only as a daily flag and has no source for RRT, and too few of its anchors were positive for hypotension (6), thrombocytopenia (7) or death within 24 hours of the landmark (8). zigongICU's ventilation, urine output and blood pressure come from the date-only nursing chart, and it had 12 positive anchors for RRT. Neither database supplied culture data for the sepsis definition (zigongICU has none, and RICD's culture results were not harmonized).

On the scored pairs, Didymos's AUROC ranged from 0.536 for hyperkalaemia in RICD to 0.934 for death within 24 hours in RICD (median 0.72), against 0.83–0.98 for the same estimators on the seven databases (Supplementary Fig. 19a). It exceeded LightGBM trained on the seven databases on 10 of 15 pairs (differences –0.100 to +0.155; lowest Holm $P = 0.096$) and TabPFN-3.5 on 6 of 15. Against the best clinical score per pair, chosen after the fact, it was ahead on 9 of 15, by up to 0.153, and behind RICD's own recorded SOFA on its two landmark mortality endpoints (–0.054 and –0.065). A LightGBM trained on each database's own labelled stays was ahead on 13 of 15 pairs, most clearly for RICD's three landmark endpoints (–0.089 to –0.218), zigongICU's new vasopressor (–0.191) and its remaining stay of more than 7 days (–0.101; Holm $P \leq 0.012$ each), whereas Didymos was ahead for zigongICU's death within 24 hours of the landmark (+0.117; Holm $P = 0.0044$).

Transferred risks needed local recalibration (Supplementary Fig. 19c). Didymos's O/E ranged from 0.04 for a new vasopressor in RICD to 3.46 for its remaining stay of more than 7 days, and a logistic map within each database, cross-fitted by patient, brought every O/E to 0.93–1.01; the transferred LightGBM's O/E (0.23–4.41) needed the same. By mean rank among the ten predictors that never saw the new databases (post hoc), TabPFN-3.5 came first (3.27; 95% interval 2.07 to 4.67) and Didymos second (3.80; 2.60 to 5.07), ahead of the transferred LightGBM (4.33), the probe (4.80) and NEWS, MEWS and SOFA (5.5–5.9) (Supplementary Fig. 19d). Didymos ranked third of 11 in zigongICU (3.12), behind the local model and TabPFN-3.5, and seventh in RICD (6.43), where SOFA and NEWS were ahead.

Conclusion. Without any training on the new hospitals, Didymos forecast at the level of supervised models trained on other databases, with lower discrimination than at home and risks that needed local recalibration. A model trained on the hospital's own labels remained ahead, most in RICD, a neurorehabilitation centre with sparse monitoring and long stays, where local practice differs most from the training databases. This gap

motivated the adaptation analysis (Supplementary Note 14).

Supplementary Note 14 | Adaptation to the new hospitals without labels

Question. Which levers close the gap at a new hospital: mapping the codes the vocabulary lacks, continuing training on the hospital's unlabelled stays, or combining the adapted model with a few labels?

Analysis. Under an amendment fixed before any adapted model was trained, each database's patients were split once into a locked test set (50%), unlabelled stays for continued training (35%) and a labelled slice (15%), and every arm was scored on the same test anchors with the transfer analysis's labels (Methods, 'Transfer to unseen hospitals and adaptation'; Fig. 9a). The arms were the zero-shot model with the original encoding, the same model with mapped codes (Supplementary Table 8), the adapted model, the adapted model recalibrated on the labelled slice, and the adapted model stacked with a LightGBM trained on that slice. The references were LightGBMs trained on the labelled slice or on all adaptation stays with their labels, TabPFN-3.5 and the clinical scores. Pairs with fewer than 10 test positives were reported descriptively.

Results. The split assigned 2,015, 1,388 and 645 RICD patients (2,646, 1,859 and 877 stays) and 1,384, 927 and 431 zigongICU patients (1,460, 973 and 466 stays) to the three parts. The locked test set held 490 rolling and 683 landmark anchors, with 13 pairs of at least 10 test positives.

Mapping codes removed almost every unknown event: RICD's continuous infusions from 85.1% to 0.0%, its intermittent doses from 31.6% to 2.9% and its output and location records to 0.0%, and zigongICU's infusions from 55.5% to 0.6% and doses from 6.5% to 1.2% (Supplementary Fig. 20a). Labels under the mapped encoding agreed with the original ones (lowest agreement over programs 0.982), and discrimination barely moved (-0.043 to $+0.025$ on the 13 pairs; Holm $P = 1.00$ for each). The largest change, $+0.18$ for a new vasopressor in RICD, came from a pair with 8 test events.

Continued training lowered the next-token loss on held-out patients of the new databases from 2.100 to 1.131 nats per token with the chosen cell, a learning rate of 10^{-4} and three passes (RICD 2.095 to 1.321; zigongICU 2.103 to 0.957); one pass at the same rate gave 1.294, and three and one passes at 3×10^{-5} gave 1.291 and 1.530 (Supplementary Fig. 20c). The chosen cell took 228 steps of 131,072 tokens, 7.5 minutes on one GPU. On the seven databases' tuning split, the loss rose by 0.012–0.033 nats per token.

On the locked test anchors, the adapted model exceeded the zero-shot model on 8 of 13 pairs (-0.029 to $+0.125$; lowest Holm $P = 0.21$; Fig. 9c,d). The gains were largest for outcomes shaped by local practice: a remaining stay of more than 7 days in RICD ($+0.119$; 0.014 to 0.214) and zigongICU ($+0.088$; 0.015 to 0.161), a new vasopressor in zigongICU ($+0.125$; 0.005 to 0.233) and hyperkalaemia in RICD ($+0.095$; 0.016 to 0.213). RICD's mortality endpoints gained 0.043 and 0.057, whereas zigongICU's lost 0.014–0.024. On the two descriptive pairs, AUROC rose from 0.628 to 0.851 for a new vasopressor in RICD (8 events) and from 0.702 to 0.851 for hyperkalaemia in zigongICU (9 events).

Recalibrated on the labelled slice, the adapted model was level with a LightGBM trained on all adaptation labels, half of each database's patients: ahead on 6 of 13 pairs, with differences of -0.075 to $+0.164$ (lowest Holm $P = 0.081$). It was ahead of TabPFN-3.5 on 8 of 13, most for a remaining stay of more than 7 days in RICD ($+0.167$; 0.064 to 0.264; Holm $P = 0.021$). Stacking the adapted model with a LightGBM trained on the labelled slice gave the best predictor: mean rank 3.15 (95% interval 2.15 to 4.23) among 15 predictors, with a mean AUROC of 0.795, and 2.15 (1.38 to 3.08) when the three local LightGBMs were left out (Fig. 9e). The adapted model's AUROC exceeded every clinical score's on 12 or 13 of the 13 pairs (Holm $P \leq 0.007$).

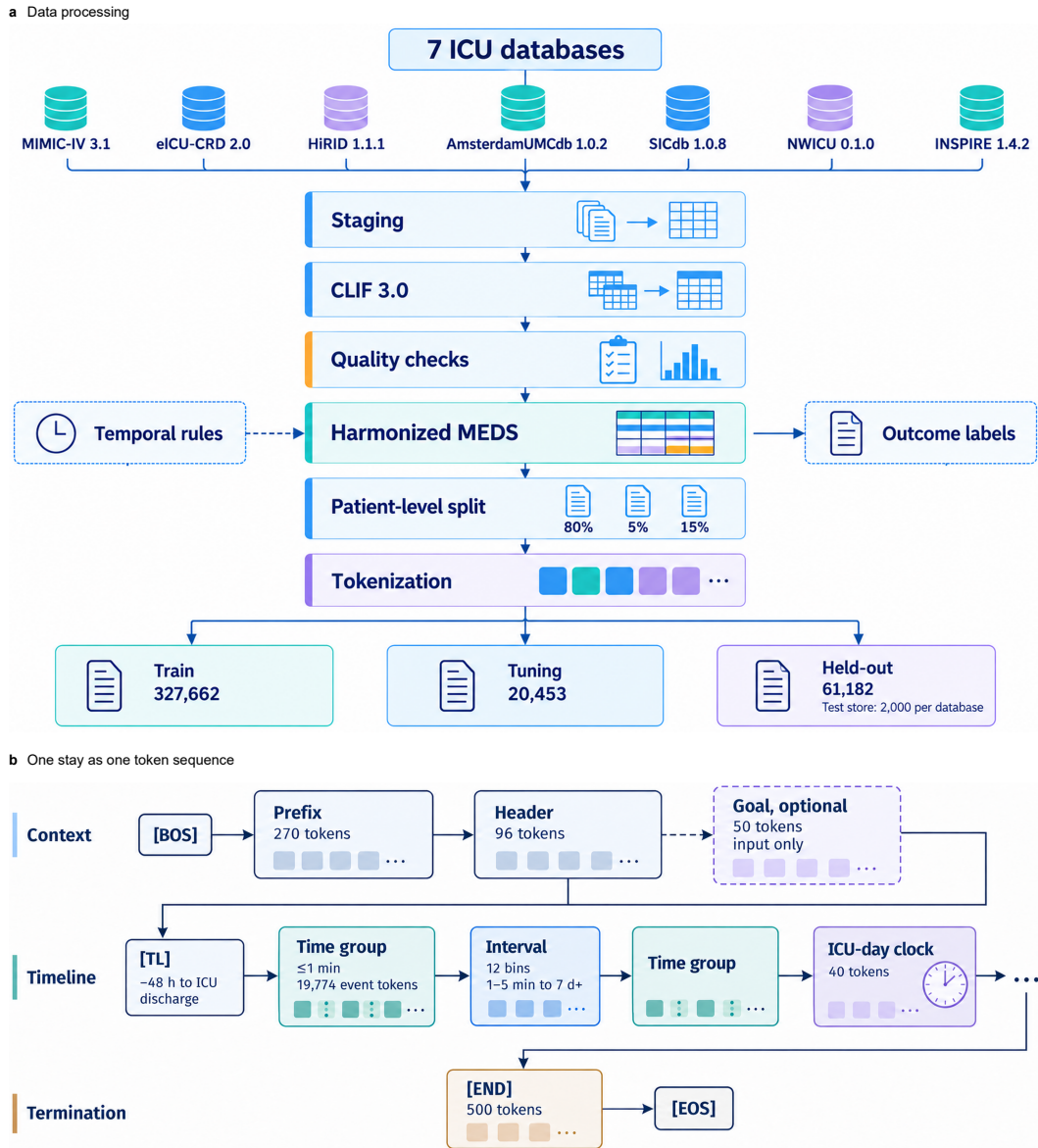
Adaptation improved calibration where the labels concern the whole stay. At the landmark, the adapted model's O/E was 0.71–1.39 before any recalibration, against 0.32–3.60 for the zero-shot model, and 0.65–1.24

after recalibration (Supplementary Fig. 21a). Rare events within 24–48 hours stayed under-predicted after recalibration on about 250 labelled anchors per database (AKI 0.35 and hyperkalaemia 0.28–0.32). A LightGBM trained on 25 to 877 labelled stays reached the recalibrated adapted model on only 3 of 15 pairs, at about 250–390 stays (Supplementary Fig. 21b).

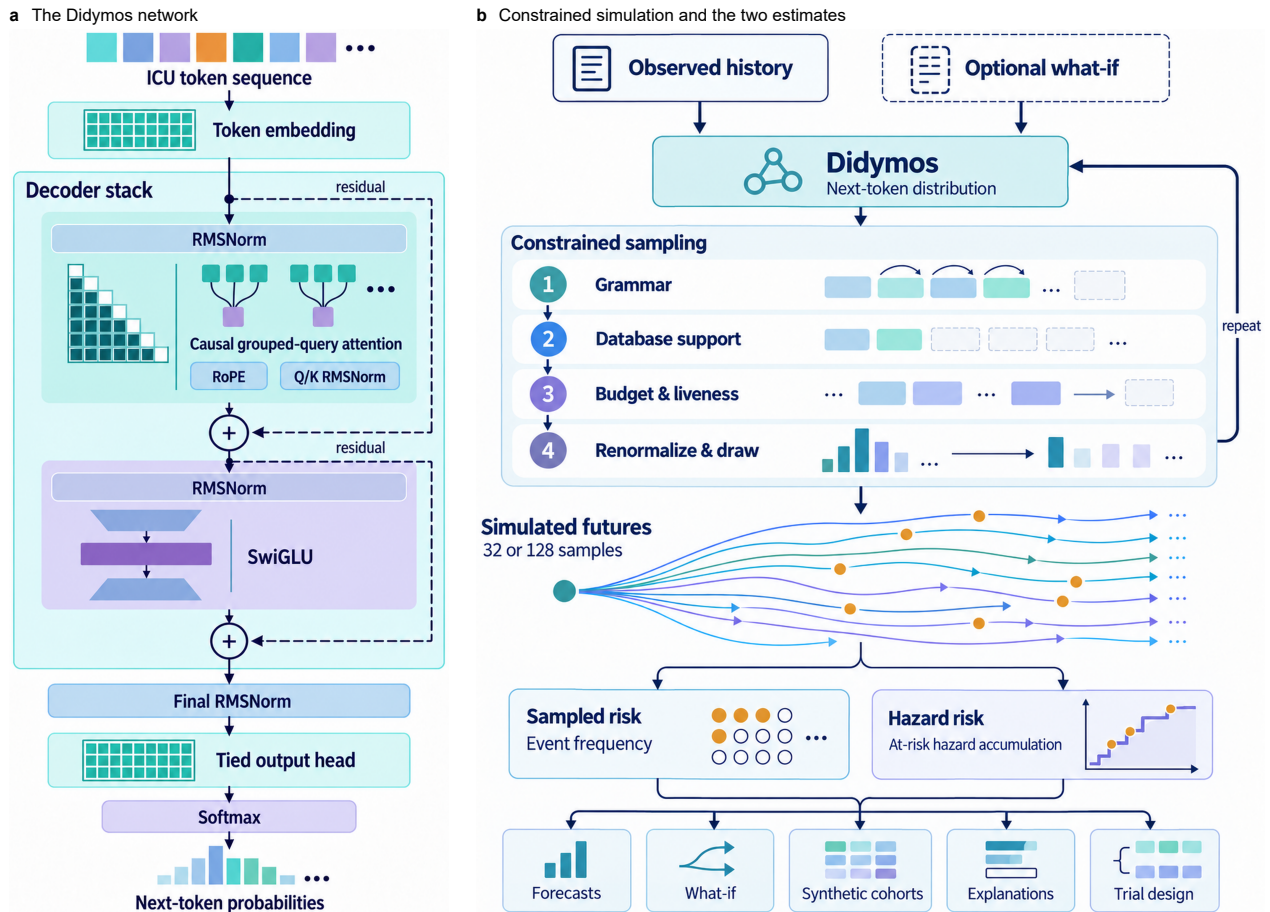
Continued training gave the adapted stays a membership signal (AUC 0.581 in RICD and 0.542 in zigongICU; pooled 0.564, 0.548 to 0.579, against 0.502 before adaptation; Supplementary Fig. 20d). On the seven training databases, the adapted model’s membership signal matched the original’s (pooled true-positive rate at a 1% false-positive rate 0.0151, against 0.015; AUC 0.548–0.559 by database at the larger size), and its other privacy tests were within their limits. Two groups’ upper bounds of the true-positive rate lay just above the pre-specified 0.03, eICU-CRD’s at the registered size (0.0303) and AmsterdamUMCdb’s at the larger size (0.0313, with all 2,747 of its held-out stays), so its futures were drawn for this analysis only (Supplementary Fig. 22; Methods).

Conclusion. At a new hospital, the effective lever is the hospital’s own unlabelled stays. Mapping codes helps little, whereas minutes of continued training lift Didymos on most endpoints, most where its zero-shot forecasts were weakest, and leave its performance on the training databases nearly unchanged. Combined with a small labelled sample, the adapted model gives the best predictor tested. The adapted stays acquire the membership signal that any training confers, so the adapted weights belong under the hospital’s governance.

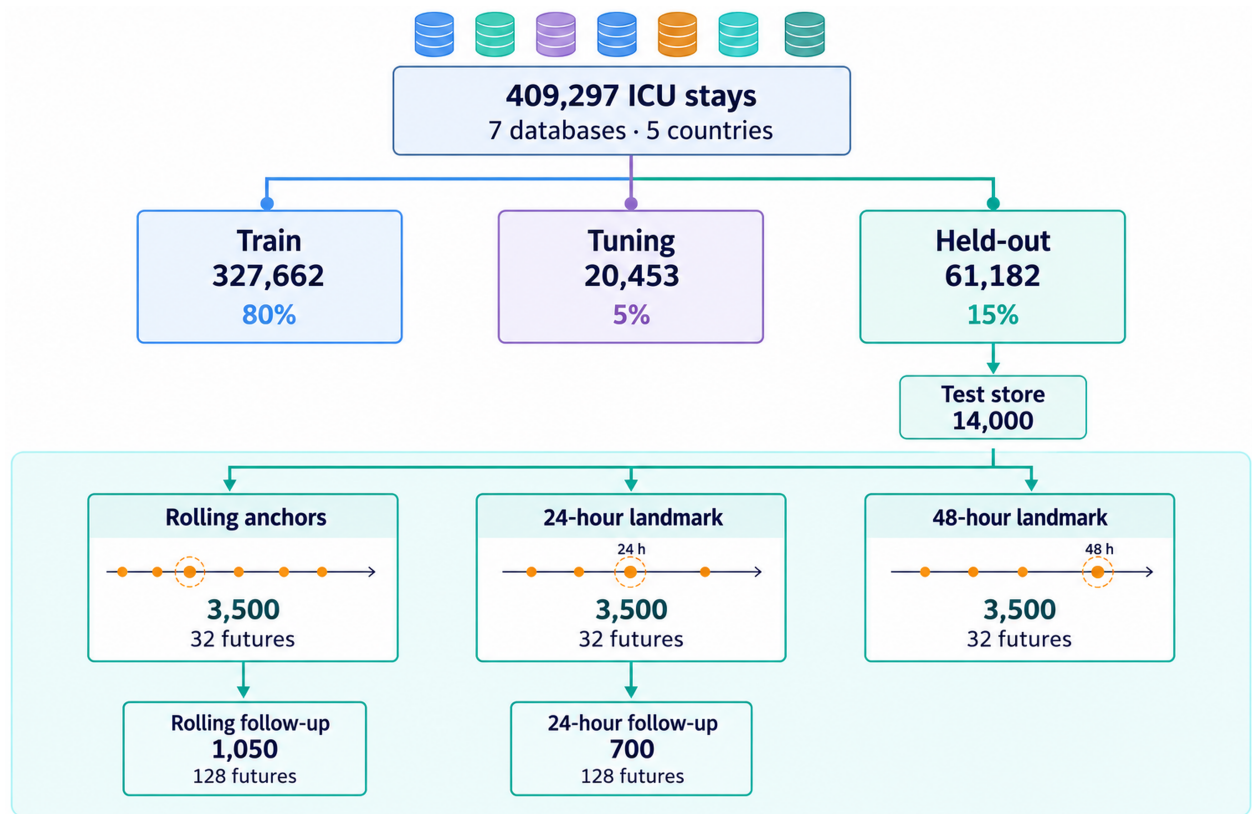
Supplementary Figures



Supplementary Fig. 1 | From seven source databases to model-ready token sequences. **a**, Data processing (GPT-image). Each database's archives were staged, converted to CLIF tables (official pipelines for MIMIC-IV and AmsterdamUMCdb; mapping-table adapters for the other five) and checked by schema and vocabulary validation, cross-database distributions and reconciliation with community MEDS pipelines. The data were harmonized into MEDS under uniform temporal rules, with one subject per adult ICU stay, from 48 hours before ICU admission to ICU discharge. Derived concepts (KDIGO stage, treatment episodes, SOFA, Adult Sepsis Event) serve as labels only. Stays were split within each database by patient (80%, 5% and 15%) and tokenized with a vocabulary fitted on the train split; the grammar accepts all 409,297 stays, and decoding returns the same tokens. **b**, One stay as one sequence (GPT-image): the prefix and header, an optional goal block, the timeline of time groups and interval tokens with ICU-day clock blocks, and the end block. The counts are the tokens of each class of the vocabulary. The artwork is drawn from its Mermaid sources (Supplementary Data 4) and 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; KDIGO, Kidney Disease: Improving Global Outcomes; SOFA, Sequential Organ Failure Assessment.



Supplementary Fig. 2 | Didymos's network and how it simulates futures. **a**, The network (GPT-image): a decoder-only transformer of 12 blocks, width 768 and 123.5 million parameters, over a vocabulary of 25,601 tokens and a context of 32,768 tokens, with grouped-query attention (12 query and 4 key-value heads), rotary positions and SwiGLU feed-forward layers; the token embedding is shared with the output head. Didymos was trained once on the train split of all seven databases by next-token cross-entropy, without task labels; the number of passes over each database was set from a membership-inference audit. **b**, Simulation from a stay's record up to the prediction time (GPT-image). At each step, the next-token probabilities are restricted by the grammar, by the database's support and by the token budget, then renormalized and sampled at temperature 1; a what-if writes a treatment start at the cutoff, or forbids it. Each simulated future gives the sampled estimate (labelled "sampled risk": whether the event occurs by the horizon) and the hazard estimate ("hazard risk": its step hazards of the event, accumulated while at risk). The artwork is drawn from its Mermaid sources (Supplementary Data 4) and credited (GPT-image); the authors confirm its accuracy and scientific validity and take full responsibility for it. RMSNorm, root-mean-square normalization; SwiGLU, Swish-gated linear unit.

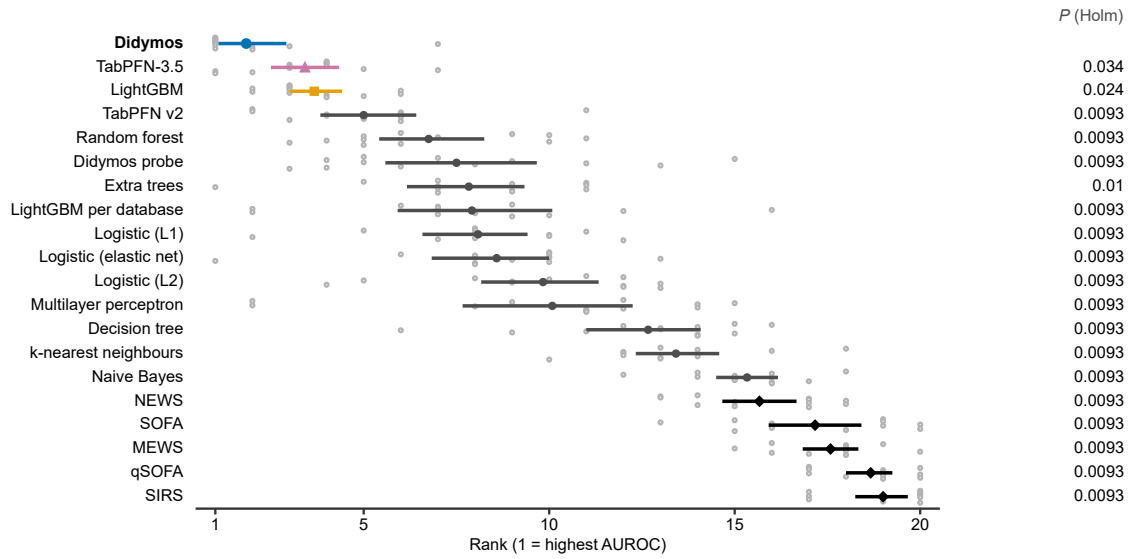


Supplementary Fig. 3 | Data and evaluation sets. The seven databases' adult ICU stays (age at least 18 years) were harmonized and tokenized (Supplementary Fig. 1), then split by patient within each database into train, tuning and held-out sets (80%, 5% and 15%). Didymos was trained on the train split without task labels; the comparators were trained on up to 5,000 training stays per database. The tuning split served model selection, the comparators' early stopping and the recalibration maps (100 rolling anchors per database). Every reported prediction time and label comes from the held-out test store, 2,000 stays per database. The pre-specified analyses used three anchor sets with 32 simulated futures per anchor: rolling anchors, 500 per database at random ICU hours, half enriched for an event in the next 48 hours and weighted by the inverse inclusion probability; and the 24- and 48-hour landmarks, up to 500 stays per database still in the ICU and alive at that time. The follow-up analyses 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 others, weighted to the whole landmark set). Every comparator was scored on the same anchors at the same prediction time. Artwork credited (GPT-image); the authors confirm its accuracy and scientific validity and take full responsibility for it.

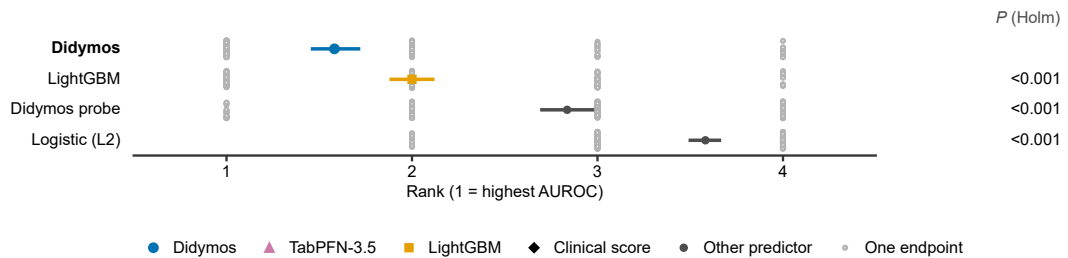
	Rolling anchors									24-hour landmark			
	Death, 24 h	New IMV, 24 h	New vasopressor, 24 h	New RRT, 48 h	AKI stage ≥2, 48 h	Sepsis, 24 h	Hypotension, 4 h	Hyperkalaemia, 24 h	Thrombocytopenia, 48 h	ICU death	In-hospital death	ICU stay >7 days	Death, 24 h
Didymos, hazard estimate	0.96	0.89	0.90	0.96						0.92			0.98
Didymos, sampled estimate	0.92	0.85	0.91	0.95	0.89	0.97	0.87	0.83	0.93	0.91	0.88	0.86	0.98
Didymos probe	0.95	0.79	0.74	0.93	0.82	0.95	0.86	0.75	0.86	0.90	0.87	0.85	0.97
LightGBM	0.96	0.81	0.85	0.93	0.83	0.95	0.90	0.81	0.92	0.91	0.86	0.85	0.96
LightGBM per database	0.94	0.79	0.83	0.90	0.84	0.70	0.85	0.82	0.91	0.90	0.85	0.82	
Logistic (L2)	0.94	0.75	0.82	0.88	0.82	0.94	0.87	0.72	0.93	0.86	0.82	0.76	0.86
TabPFN-3.5	0.96	0.80	0.84	0.93	0.83	0.96	0.89	0.81	0.91	0.93	0.88	0.86	
TabPFN v2	0.96	0.76	0.85	0.91	0.82	0.95	0.88	0.80	0.92	0.91	0.87	0.83	
Logistic (L1)	0.94	0.78	0.84	0.87	0.82	0.95	0.86	0.74	0.94	0.89	0.85	0.79	0.93
Logistic (elastic net)	0.94	0.77	0.83	0.90	0.82	0.94	0.85	0.74	0.94	0.89	0.85	0.82	0.92
Decision tree	0.86	0.64	0.77	0.80	0.80	0.86	0.88	0.72	0.90	0.80	0.77	0.75	0.81
Random forest	0.95	0.78	0.81	0.92	0.82	0.93	0.88	0.81	0.89	0.91	0.86	0.84	0.94
Extra trees	0.95	0.79	0.81	0.88	0.81	0.95	0.91	0.77	0.88	0.89	0.85	0.83	0.95
Naive Bayes	0.77	0.69	0.72	0.84	0.75	0.86	0.74	0.66	0.80	0.73	0.70	0.70	0.81
k-nearest neighbours	0.88	0.62	0.74	0.77	0.80	0.91	0.84	0.72	0.86	0.83	0.79	0.81	0.87
Multilayer perceptron	0.93	0.81	0.81	0.80	0.77	0.96	0.87	0.70	0.87	0.89	0.83	0.81	0.93
SOFA	0.75	0.68	0.66	0.81	0.65	0.52	0.70	0.64	0.79	0.70	0.69	0.70	0.80
NEWS	0.88	0.75	0.75	0.68	0.61	0.66	0.74	0.60	0.69	0.80	0.72	0.65	0.85
MEWS	0.85	0.62	0.69	0.61	0.60	0.72	0.71	0.59	0.66	0.79	0.70	0.62	0.86
qSOFA	0.80	0.60	0.69	0.60	0.58	0.58	0.74	0.53	0.67	0.73	0.65	0.61	0.77
SIRS	0.71	0.64	0.66	0.60	0.54	0.68	0.55	0.58	0.69	0.71	0.66	0.59	0.72
SAPS II										0.78	0.75	0.66	0.88
OASIS										0.73	0.71	0.69	0.85
KDIGO stage				0.82	0.62								
Creatinine				0.88	0.59								
MAP (lower)			0.68				0.84						
ROX (lower)		0.64											
Shock index	0.76		0.69										0.72
Lactate	0.70					0.39							0.74
Potassium								0.79					
Platelets (lower)									0.91				

Supplementary Fig. 4 | Every predictor's AUROC on the same anchors. Columns: the 13 primary endpoints on the rolling anchors (1,050) and the 24-hour landmark (700, case-cohort weighted). Rows, in four groups: Didymos, with the hazard estimate (on the six endpoints where it applies) and the sampled estimate from 128 simulated futures, and the Didymos probe, a ridge logistic regression on the model's last-layer state trained with labels; learned comparators trained on labelled real stays (LightGBM pooled with the database as a feature and per database; logistic regression with L2, L1 and elastic-net penalties; TabPFN-3.5 and TabPFN v2; a decision tree, random forest, extra trees, naive Bayes, *k*-nearest neighbours and a multilayer perceptron); clinical scores (SOFA, NEWS, MEWS, qSOFA and SIRS, and SAPS II and OASIS at the 24-hour landmark); and single bedside values chosen per endpoint, for which lower values count as higher risk for MAP, the ROX index and platelets. Cell colour runs from AUROC 0.5 (dark blue) to 1 (yellow) on the cividis scale; blank cells are not defined for the endpoint or not fitted. AUROC, area under the receiver operating characteristic curve; KDIGO, Kidney Disease: Improving Global Outcomes; MAP, mean arterial pressure; ROX, the ratio of oxygen saturation to inspired oxygen over respiratory rate.

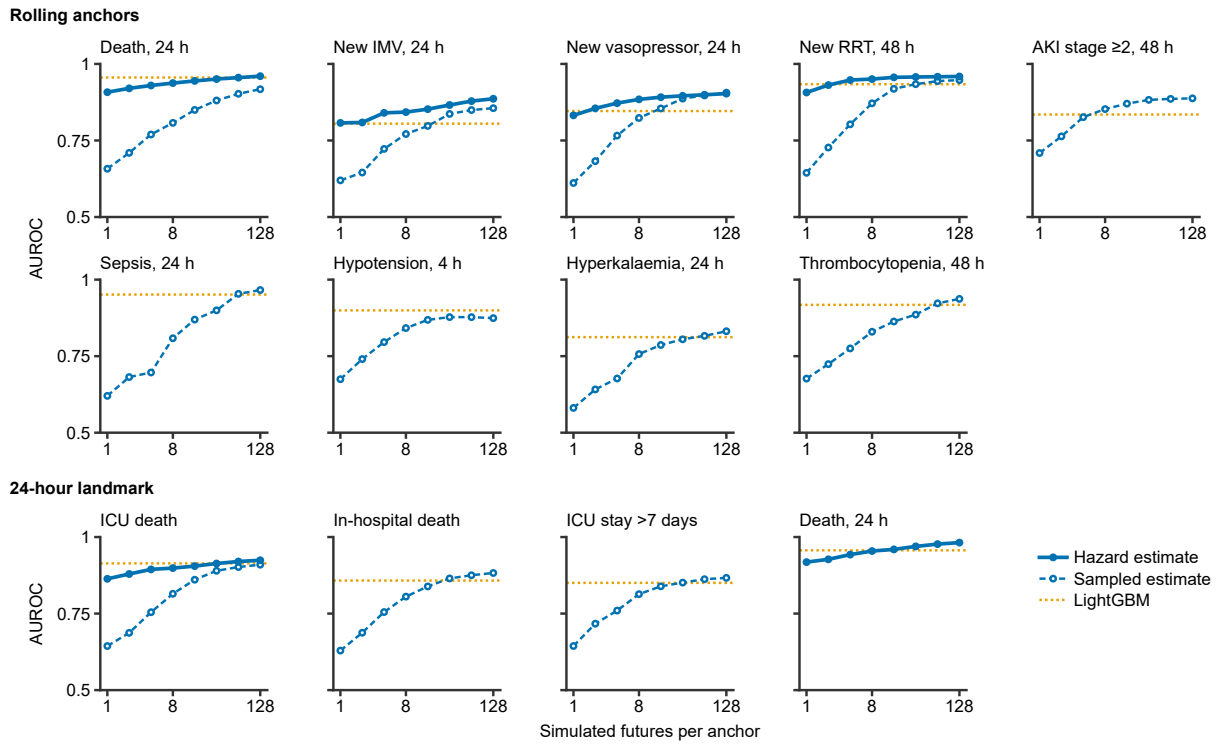
a Primary endpoints (12): mean rank among 20 predictors



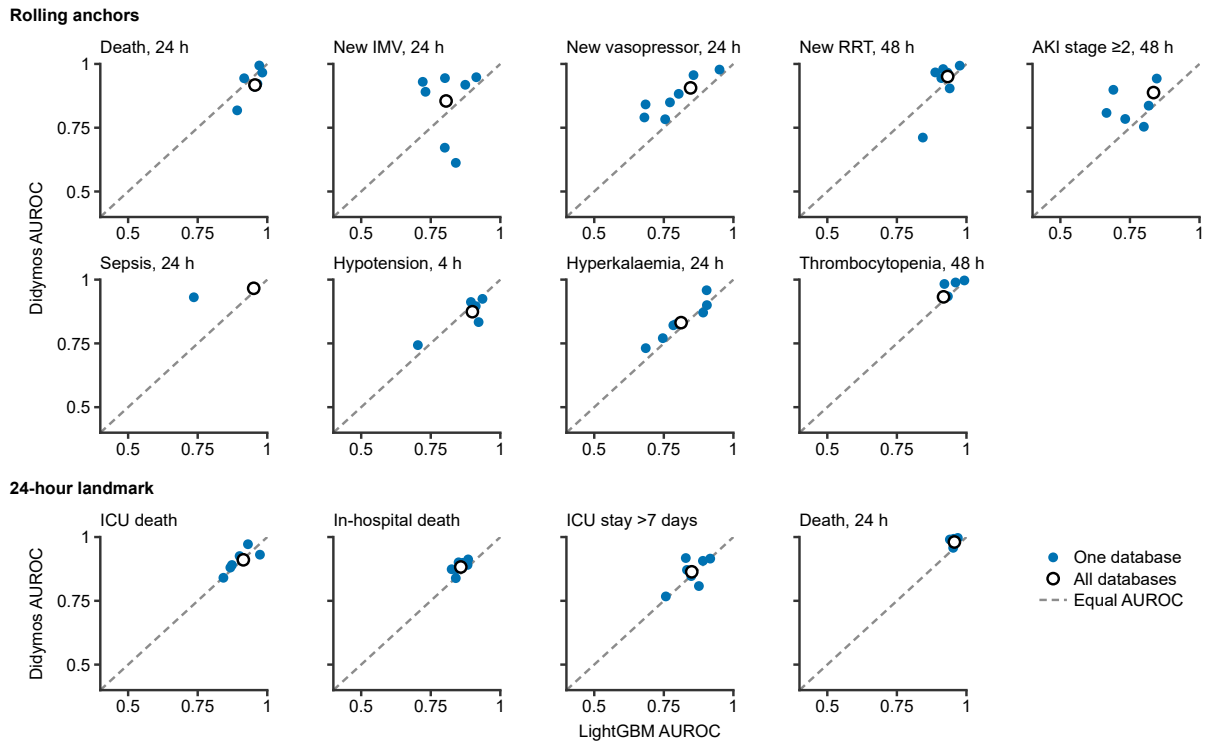
b Every scored endpoint (165): mean rank



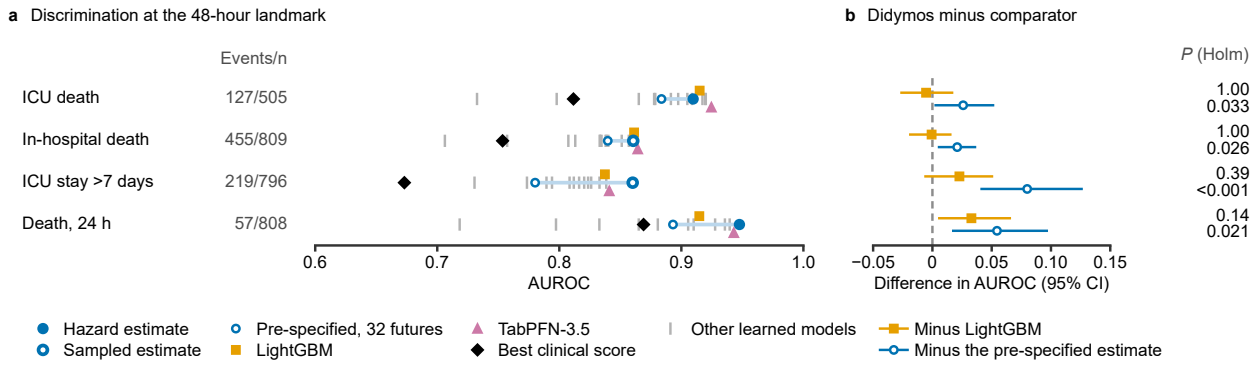
Supplementary Fig. 5 | Didymos ranks first across endpoints, among 20 predictors on the primary endpoints and among those scored on all 165. Ranks as machine-learning benchmarks report them (post hoc). On each endpoint, rank 1 is the highest AUROC, and ties take mid-ranks. Coloured marks are mean ranks with 95% intervals from 10,000 bootstrap resamples of the endpoints; small open circles are ranks on single endpoints. *P* (Holm): Didymos against each predictor, a two-sided Wilcoxon signed-rank test on the paired AUROC differences across endpoints, Holm-adjusted across the comparisons. **a**, The 12 primary endpoints on which every learned comparator was fitted, for 20 predictors: Didymos's current estimate (Fig. 2), the Didymos probe, 13 learned models trained with labels and five clinical scores; death within 24 hours at the landmark is left out because TabPFN-3.5 was not fitted there. Didymos's mean rank is 1.83 (95% interval 1.08–2.92), and it ranks first on 8 endpoints; it is ahead of TabPFN-3.5 on 9 endpoints, of LightGBM on 11 and of every clinical score on all 12 (Friedman test across the predictors, $P < 0.001$). The AUROCs are in Table 2. **b**, All 165 endpoints of the pre-specified analysis (Supplementary Fig. 9), for the four predictors scored on every one: Didymos (sampled estimate, 32 futures), LightGBM and logistic regression (both pooled over databases) and the Didymos probe. Didymos's mean rank is 1.58 (1.45–1.72); it ranks first on 102 endpoints and is ahead of LightGBM on 113 (Friedman test, $P < 0.001$). The endpoints share anchors and are correlated, so these tests describe consistency across tasks; they are not independent replications. AUROC, area under the receiver operating characteristic curve.



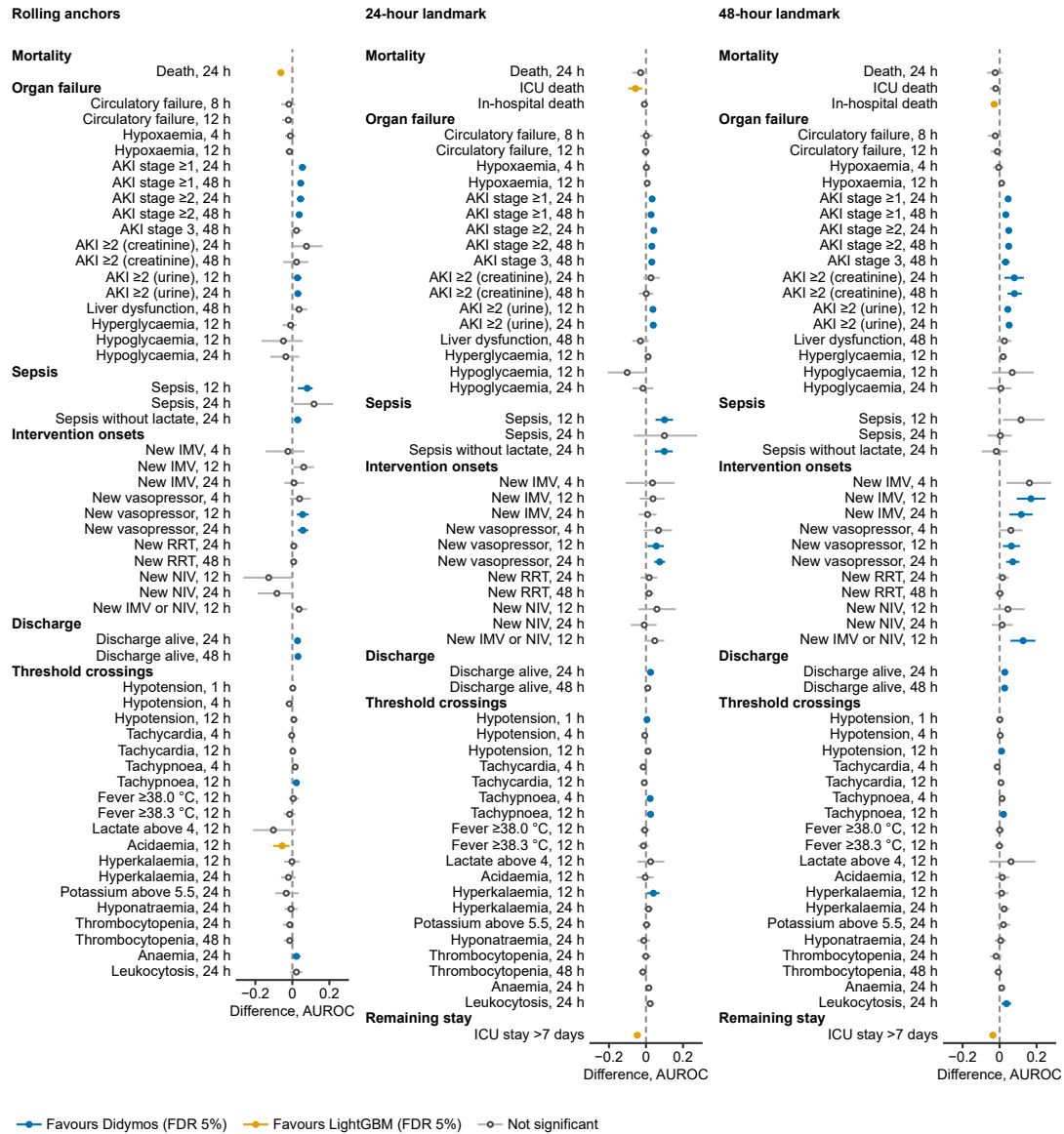
Supplementary Fig. 6 | Discrimination grows with the number of simulated futures, fastest for the sampled estimate. AUROC from $k = 1, 2, 4, \dots, 128$ simulated futures per anchor, each the mean over 20 random subsets of k futures, for the hazard estimate (solid lines, filled circles) on the six endpoints where it applies, the sampled estimate (dashed lines, open circles) and LightGBM on the same anchors (orange dotted lines). The anchors are the follow-up anchors of Supplementary Fig. 7. The sampled-estimate curve was not computed for death within 24 hours at the landmark. AUROC, area under the receiver operating characteristic curve.



Supplementary Fig. 7 | Discrimination by database: Didymos's sampled estimate against LightGBM. AUROC per database of Didymos's sampled estimate (128 simulated futures) against LightGBM's, for the 13 primary endpoints on the follow-up anchors: rolling anchors (1,050) and the 24-hour landmark (700, case-cohort weighted). Filled blue circles, one database (shown when it has at least 5 events and 5 non-events); open circles, all databases pooled; dashed line, equal AUROC (points above it favour Didymos). Sepsis is labelled in two databases, one of them with fewer than 5 events. AUROC, area under the receiver operating characteristic curve; IMV, invasive mechanical ventilation; RRT, renal replacement therapy; AKI, acute kidney injury.

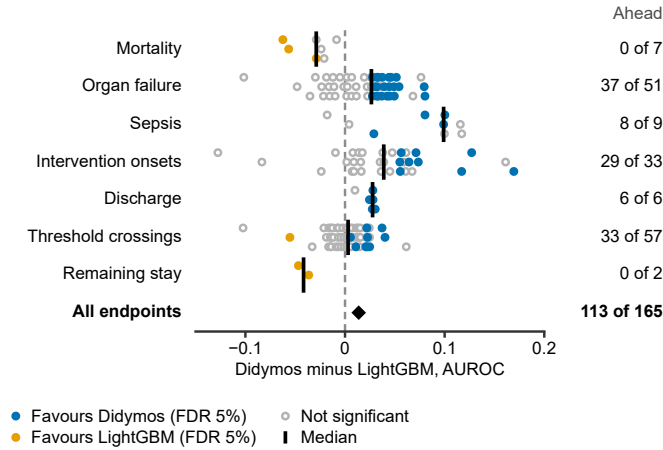


Supplementary Fig. 8 | The 48-hour landmark with the follow-up's settings. **a**, AUROC on a case-cohort sample of the pre-specified 48-hour landmark anchors (every in-hospital death, 455, and 354 other stays weighted to the landmark set; events/n per endpoint, where the endpoint is defined). Didymos's estimate from 128 simulated futures at four times the pre-specified token caps is shown as filled blue circles (hazard estimate, where it applies) or thick open circles (sampled estimate); its pre-specified estimate (32 futures, sampled) 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 the other learned models (logistic regressions, tree ensembles, k -nearest neighbours, a multilayer perceptron, naive Bayes, TabPFN v2, LightGBM per database and the Didymos probe). **b**, Didymos minus LightGBM (orange squares) and minus its pre-specified estimate (open blue circles), 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 four endpoints (upper value, LightGBM; lower, the pre-specified estimate). AUROC, area under the receiver operating characteristic curve.

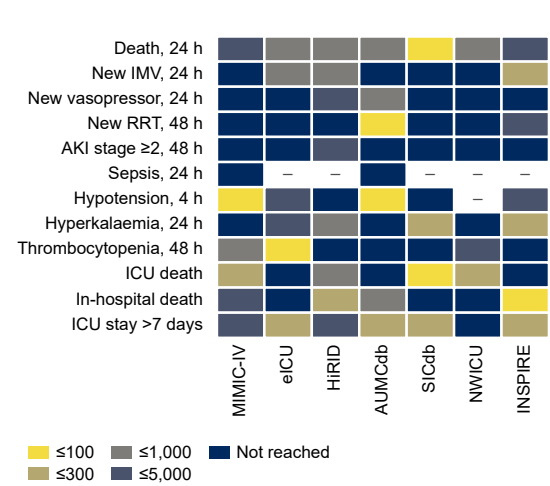


Supplementary Fig. 9 | Didymos against LightGBM on every endpoint the pre-specified analysis scored. The 165 comparisons on the three pre-specified anchor sets: rolling anchors (500 per database, 53 endpoints) and the 24- and 48-hour landmarks (up to 500 stays per database each, 56 endpoints each), the pre-registered primary and secondary endpoints, grouped by the catalog's families. Didymos's sampled estimate uses 32 simulated futures; LightGBM was trained on labelled stays; both are scored on the anchors every predictor scored. Points are differences in AUROC with 95% CIs from a paired bootstrap over stays. Filled points have a Benjamini–Hochberg q value of at most 0.05 across the 165 comparisons (exploratory, defined after the per-endpoint intervals were known): blue favours Didymos, orange LightGBM; P values come from 10,000 paired resamples. Didymos is ahead on 113 of 165 (68%); joint-bootstrap 95% CI 62–73%; 49 comparisons favour it at a 5% false-discovery rate and 6 favour LightGBM, and under Holm, 30 and 3. On the mortality endpoints, the sampled estimate at 32 futures sits below LightGBM; the follow-up (128 futures, with the hazard estimate where it applies) brings Didymos level with LightGBM or ahead on the rolling anchors and the 24-hour landmark (Fig. 2), and at the 48-hour landmark when re-run (Supplementary Fig. 8), pointing to Monte Carlo noise and simulation length. The comparisons are listed in Supplementary Data 1. AUROC, area under the receiver operating characteristic curve.

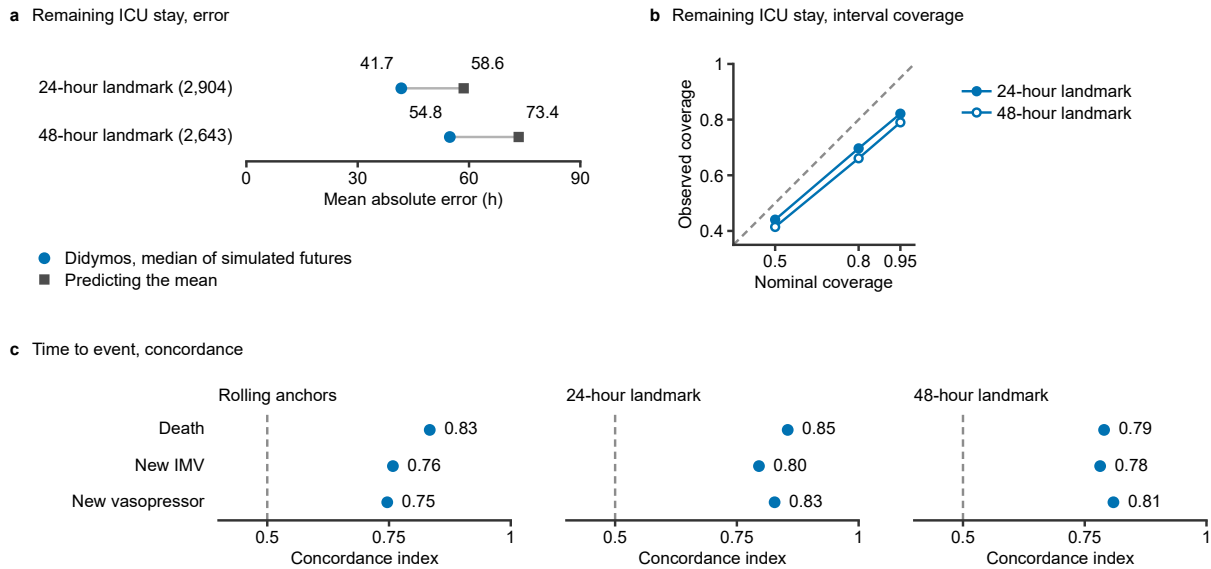
a Every scored endpoint by clinical family



b Labelled stays needed to match Didymos

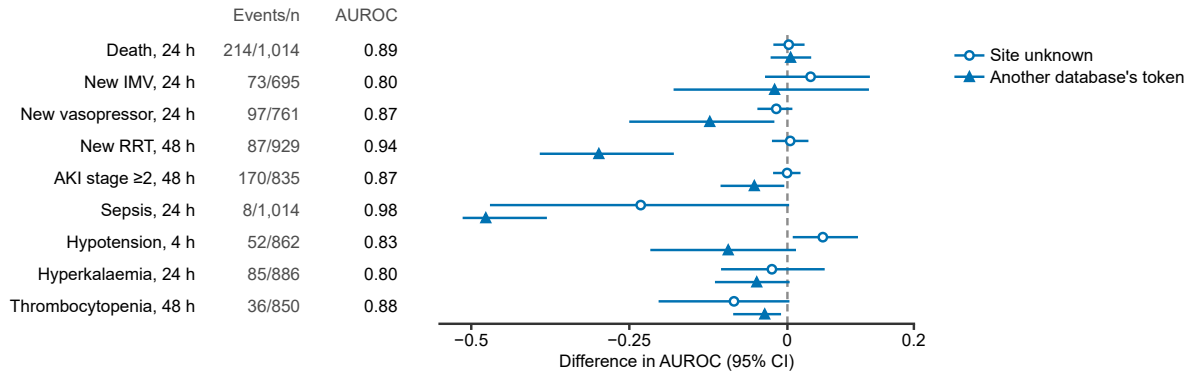


Supplementary Fig. 10 | Didymos’s advantage by clinical family, and the labelled stays a single-database model needs to match it. **a**, The comparisons of Supplementary Fig. 9 by catalog family. Points are endpoints on an anchor set, coded as in Supplementary Fig. 9; black ticks are family medians; the count of comparisons in which Didymos is ahead is at the right. Across all 165, the median difference is +0.014 AUROC (joint-bootstrap 95% CI 0.010–0.019; black diamond). **b**, Label efficiency for each pre-registered primary endpoint and database: the number of a database’s own labelled training stays at which a LightGBM trained in that database alone first reaches Didymos’s AUROC there (pre-specified; interpolation on $\log n$ fixed before it was computed). The model was trained on 100, 300, 1,000, 3,000 and all sampled training stays, up to 5,000. “Not reached”: below Didymos even with all of them; blank: no estimate (the endpoint not labelled in that database, or too few events). In 50 of 78 endpoint–database cells, more than 1,000 labelled stays were needed or Didymos was never reached; in 38, never. Per-database estimates use up to 500 held-out anchors each and carry no interval.

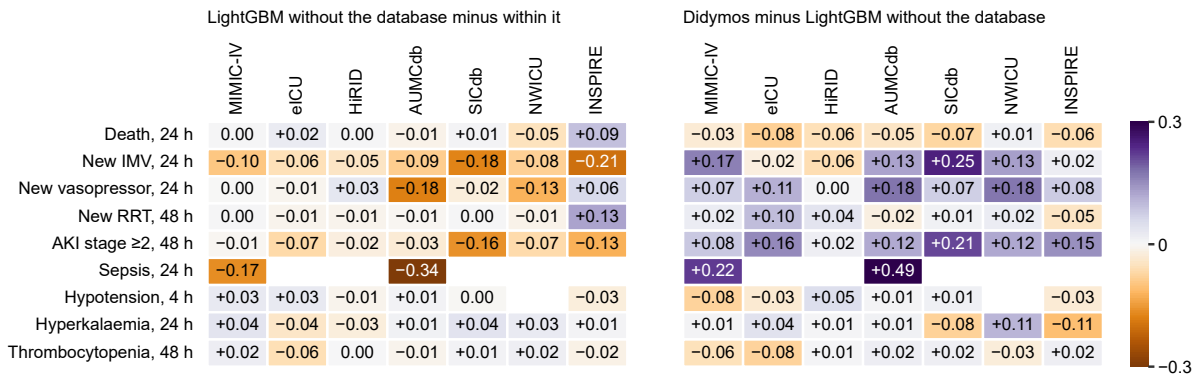


Supplementary Fig. 11 | Remaining ICU stay and time to event in the pre-specified analysis. All panels use 32 simulated futures per anchor. **a**, Mean absolute error of the remaining ICU stay at the 24- and 48-hour landmarks (anchors in brackets): Didymos, the median of the simulated remaining stays (blue circles), and predicting the mean (grey squares). **b**, Coverage of the central 50%, 80% and 95% intervals of the simulated remaining stay, against nominal (dashed line); below the line, the simulated spread is narrower than the observed one. **c**, Harrell's concordance index for death, new IMV and new vasopressor on the three pre-specified anchor sets; the risk score is the share of simulated futures with the event within 48 hours, and the dashed line marks chance. IMV, invasive mechanical ventilation.

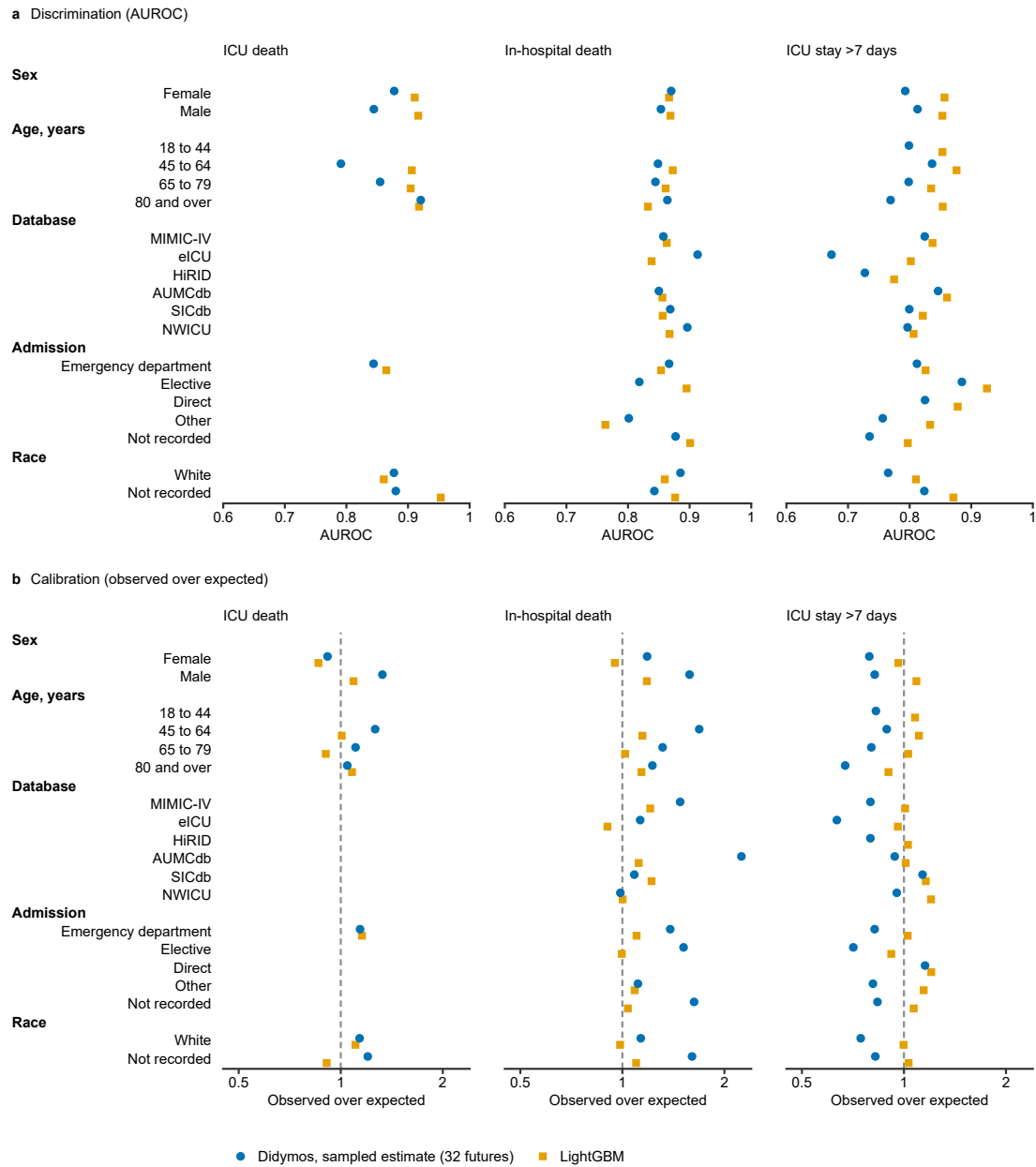
a Site prompt



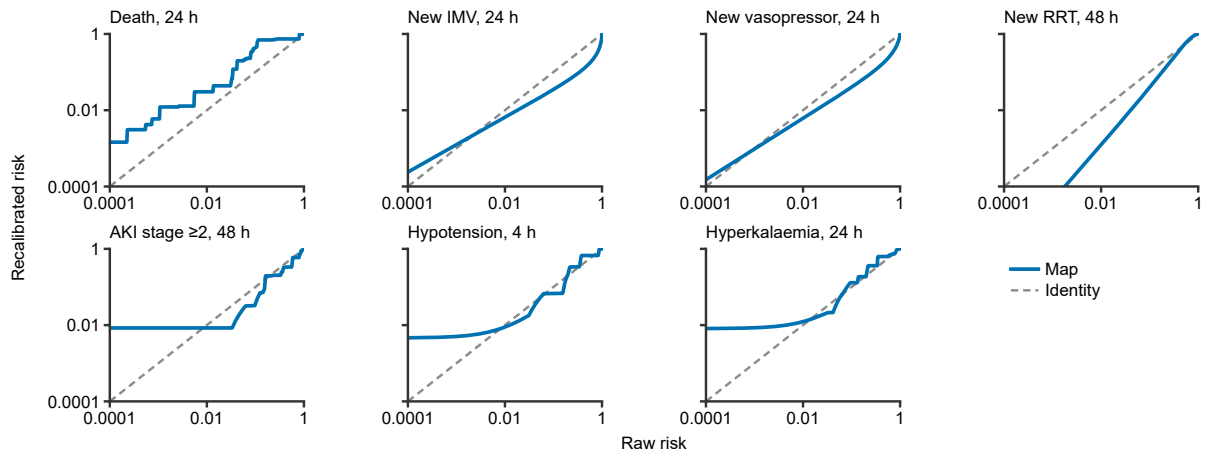
b Leaving a database out



Supplementary Fig. 12 | The site prompt, and supervised transfer to a database left out of training. **a**, On the follow-up rolling anchors, with 32 simulated futures per prompt: AUROC with the database and site tokens unknown (open circles), or with the database token swapped for another database's (triangles), minus AUROC with the recorded tokens, with 95% CIs from a paired bootstrap; the recorded-token AUROC and events/n are listed. **b**, Leaving a database out, on the pre-specified rolling anchors (500 per database). Left, LightGBM trained on the six other databases minus LightGBM trained within the database; right, Didymos's sampled estimate (32 futures, no labels) minus LightGBM trained without the database. Purple cells are positive, orange negative; blank cells have fewer than 5 events in that database. Didymos saw every database's training stays without labels, so neither panel tests transfer of the generative model to a new site. AUROC, area under the receiver operating characteristic curve.

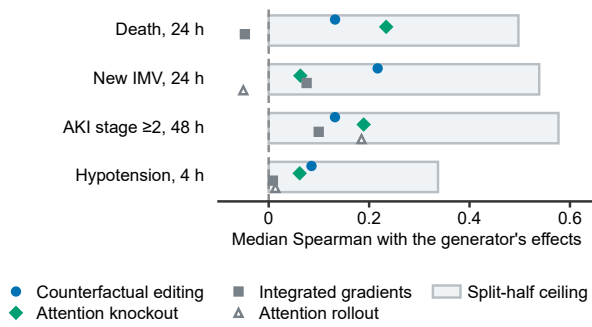


Supplementary Fig. 13 | Subgroups in the pre-specified analysis at the 24-hour landmark. a, AUROC and b, observed over expected (log scale; dashed line, 1) for ICU death, in-hospital death and an ICU stay over 7 days: Didymos's sampled estimate from 32 simulated futures, before recalibration (blue circles), and LightGBM (orange squares). Subgroups are sex, age, database, admission type and race. A subgroup is scored when it has at least 30 events and 30 non-events, and is blank otherwise. Race is recorded in MIMIC-IV, eICU-CRD and NWICU only, so "not recorded" includes the other databases. Events and n per cell are in the source data (Supplementary Data 3). AUROC, area under the receiver operating characteristic curve.

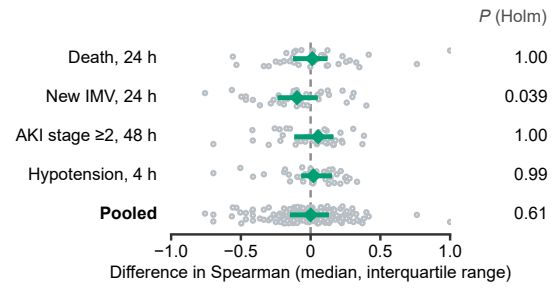


Supplementary Fig. 14 | The recalibration maps in force. Raw against recalibrated risk (log scales; dashed line, identity) for the seven endpoints whose maps were adopted. The maps were fitted on the tuning split's rolling anchors (100 per database, 32 futures): isotonic for death within 24 hours, AKI, hypotension and hyperkalaemia, and logistic (a linear map of the logit) for the three treatment starts. At low raw risk, the RRT map falls below the axis range. AKI, acute kidney injury; IMV, invasive mechanical ventilation; RRT, renal replacement therapy.

a Agreement with the generator's effects

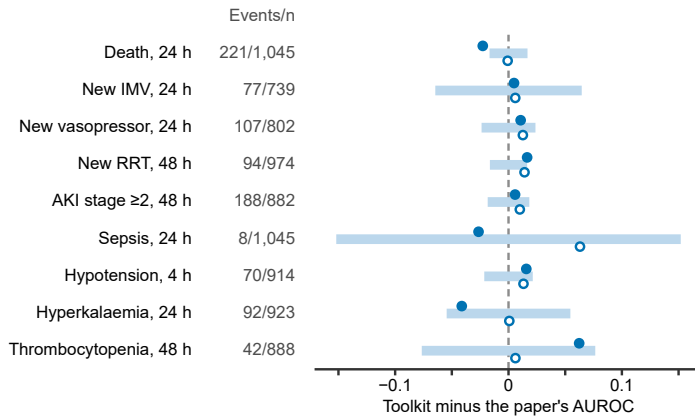


b Per anchor, knockout minus editing

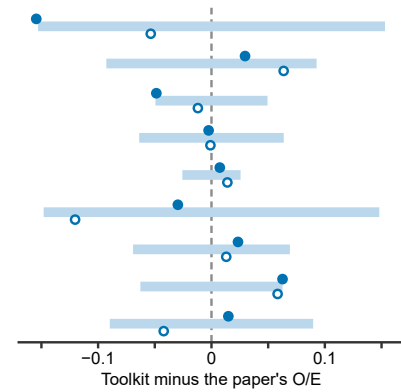


Supplementary Fig. 15 | Attention knockout agrees with the generator's own effects as well as counterfactual editing. For 40 anchors per endpoint from the higher-risk half, each group's edit was simulated again with 48 new futures and a new seed, giving the generator's own effect of each group. **a**, Median Spearman correlation, across anchors, between each method's group scores and the redrawn effects: counterfactual editing (blue circles), attention knockout (green diamonds), integrated gradients (grey squares) and attention rollout (open triangles). Bars show the split-half reliability of the generator's effects (Spearman–Brown corrected), which bounds any agreement. **b**, Per anchor, knockout's minus editing's Spearman correlation with the redrawn effects (small circles; 158 anchors scored by both), with the median (diamond) and interquartile range. Two-sided Wilcoxon signed-rank tests, Holm-adjusted across the four endpoints; the pooled test is unadjusted. IMV, invasive mechanical ventilation; AKI, acute kidney injury.

a Discrimination (AUROC)

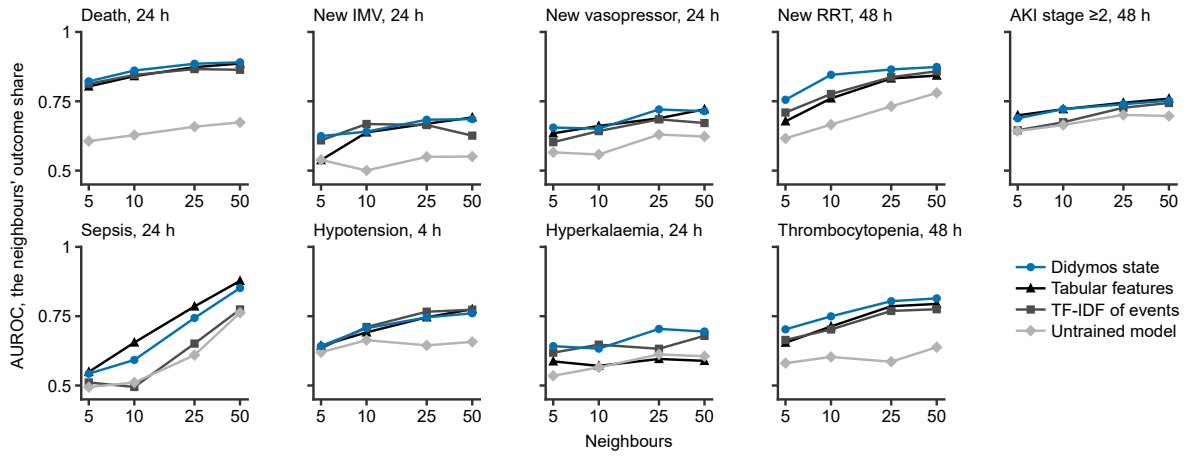


b Calibration (observed over expected)

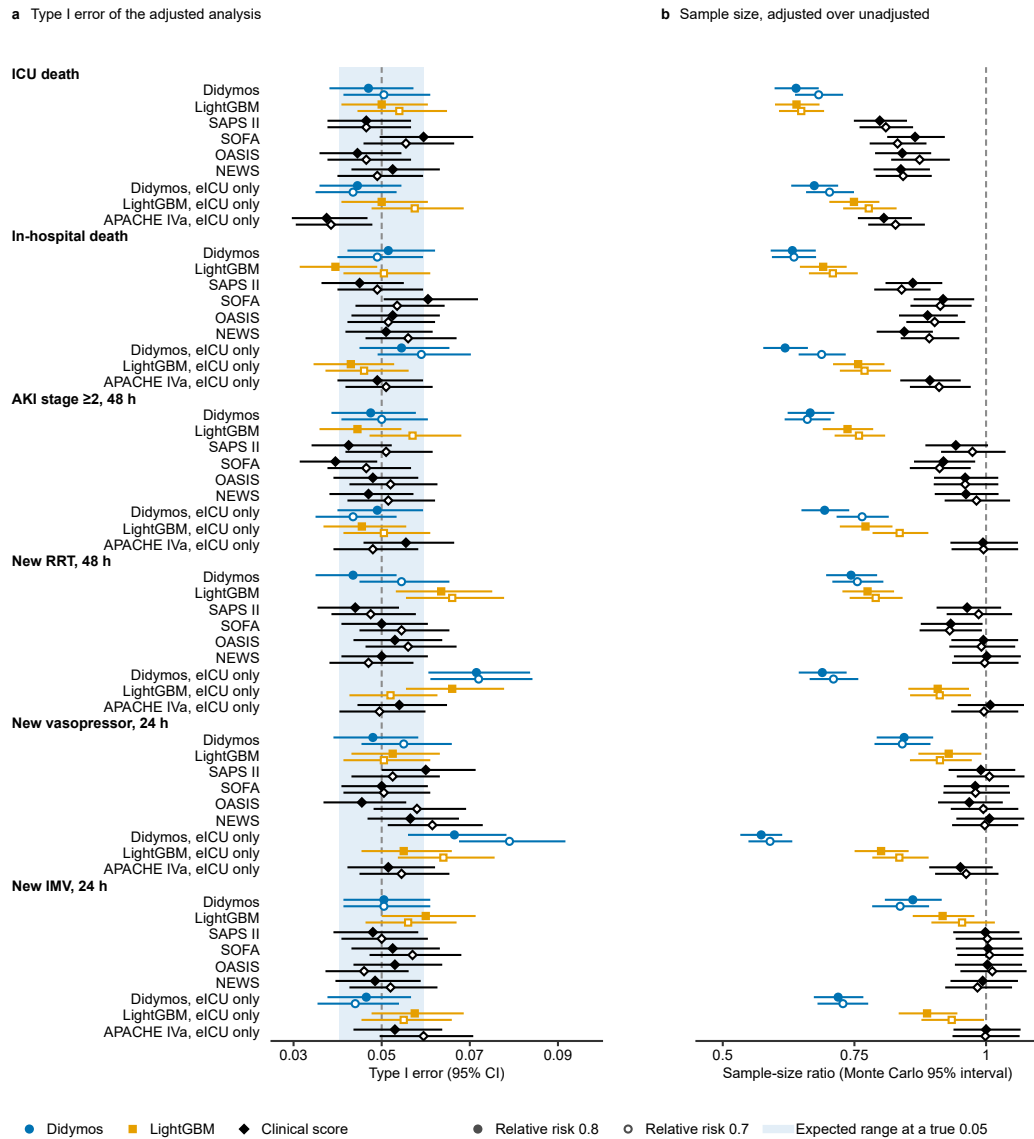


● Against the pre-specified run ○ Against the follow-up run — Rollout-noise band

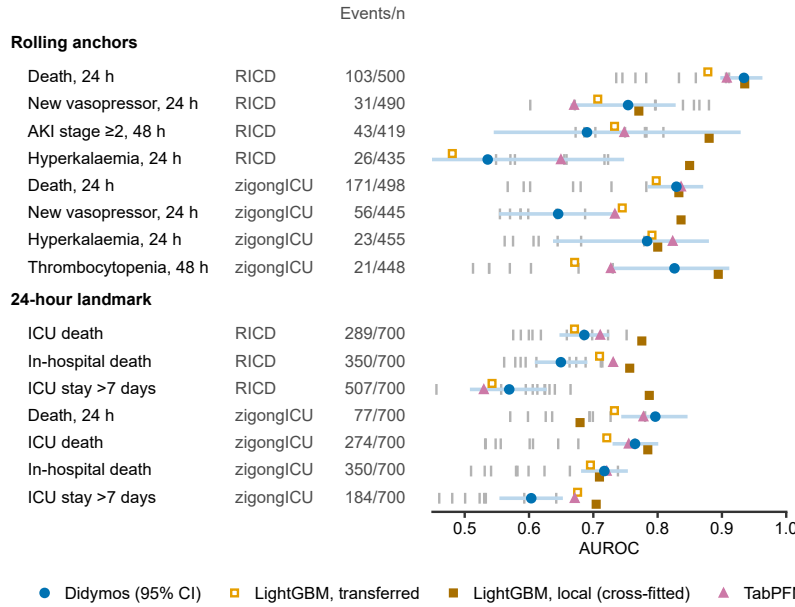
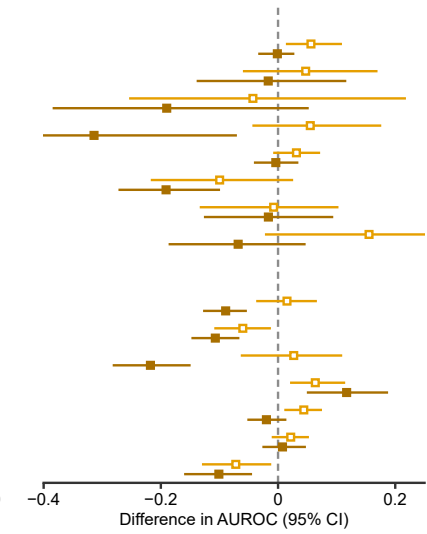
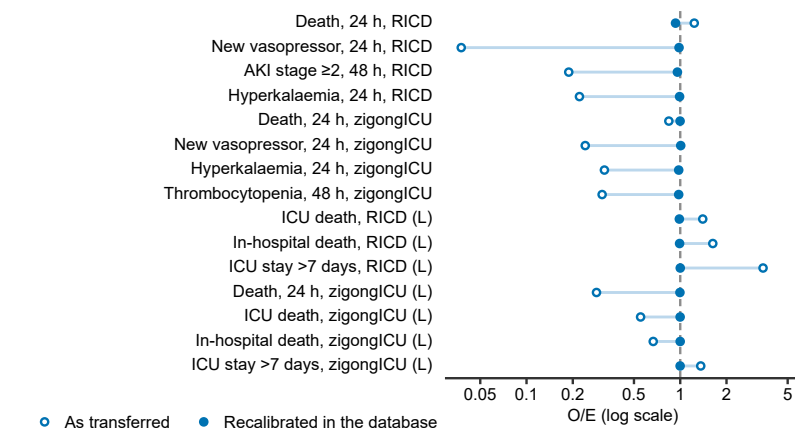
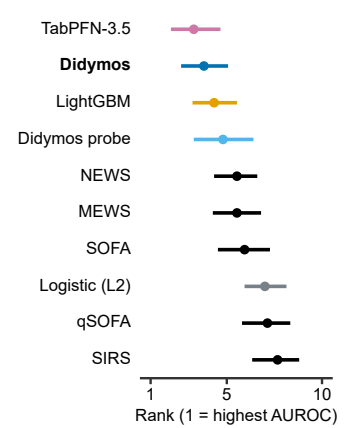
Supplementary Fig. 16 | The toolkit reproduces the paper's forecasts within simulation noise of a paper run. The toolkit's Forecast with new draws (32 simulated futures) on the follow-up rolling anchors, minus each of the paper's two runs on the same anchors: the pre-specified run (32 futures; filled circles) and the follow-up run (its 32-future curve for AUROC and its 128-future O/E; open circles). **a**, AUROC; **b**, observed over expected. The band is plus or minus $1.96 \sqrt{2}$ times the standard deviation of the toolkit's estimate when futures are resampled within anchors (1,000 resamples): the spread expected between two independent runs. The band was defined after the pre-specified rule's result was known; that rule and its result are reported in the Methods and Supplementary Table 7. O/E, observed over expected.



Supplementary Fig. 17 | Nearest neighbours as a predictor, by the representation used to find them. The AUROC of the share of a patient's k nearest neighbours with the event, used as a risk score ($k = 5$ to 50 ; leave one out, on the pre-specified rolling anchors, with neighbours from all databases). Neighbours were found by Didymos's last-layer state (blue circles), the comparators' tabular features (black triangles), TF-IDF of the prompt's event tokens (dark grey squares) and the state of an untrained model of the same architecture, a negative control (light grey diamonds). TF-IDF, term frequency–inverse document frequency; AUROC, area under the receiver operating characteristic curve.

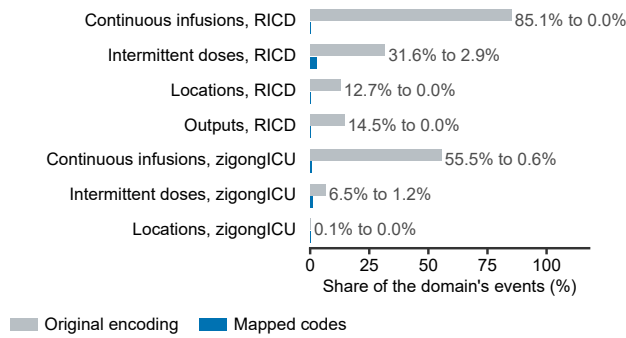


Supplementary Fig. 18 | Trial simulations: type I error and sample size. The trials draw anchors with replacement from the 24-hour landmark, for each endpoint and prognostic score: ICU and in-hospital death from the case-cohort sample, the other endpoints from the natural sample. They use all databases, or eICU-CRD only with APACHE IVa, eICU-CRD's own score. Didymos's score is the share of simulated futures with the event (128 futures for the two deaths, 32 for the others). **a**, Type I error of the covariate-adjusted analysis with no treatment effect, at the sample sizes designed for relative risks of 0.8 (filled markers) and 0.7 (open markers): 2,000 simulated trials each, with exact (Clopper–Pearson) 95% CIs; the shaded band is the range expected from 2,000 trials at a true 0.05, and the dashed line marks 0.05. **b**, The sample size an adjusted analysis needs relative to an unadjusted one (80% power, two-sided 0.05), with Monte Carlo 95% intervals from a parametric bootstrap of the two simulated powers (10,000 draws); the dashed line marks no reduction. Didymos, blue circles; LightGBM, orange squares; clinical scores, black diamonds. The landmark population is a case-cohort sample for the two death endpoints, so their sizes are not population values. APACHE, Acute Physiology and Chronic Health Evaluation.

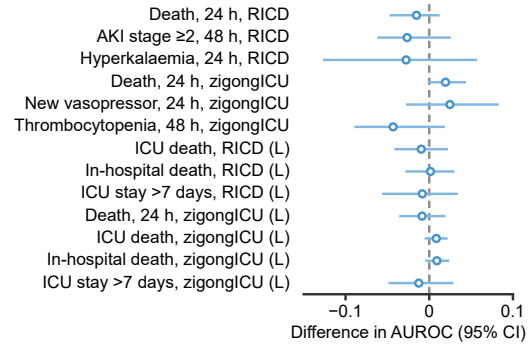
a Discrimination on every transfer anchor**b** Didymos minus LightGBM**c** Observed over expected, before and after recalibration**d** Mean rank, transferred models

Supplementary Fig. 19 | Zero-shot transfer to two databases never seen in training. RICD (Moscow, Russia; intensive care and neurorehabilitation) and zigongICU (Zigong, China; ICU patients with infection) were encoded with the frozen vocabulary, and prompts carried the database and site as unknown. **a**, AUROC on every transfer anchor (rolling anchors, 998; a case-cohort sample of the 24-hour landmark, 1,400; 128 simulated futures): Didymos with 95% CI (blue circles; the hazard estimate for death within 24 hours, a new vasopressor and ICU death, the sampled estimate otherwise), LightGBM trained on the seven databases (open orange squares), a local LightGBM trained within the database and five-fold cross-fitted by patient (filled dark orange squares), TabPFN-3.5 trained on the seven databases (purple triangles) and the other models and clinical scores (grey ticks). The endpoints shown met the availability rule fixed before any future was drawn (a future timed to the hour and at least 20 positive anchors). **b**, Didymos minus transferred (open) and local (filled) LightGBM, with 95% CIs from a paired bootstrap over stays (1,000 resamples). **c**, Observed over expected (log scale; dashed line, 1), as transferred (open circles) and after a logistic recalibration within the database, cross-fitted by patient (filled circles); (L), 24-hour landmark. **d**, Mean rank by AUROC over the 15 pairs among the predictors that never saw the new databases (post hoc), with 95% intervals from 10,000 bootstrap resamples of the pairs. AUROC, area under the receiver operating characteristic curve; AKI, acute kidney injury.

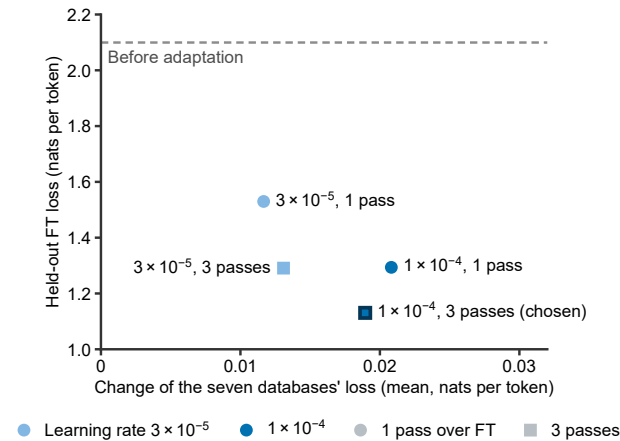
a Events without a vocabulary token



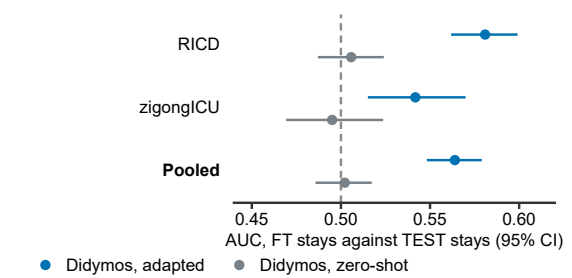
b Mapped codes minus original, zero-shot



c The training grid



d Membership signal of the FT stays



Supplementary Fig. 20 | Adaptation: mapped codes, the training grid and the membership signal of the adapted stays. **a**, Share of each domain's events encoded as an unknown token in the original encoding (grey) and after mapping each code that the vocabulary lacks to its nearest kept code (blue); domains with unknown events are shown. **b**, AUROC with mapped codes minus with the original encoding, both zero-shot, on the locked test anchors (the 13 endpoint–database pairs with at least 10 test positives; (L), 24-hour landmark), with 95% CIs from a paired bootstrap over stays. **c**, The four cells of the grid fixed in advance (learning rate by passes over the unlabelled stays): next-token loss on the held-out 10% of the unlabelled patients against the mean change of the loss on the seven databases' tuning split. The cell with the lowest held-out loss was chosen (dark outline); the dashed line marks the loss before adaptation. **d**, Membership signal of the stays used for continued training: AUC of the loss-based membership attack separating them (members) from the test stays (non-members), for adapted Didymos (blue) and for Didymos before adaptation, which saw neither (grey), with 95% CIs; the dashed line marks 0.5. AUROC, area under the receiver operating characteristic curve; AUC, area under the curve.

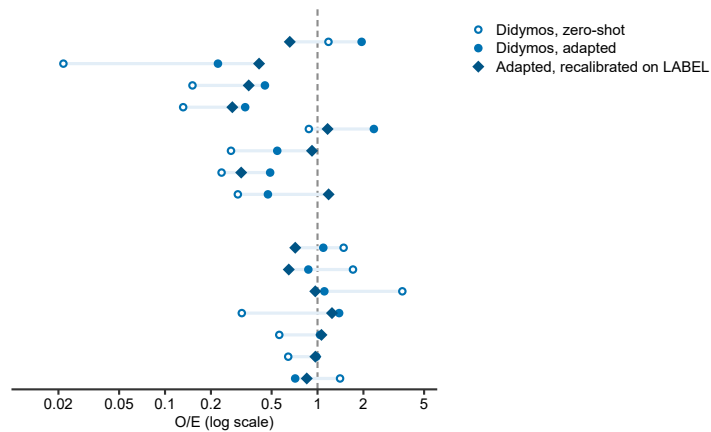
a Observed over expected on the TEST anchors

Rolling anchors

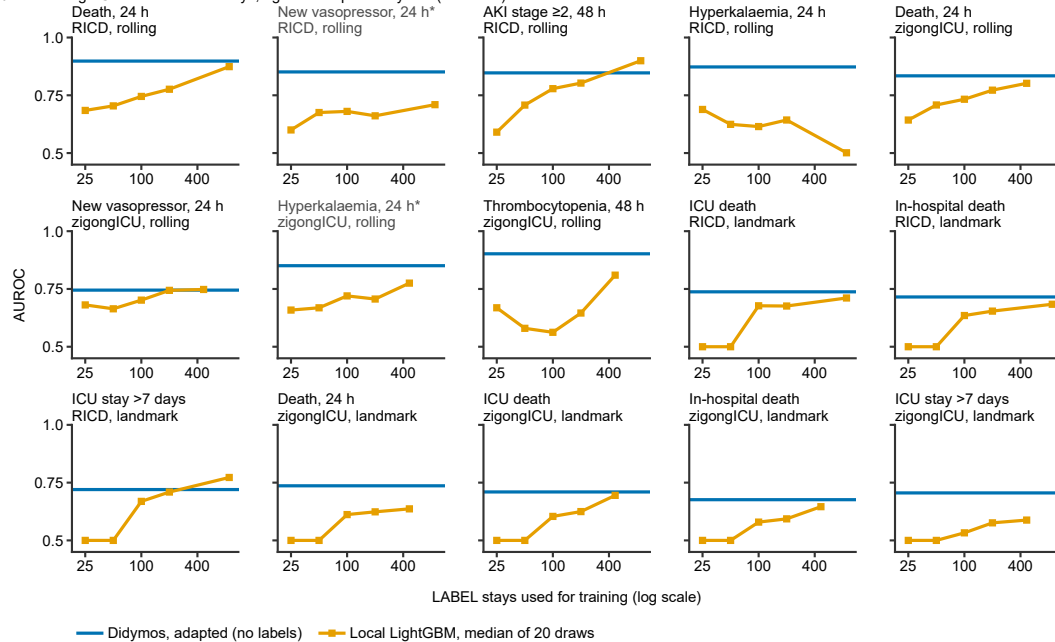
Death, 24 h	RICD
New vasopressor, 24 h*	RICD
AKI stage ≥ 2 , 48 h	RICD
Hyperkalaemia, 24 h	RICD
Death, 24 h	zigongICU
New vasopressor, 24 h	zigongICU
Hyperkalaemia, 24 h*	zigongICU
Thrombocytopenia, 48 h	zigongICU

24-hour landmark

ICU death	RICD
In-hospital death	RICD
ICU stay >7 days	RICD
Death, 24 h	zigongICU
ICU death	zigongICU
In-hospital death	zigongICU
ICU stay >7 days	zigongICU



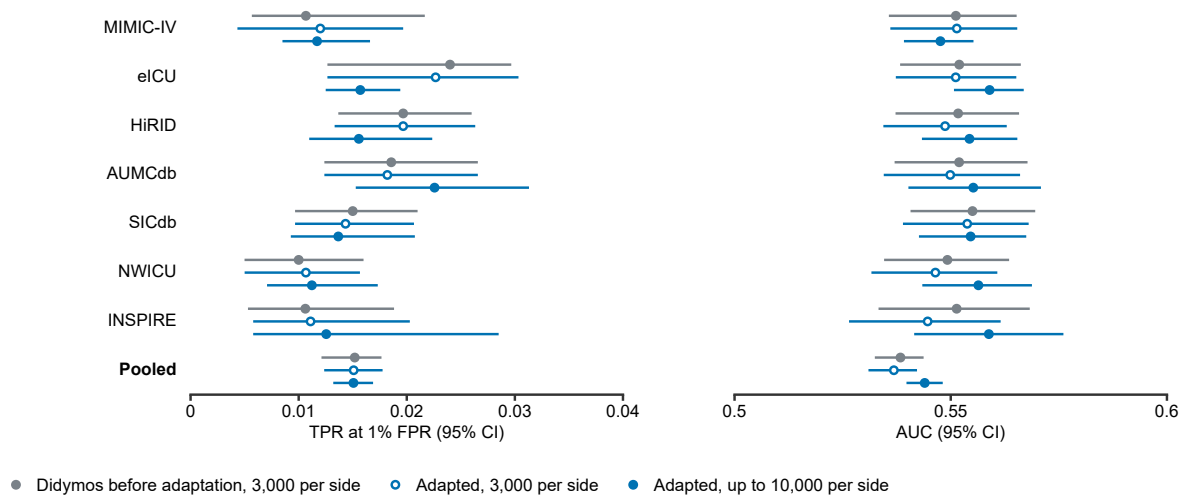
b A local LightGBM on n LABEL stays, against adapted Didymos (no labels)



Supplementary Fig. 21 | Calibration after adaptation, and the labelled stays a local model needs. **a**, Observed over expected on the locked test anchors (log scale; dashed line, 1) for Didymos zero-shot (open circles), adapted (filled circles) and adapted and recalibrated on the labelled slice (diamonds; a logistic map of the logit risk, or its intercept alone below 10 labelled positives). Asterisks mark endpoints with fewer than 10 test positives. **b**, Per endpoint and database, AUROC on the test anchors of a local LightGBM trained on n stays of the labelled slice (orange; median of 20 random draws, and all labelled stays at the right) against adapted Didymos, trained without labels (blue line). At the landmark each stay gives one training row, so 25 or 50 stays are too few for LightGBM to split, and its AUROC is 0.5 there. AUROC, area under the receiver operating characteristic curve; AKI, acute kidney injury.

a True-positive rate at 1% false positives

b Area under the curve



Supplementary Fig. 22 | The adapted model's membership test on the seven training databases. **a**, True-positive rate at a 1% false-positive rate and **b**, AUC of the loss-based membership attack (training stays as members, held-out stays as non-members), with 95% bootstrap CIs. Grey, Didymos before adaptation (3,000 stays per side); open blue circles, the adapted model at the registered size (3,000 per side); filled blue circles, the adapted model with up to 10,000 per side, a size fixed before it ran and limited by each database's held-out stays (2,747 in AmsterdamUMCdb and 2,072 in INSPIRE). The pre-specified bounds, their outcome and the study lead's decision are reported in the Methods and Supplementary Table 7. AUC, area under the curve.

Supplementary Tables

Supplementary Table 1 | Headline results beside their pre-specified values. Follow-up anchors: rolling anchors (1,050; 128 simulated futures) and the 24-hour landmark (700, case-cohort weighted; 128 futures with four times the pre-specified token cap). “Didymos” is the current estimate (H, hazard estimate; S, sampled estimate). Differences are paired, with 95% CIs from a bootstrap over stays. The pre-specified value is Didymos’s sampled estimate from 32 futures on the same anchors. Below, the 48-hour landmark in the pre-specified analysis (32 futures, against LightGBM); its re-run with the follow-up’s settings is shown in Supplementary Fig. 8.

Endpoint	Didymos	LightGBM	TabPFN-3.5	– LightGBM (95% CI)	– TabPFN-3.5 (95% CI)	Pre-specified
<i>Rolling anchors</i>						
Death, 24 h	0.960 H	0.956	0.955	+0.004 (–0.006 to 0.015)	+0.005 (–0.005 to 0.015)	0.903
New IMV, 24 h	0.886 H	0.805	0.797	+0.081 (0.029 to 0.152)	+0.089 (0.026 to 0.189)	0.837
New vasopressor, 24 h	0.903 H	0.846	0.842	+0.057 (0.017 to 0.096)	+0.061 (0.021 to 0.102)	0.889
New RRT, 48 h	0.959 H	0.934	0.933	+0.025 (0.005 to 0.048)	+0.026 (–0.002 to 0.056)	0.931
AKI stage ≥2, 48 h	0.888 S	0.835	0.835	+0.053 (0.022 to 0.085)	+0.053 (0.018 to 0.089)	0.887
Sepsis, 24 h (8 events)	0.966 S	0.951	0.959	+0.015 (–0.063 to 0.081)	+0.007 (–0.011 to 0.024)	0.989
Hypotension, 4 h	0.874 S	0.900	0.894	–0.025 (–0.064 to 0.005)	–0.020 (–0.057 to 0.007)	0.875
Hyperkalaemia, 24 h	0.831 S	0.812	0.807	+0.019 (–0.031 to 0.077)	+0.025 (–0.022 to 0.071)	0.847
Thrombocytopenia, 48 h	0.932 S	0.918	0.910	+0.015 (–0.002 to 0.031)	+0.022 (0.003 to 0.044)	0.829
<i>24-hour landmark</i>						
ICU death	0.924 H	0.914	0.932	+0.010 (–0.013 to 0.031)	–0.008 (–0.028 to 0.010)	0.853
In-hospital death	0.883 S	0.858	0.876	+0.025 (0.007 to 0.043)	+0.006 (–0.010 to 0.023)	0.856
ICU stay >7 days	0.864 S	0.850	0.864	+0.013 (–0.016 to 0.044)	–0.000 (–0.035 to 0.031)	0.843
Death, 24 h	0.982 H	0.957	–	+0.025 (0.010 to 0.044)	–	–
<i>48-hour landmark (pre-specified only)</i>						
ICU death	0.882 S			–0.021 (–0.053 to 0.010)		
In-hospital death	0.830 S			–0.029 (–0.049 to –0.010)		
ICU stay >7 days	0.797 S			–0.036		

Supplementary Table 2 | Didymos beside the closest generative models of health records and intensive care twins. Each row summarizes what the cited work reports in its abstract; the columns describe scope, not quality. Preprints are marked. ICU, intensive care unit; LLM, large language model.

Model (year)	Data and setting	Approach	What is evaluated
ETHOS (2024) ²¹	MIMIC-IV; one US centre	Transformer over tokenized patient health timelines; predictions by simulating futures, without fine-tuning	Zero-shot inference across tasks
ETHOS-ARES (2025) ²²	MIMIC-IV with its emergency department extension; 285,622 timelines	Dynamic, personalized risk for clinician-defined events, with an explainability module	Hospital and ICU admission and prolonged stay against early warning scores and machine-learning models; subgroups and calibration
Foresight (2024) ²³	Two UK hospital datasets and MIMIC-III; 811,336 patients	Free text converted to coded concepts, then a generative transformer	Forecasts of the next disorder or concept; clinician review of synthetic timelines
Delphi-2M (2025) ²⁴	UK Biobank (0.4 million participants); external validation in Danish registries (1.9 million)	Generative transformer for the incidence of more than 1,000 diseases	Disease rates against single-disease models; synthetic trajectories; explanations and learnt biases
CoMET (2025, preprint) ²⁵	Epic Cosmos, 310 US health systems; pretraining on 118 million patients	Decoder-only medical event models up to 1 billion parameters; scaling laws	78 tasks, matched or exceeded task-specific supervised models without fine-tuning
DT-GPT (2025) ²⁶	Lung cancer, ICU (MIMIC-IV) and Alzheimer's disease datasets	Fine-tuned LLM forecasting clinical variables	Forecast error against machine-learning models; distributions and correlations of variables; zero-shot forecasting
DT-ICU (2026, preprint) ²⁷	MIMIC-IV	Multimodal, multitask model updated over the stay	Continuous risk estimation against baselines; modality ablations
EHR-MPC (2026, preprint) ²⁸	Sepsis cohort from 8 hospitals of one US health system	Generative EHR twin with model predictive control	Off-policy and simulation-based evaluation of treatment policies
Mechanistic twin (2020) ²⁹	One US medical ICU; 29 patients prospectively	Causal graph of organ systems and treatments for the first 24 hours of sepsis	Agreement of predicted and observed treatment responses
Didymos (this study)	Seven ICU databases in five countries (USA, Switzerland, the Netherlands, Austria, South Korea); 409,297 stays	One decoder trained once without task labels, with grammar and per-database support constraints; sampled and hazard estimates	13 primary and 165 scored endpoints against 13 learned models, 7 clinical scores and the databases' own scores; calibration mechanism; simulated treatment responses against matched observational contrasts with a null arm; membership inference of unconditional, prompted and intervened sampling; attribution sanity checks; toolkit and trial simulation

Supplementary Table 3 | TRIPOD+AI checklist.³⁰ Item numbers follow the statement; topics are abbreviated. The study develops and evaluates a generative model whose predictions are read from simulated futures.

Item	Topic	Where reported
1	Title	Title
2	Abstract	Abstract
3a	Healthcare context, rationale, existing models	Introduction, paragraphs 1–3
3b	Target population, intended purpose and users	Introduction, paragraph 1; Discussion (intended users)
3c	Known health inequalities	Methods, Data sources (race recorded in three databases); Results (subgroups); Supplementary Fig. 13
4	Objectives	Introduction, last paragraph
5a	Data sources	Methods, Data sources and cohort; Table 1
5b	Dates of data	Methods, Data sources and cohort; Table 1
6a	Setting, number and location of centres	Methods, Data sources and cohort; Table 1
6b	Eligibility	Methods, Data sources and cohort
6c	Treatments and their handling	Methods, Representation and grammar; What-if simulation
7	Data preparation and quality checks	Methods, Harmonization and preprocessing
8a	Outcome definitions and horizons	Methods, Prediction tasks and anchor designs
8b	Qualifications of outcome assessors	Not applicable: outcomes are computed from records; raters' qualifications in Methods, Rater study
8c	Blinding of outcome assessment	Methods, Prediction tasks (one label engine for real and simulated futures); Rater study (blinded)
9a	Choice of predictors	Didymos uses the whole record before the prediction time; comparators' features in Methods, Comparators
9b	Predictor definitions and timing	Methods, Representation and grammar (prediction inputs); Comparators
9c	Qualifications of predictor assessors	Not applicable
10	Study size	Methods, Data sources and cohort; Prediction tasks and anchor designs; Rater study (sizes from simulation)
11	Missing data	Methods, Representation and grammar (events as recorded, no imputation); Comparators (imputation and indicators; missing score components scored as normal)
12a	Use and partitioning of data	Methods, Data sources and cohort; Model, training and selection (held-out calibration screen)
12b	Handling of predictors	Methods, Representation and grammar; Comparators
12c	Model type, building, tuning, internal validation	Methods, Model, training and selection; Comparators
12d	Heterogeneity across clusters	Methods, Statistics (bootstrap stratified by database); Results; Fig. 2d; Supplementary Figs. 7 and 12
12e	Performance measures	Methods, Calibration and recalibration; Statistics and reproducibility
12f	Model updating	Methods, Calibration and recalibration
12g	How predictions are calculated	Methods, Simulation and the two estimates; Code availability
13	Class imbalance	Methods, Prediction tasks (enriched anchors with inclusion weights); Comparators (stratified sampling with odds correction)
14	Fairness	Results (subgroups); Supplementary Fig. 13; no fairness-specific method was applied
15	Model output	Methods, Simulation and the two estimates (probabilities; no classification thresholds)

Item	Topic	Where reported
16	Differences between development and evaluation data	Methods, Data sources and cohort (held-out patients of the same databases); Discussion
17	Ethics approval and consent	Methods, Ethics
18a	Funding	Acknowledgements
18b	Conflicts of interest	Competing interests
18c	Protocol	Methods, Study design and reporting
18d	Registration	Methods, Study design and reporting
18e	Data sharing	Data availability
18f	Code sharing	Code availability
19	Patient and public involvement	Methods, Study design and reporting
20a	Flow of participants	Supplementary Fig. 3; Table 1; Results
20b	Characteristics by data source	Table 1
20c	Comparison with development data	Table 1 (splits within each database)
21	Participants and events in each analysis	Figure legends (events/n); Methods; Supplementary Data 3
22	Full model specification	Methods, Model, training and selection; Code availability; Data availability
23a	Performance with intervals, including subgroups	Results; Fig. 2, Fig. 3; Supplementary Figs. 4–13
23b	Heterogeneity across clusters	Fig. 2d; Supplementary Figs. 7 and 12
24	Results of model updating	Results (calibration); Fig. 3b,c; Supplementary Fig. 14
25	Interpretation	Discussion
26	Limitations	Discussion
27a	Handling of poor-quality or unavailable inputs	Methods, Simulation (support masks); Results and Supplementary Fig. 12 (unknown site prompts)
27b	User interaction and expertise	Methods, The Didymos toolkit; Discussion (intended users)
27c	Next steps	Discussion, last paragraph

Supplementary Table 4 | Candidate checkpoints and the selection decision. Every candidate trained during development, each decided by a rule written before the candidate was measured. Tuning loss in nats per target token on the whole tuning split. The privacy audit is the membership-inference audit of training (Methods).

Candidate	Size, context	Change	Tuning loss	Privacy audit	Decision
S	71.5M, 8,192	First full run	1.614	Not audited	Larger model better on all databases
M	123.5M, 8,192	Default size	1.582	Not met (INSPIRE AUC 0.650 at 5.9 passes)	Passes rebalanced
M32	123.5M, 32,768	Longer context; rotary base 10^6 ; 2.5–3 passes	1.561	Met	First candidate; reference
ML32	269.5M, 32,768	2.2 times larger, same passes	1.563	Met, with gains for members only	Not adopted
MC32	123.5M, 32,768	ICU-day clock blocks	1.564	Met	Finalist by rule
MCN32	123.5M, 32,768	Plus input-noise augmentation	1.607	Met	Not adopted (higher loss)
MU32, MUH32	123.5M, 32,768	Muon optimizer, two learning rates	1.527, 1.509	Inconclusive; not met	Over the privacy bound
MUF32	123.5M, 32,768	Muon; passes projected to AUC 0.565	1.5388	Inconclusive (TPR upper bound 0.031)	Not adopted
MUC32 (Didymos 1.0)	123.5M, 32,768	Muon; passes projected to AUC 0.555	1.5525	Met (AUC 0.549–0.555)	Adopted

Supplementary Table 5 | The primary endpoints and their definitions. One label engine applied each definition to an anchor's real future and to every decoded simulated future alike. A binary endpoint counts an event that begins after the prediction time, within the horizon and before ICU discharge; an anchor already in the event at the prediction time (for example, already ventilated, for new IMV) is not eligible for it. A threshold on a value sits at the decile edge of that value's tokens nearest the clinical cut-off, so that no label depends on where a decoded value falls within a bin. Estimate: the one reported for the follow-up (H, hazard estimate; S, sampled estimate); the pre-specified analysis used the sampled estimate throughout. ASE, Adult Sepsis Event; KDIGO, Kidney Disease: Improving Global Outcomes; MAP, mean arterial pressure.

Endpoint	Definition	Estimate
<i>Rolling anchors (prediction times drawn along the ICU stay)</i>		
Death, 24 h	ICU death within 24 hours	H
New IMV, 24 h	Start of an episode of invasive mechanical ventilation within 24 hours	H
New vasopressor, 24 h	Start of a vasopressor infusion episode within 24 hours	H
New RRT, 48 h	Start of an episode of renal replacement therapy (continuous RRT or intermittent haemodialysis; a gap of 24 hours separates episodes) within 48 hours	H
AKI stage ≥ 2 , 48 h	KDIGO stage 2 or 3, by creatinine, urine output or continuous RRT, ² within 48 hours	S
Sepsis, 24 h	Onset of an ASE (a blood culture with at least four days of antimicrobials, and acute organ dysfunction) ³¹ within 24 hours; labelled in MIMIC-IV and AmsterdamUMCdb, which record cultures and administered antimicrobials	S
Hypotension, 4 h	MAP below 67 mmHg (invasive) or 62 mmHg (non-invasive), the decile edges nearest 65 mmHg, in consecutive measurements no more than 2 hours apart spanning at least 30 minutes, within 4 hours	S
Hyperkalaemia, 24 h	Potassium above 4.9 mmol/L, the top decile edge and the edge nearest 5.5 mmol/L, within 24 hours	S
Thrombocytopenia, 48 h	Platelet count below 71×10^3 per μL , the lowest decile edge and the edge nearest 50×10^3 per μL , within 48 hours	S
<i>24- and 48-hour landmarks (stays in the ICU and alive at the landmark)</i>		
ICU death	Death before ICU discharge	H
In-hospital death	Hospital discharge recorded as death, or death at ICU discharge; HiRID and AmsterdamUMCdb record mortality only at ICU discharge	S
ICU stay >7 days	A remaining ICU stay of more than 168 hours after the landmark; a simulated future stopped earlier by the token cap is censored (Kaplan–Meier)	S
Death, 24 h (follow-up)	ICU death within 24 hours of the landmark	H

Supplementary Table 6 | The comparators and their settings. Every comparator was scored on the same held-out anchors as Didymos and at the same prediction time. Learned models saw only data up to that time. They were trained on rows of real training stays drawn with each anchor design (at a landmark, every stay still in the ICU; on rolling anchors, up to four random ICU hours per stay), from up to 5,000 stays per database, with settings chosen and early stopping done on rows of the tuning split. Their features, decoded from the same tokens Didymos sees, 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 for the 64 most charted numeric channels, the active organ-support treatments, the statics documented at admission, the database and the hours since ICU admission. The logistic regressions, the decision tree, random forest, extra trees, naive Bayes, k -nearest neighbours and the multilayer perceptron received median imputation, missing-value indicators and standardization fitted on the training rows. Clinical scores needed no training and used the harmonized raw values recorded or resulted in the 24 hours before the prediction time, with a missing component scored as normal. AVPU, alert, voice, pain, unresponsive; GCS, Glasgow Coma Scale; ROX, the ratio of oxygen saturation to inspired oxygen over respiratory rate.

Family	Model	Settings
Gradient boosting	LightGBM, ³² pooled (the database as a feature) and per database	At most 1,000 trees, learning rate 0.05, 31 leaves, early stopping
Tabular foundation models	TabPFN-3.5 ³³ (weights run locally and offline)	All features; 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	Up to 10,000 rows and the features its input limit allows
Linear	Logistic regression, L2 penalty	Standardized features
	Logistic regression, L1 (liblinear) and elastic-net (l1 ratio 0.5, saga) penalties	C in {0.01, 0.1, 1}
	Didymos probe	Ridge logistic regression on Didymos's frozen last-layer state at the end of the prompt (last token and mean over the last window); C by five-fold cross-validation
Other trees	Decision tree (CART)	Depth 3, 5 or 8; at least 50 rows per leaf
	Random forest; extra trees	300 trees; the square root of the features at each split; at least 5 rows per leaf
Other learners	Gaussian naive Bayes	–
	k -nearest neighbours	k of 25 or 100, distance-weighted
	Multilayer perceptron	Hidden layers of 128 and 64 units; early stopping
Clinical scores	SOFA ³⁴	Worst values in the window
	SAPS II (without its chronic-disease items) and OASIS, at the landmarks ^{35,36}	Worst values; the published probability equations
	NEWS (NEWS2, oxygen-saturation scale 1), MEWS, qSOFA and SIRS ^{37–40}	Latest values; AVPU derived from the GCS
Databases' own scores	APACHE IVa (eICU-CRD), SAPS 3 (SICdb), ASA class (INSPIRE); RICD's recorded SOFA (transfer analysis) ^{5,6}	As recorded; for APACHE IVa, its predicted ICU and hospital mortality and ICU stay
Bedside values	One variable per endpoint, at its latest value, a missing value set to normal	Death within 24 h: shock index, lactate; new IMV: ROX index; new vasopressor: shock index, MAP; new RRT and AKI: KDIGO stage, creatinine; hypotension: MAP; hyperkalaemia: potassium; thrombocytopenia: platelets; sepsis: lactate

Supplementary Table 7 | Pre-specified decision rules beside the observed estimates. Every rule written into the protocol or its amendments before the analysis it governs, with what was observed. Rules added after a first result are marked post hoc. The main text presents estimates; this table records how each rule came out.

Analysis	Rule	Observed	Outcome under the rule
Forecasting (pre-specified comparisons)	Descriptive: Didymos against LightGBM per primary endpoint, paired difference with its interval; no pass or fail	Supplementary Table 1; Fig. 2	Descriptive
Privacy audit of training	Per database and pooled: upper 95% bounds of membership $AUC \leq 0.60$ and of TPR at 1% FPR ≤ 0.03 ; extraction, copies, density-ratio and attribute-inference limits; copying generator must fail and private reference pass	AUC 0.549–0.555, TPR 0.010–0.024 in every database; all other tests within their limits; controls as specified	Met
Prompted sampling, loss attack (500 per side)	As above	Three databases' TPR intervals too wide	Inconclusive
Prompted sampling, loss attack at full sizes (post hoc)	As above, up to 3,000 per side	Pooled AUC 0.531 (0.524–0.537), TPR 0.013 (0.011–0.017); INSPIRE (790 per side) AUC 0.568 (0.538–0.597), TPR 0.018 (0.009–0.039)	Met pooled and in six databases; INSPIRE inconclusive; recorded as passed by the study lead, with INSPIRE stated
Prompted and intervened extraction	Member minus non-member rate of reproducing the first 32 tokens ≤ 0.01 ; Jaccard AUC upper bound ≤ 0.60	0 of members and non-members, with and without intervention; Jaccard AUC 0.516 (0.493–0.540)	Met
Toolkit conformance	AUROC and O/E of nine endpoints within the pre-specified run's 95% interval; hazard-estimate AUROC within 0.02 of the follow-up curve; what-if contrasts within the follow-up's intervals	14 of 15 checks; sepsis (8 events) outside: AUROC 0.963 against 0.989 (0.974–0.999)	14 of 15
Toolkit conformance, added check (post hoc, fixed before computed)	Within $1.96 \sqrt{2}$ times the rollout-bootstrap SD of the pre-specified run	7 of 9; death within 24 h (AUROC 0.880 against 0.903, limit 0.017; O/E 1.69 against 1.85, limit 0.15) and new RRT (by 6×10^{-7}) outside	7 of 9
Toolkit conformance, adjusted criterion (post hoc, study lead)	Within that limit of either of the paper's two runs	All nine; the paper's two runs differ by 3.7 noise SDs on death within 24 h	9 of 9
Recalibration, isotonic	Adopt where AUROC stays within 0.005 at 32 and 16 futures	Held for death within 24 h, AKI, hypotension and hyperkalaemia	Four adopted
Recalibration, logistic (post hoc)	Adopt when $b > 0$ and, at 32 and 16 futures, O/E closer to 1 and calibration error no larger	Met by new IMV, new vasopressor and new RRT	Three adopted

Analysis	Rule	Observed	Outcome under the rule
Trial simulation checks	Type I error 0.04–0.06 at relative risk 1; unadjusted simulated size within 5% of the analytic size	92 of 108 cells for each check	92 of 108
Teaching cases	Each case meets its objective; pages self-contained	Nine of nine	Met
Attribution remedies (exploratory)	Sanity correlation about 0.3 or less and positive agreement with editing	Knockout –0.02 and 0.71; attention × gradient (means) –0.03 and 0.14; gradient-weighted rollout (means) –0.04 and 0.14; last-layer attention (means) 0.01 and 0.08; rollout variants 0.92–0.97	Knockout and three gradient or last-layer variants met; rollout variants did not
Rater study, A (realism)	Pooled accuracy equivalent to chance: 90% CI within 0.40–0.60	0.408 (0.351–0.468), below chance; post hoc non-detection bound 0.47	Equivalence not shown
Rater study, B1 (simulated responses)	Non-inferiority, margin –0.5 on a 1–5 scale, insulin excluded	–0.23 (–0.56 to 0.11)	Not established
Rater study, B2 (explanations)	Editing and knockout each rated above random factors, Holm	+0.09 (–0.22 to 0.41) and –0.01 (–0.32 to 0.31); Holm $P = 1.00$ each	Not established
Rater study, C (decision support)	ICU-death Brier score, with minus without the card, $P < 0.05$	–0.016 (–0.031 to –0.001); $P = 0.037$	Met
Model selection, adoption	Privacy audit met; tuning loss $\leq$ reference + 0.01; better on authenticity (+2) or calibration (+3) and no worse on the other	Authenticity 30.5 against 25.0 of 35; calibration 21 against 22	Met; one departure from the finalist rule (Methods)
Transfer to new databases, availability	An endpoint is scored where its defining source is timed to the hour and at least 20 anchors of the design are positive	15 of 26 endpoint–database pairs (zigongICU’s nursing chart records dates only; RICD has no RRT source)	15 pairs scored
Adaptation, choice of the grid cell	Lowest next-token loss on the held-out 10% of the unlabelled patients	10^{-4} and three passes, 1.131 nats per token (others 1.291–1.530)	Chosen
Adaptation, membership audit on the seven databases	As the privacy audit of training (3,000 per side)	eICU-CRD TPR 0.0227 (0.0127–0.0303); all other groups and tests within their limits	Inconclusive
Adaptation, membership audit at a larger size (fixed before it ran)	As above, up to 10,000 per side	eICU-CRD TPR 0.0157 (0.0125–0.0194); AmsterdamUMCdb (2,747 per side, all held-out stays) 0.0226 (0.0153–0.0313); pooled 0.0151 (0.0132–0.0169)	Inconclusive; futures drawn for this analysis only, by the study lead’s decision, with no synthetic data

Supplementary Table 8 | The code mapping for the two new databases. For RICD and zigongICU only, a timeline event whose code the frozen vocabulary lacks was rewritten, before encoding, to the nearest code the vocabulary keeps. The first rule that yields a kept code applies, and every other event is untouched. The mapping is off for the seven training databases, which encode exactly as before. Below the rules, the share of each domain’s events encoded as an unknown token, in the original encoding and with the mapping (the domains with unknown events).

Rule	Mapping
Whole-code substitution	Stool output to other gastrointestinal output; a rehabilitation unit to an inpatient ward and a haemodialysis room to a procedural area, for transfers in and out
Unit conversion within a family	A dose in another unit of the same family, converted to the kept unit of the same drug and route: mass (g, mg, μ g, ng), volume (L, mL), amount (mol, mmol) and units (U, kU)
Drug-specific conversion	Enoxaparin, 100 anti-Xa IU to 1 mg; potassium chloride, 74.55 mg to 1 mmol, and 1 mEq to 1 mmol
Infusion whose unit is not kept	The kept unit of the same drug and route, first in a fixed order (mL/h, mg/h, μ g/kg/min, μ g/min, mg/min, . . .), with the rate recorded as unknown; a stop keeps its zero rate
Dose whose unit cannot be converted	The kept unit of the same drug and route, first in a fixed order (unspecified, mg, μ g, g, mL, U, . . .), with the dose recorded as unknown
Route	An unspecified route, tried after the recorded one
<i>Events encoded as unknown: original encoding $\rightarrow$ with the mapping</i>	
RICD	Continuous medications 85.1% $\rightarrow$ 0.0%; intermittent medications 31.6% $\rightarrow$ 2.9%; outputs 14.5% $\rightarrow$ 0.0%; locations 12.7% $\rightarrow$ 0.0%
zigongICU	Continuous medications 55.5% $\rightarrow$ 0.6%; intermittent medications 6.5% $\rightarrow$ 1.2%; locations 0.1% $\rightarrow$ 0.0%

Supplementary Table 9 | The P -value approximation check. For the hazard estimate, whose per-anchor values were not kept, P values come from the stored 95% interval by a normal approximation. The approximation was applied to the 36 comparisons of Didymos's sampled estimate (128 futures) for which exact paired-bootstrap P values exist (10,000 resamples): the median absolute difference in $\log_{10}P$ is 0.060 (maximum 0.72), and the decision at 0.05 is the same in 36 of 36. Comparators: LightGBM (pooled), TabPFN-3.5, and the pre-specified 32-future estimate.

Anchors	Endpoint	Comparator	P , exact	P , from the interval
Rolling	Death, 24 h	LightGBM	<0.001	<0.001
Rolling	Death, 24 h	TabPFN-3.5	<0.001	<0.001
Rolling	Death, 24 h	Pre-specified, 32 futures	0.22	0.26
Rolling	New IMV, 24 h	LightGBM	0.24	0.21
Rolling	New IMV, 24 h	TabPFN-3.5	0.027	0.031
Rolling	New IMV, 24 h	Pre-specified, 32 futures	0.41	0.38
Rolling	New vasopressor, 24 h	LightGBM	<0.001	0.0027
Rolling	New vasopressor, 24 h	TabPFN-3.5	<0.001	<0.001
Rolling	New vasopressor, 24 h	Pre-specified, 32 futures	0.091	0.067
Rolling	New RRT, 48 h	LightGBM	0.14	0.13
Rolling	New RRT, 48 h	TabPFN-3.5	0.27	0.23
Rolling	New RRT, 48 h	Pre-specified, 32 futures	0.028	0.025
Rolling	AKI stage ≥ 2 , 48 h	LightGBM	<0.001	<0.001
Rolling	AKI stage ≥ 2 , 48 h	TabPFN-3.5	0.0048	0.0034
Rolling	AKI stage ≥ 2 , 48 h	Pre-specified, 32 futures	0.90	0.90
Rolling	Sepsis, 24 h	LightGBM	0.63	0.71
Rolling	Sepsis, 24 h	TabPFN-3.5	0.40	0.44
Rolling	Sepsis, 24 h	Pre-specified, 32 futures	0.16	0.069
Rolling	Hypotension, 4 h	LightGBM	0.10	0.10
Rolling	Hypotension, 4 h	TabPFN-3.5	0.20	0.14
Rolling	Hypotension, 4 h	Pre-specified, 32 futures	0.88	0.94
Rolling	Hyperkalaemia, 24 h	LightGBM	0.49	0.45
Rolling	Hyperkalaemia, 24 h	TabPFN-3.5	0.31	0.30
Rolling	Hyperkalaemia, 24 h	Pre-specified, 32 futures	0.38	0.34
Rolling	Thrombocytopenia, 48 h	LightGBM	0.066	0.084
Rolling	Thrombocytopenia, 48 h	TabPFN-3.5	0.007	0.026
Rolling	Thrombocytopenia, 48 h	Pre-specified, 32 futures	0.0078	0.041
24-hour landmark	ICU death	LightGBM	0.83	0.79
24-hour landmark	ICU death	TabPFN-3.5	0.089	0.066
24-hour landmark	ICU death	Pre-specified, 32 futures	<0.001	0.0015
24-hour landmark	In-hospital death	LightGBM	0.0086	0.0051
24-hour landmark	In-hospital death	TabPFN-3.5	0.47	0.44
24-hour landmark	In-hospital death	Pre-specified, 32 futures	0.0014	<0.001
24-hour landmark	ICU stay >7 days	LightGBM	0.38	0.37
24-hour landmark	ICU stay >7 days	TabPFN-3.5	1.00	1.00
24-hour landmark	ICU stay >7 days	Pre-specified, 32 futures	0.33	0.33

Supplementary Table 10 | Terminology and traceability. The paper’s names, the names used in the code and the protocol, and where each figure’s numbers come from. Every figure’s source data list its input files with their sha256 and the code version (Supplementary Data 3).

Paper term	Code or protocol	Definition or source
Didymos 1.0	Checkpoint MUC32_main	The adopted model
Adapted Didymos	Checkpoint E17_lr0.0001_p3; arm F1 of amendment 17	Didymos 1.0 trained further on the unlabelled stays of RICD and zigongICU
Zero-shot; mapped codes	Arms Z0 and Z1 of amendment 17	Didymos 1.0 on the original and on the mapped encoding of the new databases
Recalibrated; stacked	Arms F1+R and F1+S of amendment 17	Adapted Didymos with a logistic map fitted on the labelled slice; stacked with a LightGBM trained on it
The Didymos toolkit	icutwin.toolkit, command icutwin_toolkit	Forecast, What-if, Explain, Similar patients, Cohort synthesizer, Trial designer, Teaching cases
The Didymos probe	Twin probe	Ridge logistic regression on Didymos’s frozen last-layer state
Simulated futures	Rollouts	Continuations of a record drawn under the grammar and the database’s support
Sampled estimate	E0, the share	Fraction of simulated futures in which the event occurs
Hazard estimate	E2, the hybrid	Step probability of the event summed along each future while at risk, averaged over futures
Rolling anchors	R (pre-specified, 3,500); R128 (follow-up, 1,050)	Prediction times sampled along held-out stays, with inclusion weights
24-hour landmark	L24 (pre-specified, 3,500); L24x (follow-up, 700)	Prediction 24 hours after ICU admission; L24x is a case-cohort sample
Synthetic patients; the Didymos synthetic cohort	Twins; the bank (13,416 stays)	Whole stays generated by Didymos
Pre-specified analysis	A1–A5, B2–B3, C1–C4	The protocol and its first six amendments
Follow-up analysis	F1–F9; amendments 9–11	Fixed before they ran, after the pre-specified results were known
Null arm	A/A arm	Two withhold arms with different seeds
Held-out calibration screen	Selection targets S1–S2	Calibration of synthetic stays and simulated futures on held-out stays during model selection
Fig. 1	Materials	materials/table1.json; architecture and tokenization records
Fig. 2	A1, F1, F2, comparators	a1/test_*/a1.json, f1/{R128,L24x}/f2.json, f10/comparators.json
Fig. 3	A2, F2, F3, V2, V2b	f1/*/f2.json, f3/*, c3/c3b.json, toolkit V1–V2 outputs
Fig. 4	C1, F5	c1/c1.json, f5/*
Fig. 5	A3, A4, C3, F6, F8, V3, V3x	a3/a3.json, c3/*, f8/*, toolkit V3 outputs
Fig. 6	C2, F4, attention diagnosis	c2/*, f4/*
Fig. 7	V1, V4, V5	Toolkit V1, V4 and V5 outputs
Fig. 8	Rater study	Pooled aggregates of the answer-only analysis
P values	Reporting statistics	stats.json; Supplementary Data 2

Supplementary Data

Supplementary Data 1 | Didymos against LightGBM on every scored endpoint.

One row per comparison (165): the anchor set (R, rolling anchors; L24 and L48, the 24- and 48-hour landmarks), the endpoint and its catalog family (F1 mortality, F2 organ failure, F3 sepsis, F4 intervention onsets, F5 discharge, F6 threshold crossings, F7 remaining stay), events and anchors, the difference between Didymos's sampled estimate (32 simulated futures) and LightGBM with its 95% CI from a paired bootstrap over stays (1,000 resamples), the two-sided P value from 10,000 paired resamples, the Benjamini–Hochberg q value across the 165 comparisons and the Holm-adjusted P value. Exploratory analysis, defined after the per-endpoint intervals were known (Supplementary Fig. 9). File: `supplementary_data_1_breadth.csv`.

Supplementary Data 2 | Reporting statistics.

The statistics shown in the figures, one file per family: Didymos against LightGBM and against TabPFN-3.5, the follow-up against the pre-specified estimate, Didymos against the best clinical score, calibration before and after recalibration, simulated against observational contrasts, the null arm, membership signal against fit, augmentation, attribution, ranks across endpoints, label efficiency, the trial simulations, the leaderboard (Table 2), the 48-hour landmark re-run, knockout against the redrawn effects, transfer to the two new databases and adaptation to them. Each file gives the estimate, its interval, the unadjusted and adjusted P values and the method. Files: `supplementary_data_2_statistics/`.

Supplementary Data 3 | Source data for every figure.

For each main and supplementary figure, the numbers plotted in each panel (estimates, intervals, events and n), and a manifest listing every input file with its sha256 and the version of the code that drew the figure. Aggregates only; no individual stay is included. Files: `supplementary_data_3_source_data/`.

Supplementary Data 4 | Schematic artwork and diagram sources.

The native files of the schematic artwork (GPT-image) in Fig. 1a and c, Fig. 4a, Fig. 7a and Supplementary Fig. 1–Supplementary Fig. 3, with their content credentials and sha256 (`provenance.json`); and the Mermaid sources of Supplementary Fig. 1 and Supplementary Fig. 2, which hold every label and number the artwork shows. The artwork is credited “(GPT-image)”; the authors confirm its accuracy and scientific validity and assume full responsibility for it. Files: `supplementary_data_4_artwork/`.

Supplementary References

- [1] 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).
- [2] Khwaja, A. KDIGO clinical practice guidelines for acute kidney injury. *Nephron Clin. Pract.* **120**, c179–c184 (2012).
- [3] Prentice, R. L. A case-cohort design for epidemiologic cohort studies and disease prevention trials. *Biometrika* **73**, 1–11 (1986).
- [4] Casella, G. & Robert, C. P. Rao-Blackwellisation of sampling schemes. *Biometrika* **83**, 81–94 (1996).
- [5] 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).
- [6] 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).
- [7] Harrell, F. E., Califf, R. M., Pryor, D. B., Lee, K. L. & Rosati, R. A. Evaluating the yield of medical tests. *JAMA* **247**, 2543–2546 (1982).
- [8] Rosenbaum, P. R. & Rubin, D. B. The central role of the propensity score in observational studies for causal effects. *Biometrika* **70**, 41–55 (1983).
- [9] Gneiting, T. & Raftery, A. E. Strictly proper scoring rules, prediction, and estimation. *J. Am. Stat. Assoc.* **102**, 359–378 (2007).
- [10] 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).
- [11] Carlini, N. *et al.* Membership inference attacks from first principles. Preprint at <https://arxiv.org/abs/2112.03570> (2021).
- [12] Carlini, N. *et al.* Extracting training data from large language models. Preprint at <https://arxiv.org/abs/2012.07805> (2020).
- [13] Sundararajan, M., Taly, A. & Yan, Q. Axiomatic attribution for deep networks. Preprint at <https://arxiv.org/abs/1703.01365> (2017).
- [14] Abnar, S. & Zuidema, W. Quantifying attention flow in transformers. Preprint at <https://arxiv.org/abs/2005.00928> (2020).
- [15] Adebayo, J. *et al.* Sanity checks for saliency maps. Preprint at <https://arxiv.org/abs/1810.03292> (2018).
- [16] Chefer, H., Gur, S. & Wolf, L. Transformer interpretability beyond attention visualization. Preprint at <https://arxiv.org/abs/2012.09838> (2020).
- [17] 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).
- [18] 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).
- [19] Grechko, A. V. *et al.* RICD: Russian Intensive Care Dataset. *Gen. Reanimatol.* **20**, 22–31 (2024).
- [20] Xu, P. *et al.* Critical care database comprising patients with infection. *Front. Public Health* **10**, 852410 (2022).
- [21] Renc, P. *et al.* Zero shot health trajectory prediction using transformer. *npj Digit. Med.* **7**, 256 (2024).
- [22] Renc, P. *et al.* Foundation model of electronic medical records for adaptive risk estimation. *GigaScience* **14**, g1af107 (2025).
- [23] 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).
- [24] Shmatko, A. *et al.* Learning the natural history of human disease with generative transformers. *Nature* **647**,

248–256 (2025).

- [25] Waxler, S. *et al.* Generative medical event models improve with scale. Preprint at <https://arxiv.org/abs/2508.12104> (2025).
- [26] Makarov, N. *et al.* Large language models forecast patient health trajectories enabling digital twins. *npj Digit. Med.* **8**, 588 (2025).
- [27] 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).
- [28] 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).
- [29] 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).
- [30] 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).
- [31] Rhee, C. *et al.* Incidence and trends of sepsis in US hospitals using clinical vs claims data, 2009–2014. *JAMA* **318**, 1241–1249 (2017).
- [32] Ke, G. *et al.* LightGBM: a highly efficient gradient boosting decision tree. In *Advances in Neural Information Processing Systems 30* (2017).
- [33] Hollmann, N. *et al.* Accurate predictions on small data with a tabular foundation model. *Nature* **637**, 319–326 (2025).
- [34] 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).
- [35] 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).
- [36] 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).
- [37] 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).
- [38] Subbe, C. P., Kruger, M., Rutherford, P. & Gemmel, L. Validation of a modified Early Warning Score in medical admissions. *QJM* **94**, 521–526 (2001).
- [39] 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).
- [40] 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).