

When explanations disagree: a consequence-calibrated audit of post-hoc attribution for clinical time-series prediction

Ahmed Mohammed Fawaz

Ahmed829536@med.psu.edu.eg

Port Said University

Research Article

Keywords: explainable AI, clinical decision support, electronic health records, model interpretability, regulatory science, intensive care

Posted Date: October 14th, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-10842403/v1>

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

[Read Full License](#)

Additional Declarations: No competing interests reported.

Abstract

Post-hoc explanation methods are increasingly attached to clinical risk models on the assumption that different methods provide materially consistent accounts of the same prediction. We tested this assumption using a consequence-calibrated audit across 13 ICU cohorts from three databases (MIMIC-III, MIMIC-IV and eICU-CRD; 28,875 admissions). Three gradient-based attribution methods showed moderate cross-method agreement (mean Kendall $\tau = 0.50$; range, 0.41–0.61), with no cohort reaching the pre-specified substantial-agreement threshold of $\tau = 0.70$; agreement varied with data source rather than clinical population. In seven semi-synthetic benchmarks with known ground truth, attribution-method rankings were condition-dependent and reversed between signal conditions. Under a structural, non-clinically adjudicated action mapping, two method pairs produced different recommended actions in 62–66% of synthetic cases and 67–69% of cases in a real cohort, whereas the third pair differed in 7–11%. A model of disagreement achieved an out-of-sample AUC of 0.73, but no tested stratum met the pre-specified trust threshold, and disagreement did not predict misclassification in the expected direction. A separate vignette survey of 35 clinicians provided method-independent evidence that the action categories represented genuine and sometimes specialty-patterned clinical disagreement. These findings support consequence-calibrated auditing of explanation behaviour rather than selection of a universally preferred attribution method.

Background

Clinical prediction models are now embedded in electronic health record systems at scale, and the question of whether clinicians should act on them has become inseparable from the question of whether they can be explained. Regulators, professional bodies and consensus guidelines increasingly treat an explanation as part of what makes a model deployable rather than as an optional adjunct.^{1,2} The dominant response has been post-hoc attribution: a model is trained for accuracy, and a separate procedure SHAP, Integrated Gradients, gradient saliency, attention weights is applied afterwards to report which input features drove each prediction.^{3,4,5} Such attributions now appear in clinical AI systems deployed at the bedside.⁶ A dissenting position holds that high-stakes decisions should use models that are interpretable by construction, precisely because post-hoc explanation of an opaque model may not be trustworthy;⁷ our results bear directly on the empirical premise of that argument.

This arrangement rests on an assumption that is rarely stated and, to our knowledge, has never been tested at scale on clinical time-series data: that the choice of attribution method is a technical detail rather than a clinical one. If two methods applied to the same model and the same patient produce materially different accounts of why that patient is at risk, then the explanation is a property of the method, not of the model — and the transparency it purports to provide is illusory. The problem has been named in the machine-learning literature, where practitioners report resolving disagreement by ad hoc heuristics,⁸ and individual method classes have been shown to fail basic sanity checks⁹ or to be

unreliable as explanations of model behaviour.¹⁰ What has been missing is an assessment on the kind of model, data and decision that clinical deployment actually involves.

Three gaps have kept the question open. First, the clinical AI literature is dominated by surveys and taxonomies of explainability rather than empirical measurement of whether explanations agree.^{11,12} Second, where benchmarking has been done rigorously it has been on medical images, where the unit of explanation is a spatial region and the clinical reader can inspect it directly; the transferability of those findings to irregularly sampled physiological time series is unknown. Third, and most consequentially, the field lacks any established way to convert a disagreement statistic into a statement about patient care. A Kendall τ of 0.5 between two feature rankings is not self-interpreting: it becomes meaningful only when tied to whether the two rankings would lead a clinician to do different things.

The cost of leaving this untested is not hypothetical. Continuous deterioration models are deployed on inpatient time series at scale,^{13,14} yet such systems have also failed silently after deployment,¹⁵ and post-deployment failure modes are frequently invisible to the discrimination metrics used to validate them.¹⁶ An explanation that is unstable across methods is a mechanism by which such failure remains invisible: it supplies a plausible narrative for any prediction, correct or not.

Here we develop and evaluate a consequence-calibrated audit instrument for post-hoc attribution in clinical time-series prediction. We first measure cross-method agreement across multiple clinical populations and independently collected databases, allowing both generalisation and source dependence to be assessed. We then calibrate disagreement against a structural action space, asking not only whether attribution rankings differ but whether those differences would imply different interventions. Because real clinical data provide no ground truth for which features truly caused an outcome, we additionally construct semi-synthetic cohorts in which the injected causal structure is known and require each benchmark to pass an admissibility check before attribution faithfulness is assessed. Finally, we test whether disagreement can be predicted before an explanation is generated and whether it identifies a reproducible trust regime. We deliberately do not seek to crown a universally superior attribution method. Rather, we test whether explanation behaviour is sufficiently condition-dependent that method standardisation is itself unsafe, and whether an audit can identify circumstances in which explanations should be treated with greater caution.

Results

Mortality prediction across three independent databases

We assembled 13 ICU cohorts spanning five neurocritical and neurosurgical populations: aneurysmal subarachnoid haemorrhage, brain tumour resection, spinal fusion and decompression, spontaneous intracerebral haemorrhage, and traumatic brain injury craniotomy. from three independently collected critical-care databases (MIMIC-III, MIMIC-IV and eICU-CRD), totalling 28,875 admissions. For each cohort we trained a multilayer perceptron on the flattened observation window to predict in-hospital mortality

from routinely recorded vital signs, using subject-level splits throughout and standardisation fitted on development folds only. Discrimination was consistent with published models on comparable tasks and populations (AUROC 0.66–0.80, mean 0.73), and outcome prevalence ranged from 8.9% to 32.1% (Table 1).

Table 1

Cohort characteristics, predictive performance and cross-method explanation agreement. All 13 cohorts that passed the leakage audit. Mean τ is the mean pairwise Kendall rank agreement between attribution methods, averaged over patients and method pairs. Events (%) is in-hospital mortality prevalence.

Database	Population	N	Events (%)	AUROC	AUPRC	Brier	Mean τ
eICU-CRD	aSAH	2,134	15.9	0.763	0.483	0.124	0.415
eICU-CRD	Spontaneous ICH	692	17.8	0.751	0.567	0.135	0.441
eICU-CRD	TBI craniotomy	2,832	12.8	0.804	0.407	0.087	0.425
MIMIC-III	aSAH	646	24.6	0.663	0.390	0.216	0.608
MIMIC-III	Brain tumour resection	1,052	14.9	0.695	0.315	0.157	0.494
MIMIC-III	Spinal fusion/decompression	2,741	8.9	0.677	0.273	0.087	0.529
MIMIC-III	Spontaneous ICH	1,331	32.1	0.723	0.627	0.235	0.551
MIMIC-III	TBI craniotomy	2,597	15.9	0.690	0.466	0.154	0.564
MIMIC-IV	aSAH	1,299	23.6	0.733	0.560	0.208	0.539
MIMIC-IV	Brain tumour resection	1,920	13.1	0.730	0.390	0.132	0.518
MIMIC-IV	Spinal fusion/decompression	4,815	10.6	0.715	0.341	0.110	0.415
MIMIC-IV	Spontaneous ICH	3,100	24.8	0.736	0.535	0.204	0.504
MIMIC-IV	TBI craniotomy	3,716	13.1	0.749	0.388	0.139	0.464
All	13 cohorts	28,875	—	0.725	0.442	0.153	0.497

Crossing five neurocritical/neurosurgical populations with three databases gave 15 candidate cohort–database combinations. One (eICU-CRD brain tumour resection) could not be extracted, its source file contained zero records and was excluded before any modelling; this is a data-availability gap, not a leakage-audit failure. Of the 14 combinations that loaded, each then passed a pre-registered leakage audit that tests for outcome information reachable from the feature window before any explanation was computed. One of the 14 (eICU-CRD spinal fusion and decompression) failed, a single feature reached

AUROC = 1.00 against the outcome, the audit's leakage signature and was excluded from all downstream analysis. The 13 cohorts reported in Table 1 are those that passed both checks. This gate is reported rather than assumed because an explanation audit on a leaking model measures the leak, not the model.

Attribution methods agree only moderately, in every cohort

We applied three gradient-based attribution methods: gradient saliency, Gradient×Input and Integrated Gradients to each cohort's model, alongside two randomised floor controls, and measured pairwise rank agreement between the feature importances they assigned to each patient (Kendall τ , averaged over patients and method pairs). Agreement was moderate and strikingly consistent in its mediocrity: mean τ = 0.50 across cohorts, ranging from 0.41 (MIMIC-IV spinal fusion and decompression) to 0.61 (MIMIC-III aneurysmal subarachnoid haemorrhage) (Fig. 1a). Not one of the 13 cohorts reached τ = 0.70, the conventional threshold for substantial agreement. This 3-method figure is stable under a broader check: an 8-method panel that also includes KernelSHAP, TimeSHAP and ExpectedGrad alongside the two floors gives a lower mean ($\tau \approx 0.21$ –0.38 per cohort), as expected since a wider, more heterogeneous method set agrees less the two figures answer different questions and are not in tension. Agreement was nonetheless far above the chance floor established by randomised and constant control attributions ($|\tau| \leq 0.027$ in every cohort), so the methods are not simply unrelated to one another: they agree about roughly half of the ordering and disagree about the rest.

The structure of that disagreement is informative. Agreement varied systematically with the source database and not with the clinical population: source explained 63% of between-cohort variance in τ ($p = 0.007$), while population explained 8% ($p = 0.95$) (Fig. 1b). The same clinical population, audited in two different databases, yielded different agreement; different populations within one database did not. This has a direct practical consequence an explanation audit conducted on one data source provides little assurance about the behaviour of the same methods on another, which is precisely the transfer that a multi-site deployment requires.

We also observed an apparent inverse relationship between model discrimination and explanation agreement (crude $\beta = -1.17$, $p = 0.005$), which would be a striking finding if it held. It does not: adjusting for source database collapses the association ($\beta = -0.22$, $p = 0.76$), and it does not reproduce within any single database (Fig. 1c). We report it because the crude version is the analysis a reader might expect, and because it illustrates how readily a confounded cross-database comparison can generate an interpretable-sounding claim about explanation quality.

Faithfulness cannot be assessed on real data — and on synthetic data, no method wins

Real clinical data has no ground truth about which features truly drove an outcome, so faithfulness is not estimable there; agreement between methods is the most that real data can support. To assess faithfulness we constructed a seven-rung ladder of semi-synthetic cohorts in which the causal features are known by construction and graded in difficulty: an easy case, a near-threshold case, and cases

exercising transient signals, early signals, feature interactions, redundant features and distractor features.

Each rung was gated for admissibility before use. A benchmark is admissible only if the model both discriminates on it and loses that discrimination when the injected causal cells are ablated otherwise the model is solving the task by some route other than the planted signal, and measuring attribution against the planted signal is meaningless. All 14 rung-by-model benchmarks (7 rungs $\times$ 2 model families) passed this gate, with ablated AUROC falling to near chance (0.36–0.55 under the counterfactual operator).

Against this ground truth, no method was reliably most faithful (Fig. 2a). Integrated Gradients or Gradient $\times$ Input led on five of seven rungs, but on the transient-signal rung the ordering inverted completely: gradient saliency achieved perfect recovery (AP = 1.00) while ranking last on the early-signal rung. Between those two rungs the method ranking was an exact reversal ($\tau = -1.00$). Absolute faithfulness was also low wherever the task was hard: on the near-threshold rung the best method reached AP = 0.14 against a base rate of 0.02, and on the interaction rung AP = 0.22. A method that recovers ground truth perfectly under one condition and fails under another cannot be recommended on the basis of an average.

Method self-consistency under input perturbation showed the same absence of a clear winner in a different form: gradient saliency was highly unstable (0.04), while Gradient $\times$ Input (0.56) and Integrated Gradients (0.57) were comparable and neither met the pre-registered stability criterion.

Disagreement changes the recommended action

Agreement statistics acquire practical meaning only when linked to an action space. We therefore mapped each attribution to the intervention it most strongly implicated and measured the proportion of cases in which two attribution methods, applied to the same patient and model, implicated different interventions. We termed this the decision-flip rate and estimated 95% confidence intervals using a subject-clustered bootstrap. A pre-registered non-inferiority margin of $\delta = 0.10$ defined consequential disagreement for this demonstration: a method pair was considered to exceed the margin when the lower confidence bound of its flip-rate estimate exceeded 0.10.

Two of three method pairs exceeded this margin: saliency versus Gradient $\times$ Input at 62.2% (95% CI, 55.0–69.4) and saliency versus Integrated Gradients at 65.6% (58.3–72.8). Gradient $\times$ Input versus Integrated Gradients produced a lower flip rate of 11.1% (6.7–16.1) and did not exceed the margin. Thus, disagreement was not uniformly consequential: two method pairs were substantially separated in the defined action space, whereas the third was comparatively close. This distinction is not captured by an aggregate similarity statistic alone.

We emphasise that the action mapping used here is a structural placeholder rather than a clinically adjudicated mapping, and the resulting flip rates should therefore be interpreted as an evaluation of the audit instrument rather than estimates of clinical decision impact.

The decision-flip finding replicates on real patient data

The analysis above uses semi-synthetic data because faithfulness requires known ground truth. The decision-flip mechanism itself does not: it requires only real attributions and a feature-to-action mapping, both of which are available for real cohorts. We therefore repeated the decision-flip analysis on the largest real cohort (MIMIC-IV spinal fusion and decompression, $n = 723$ test cases), applying a structural mapping keyed to the real 6-channel vital-sign panel used to train that cohort's model (heart rate, systolic and mean arterial pressure, respiratory rate, oxygen saturation, temperature) in place of the synthetic cohort's generic feature slots.

The far/close pattern reproduced closely in the largest real cohort. Saliency versus Gradient×Input flipped on 67.3% of cases (95% CI, 62.0–72.7), saliency versus Integrated Gradients on 68.7% (63.3–73.7), and Gradient×Input versus Integrated Gradients on 6.7% (4.0–9.3). Thus, the two more widely separated method pairs again exceeded the pre-specified margin, whereas the closest pair did not. This analysis was restricted to a single real cohort and used a structural action mapping that was not clinically adjudicated; it therefore does not establish that attribution disagreement changes clinical decisions. Rather, it shows that the qualitative far/close structure observed in the semi-synthetic benchmark is not confined to the synthetic construction.

Disagreement is predictable before it happens

If disagreement were unpredictable, the only safe policy would be to distrust all post-hoc explanation. It is not. We fitted a model predicting per-case disagreement from three pre-registered covariates the correlation among candidate features, the temporal concentration of the signal, and the model's prediction margin using generalised estimating equations with subject-level clustering. Feature correlation predicted disagreement in the expected direction; temporal concentration and prediction margin did not reach significance (see below). The model transferred to held-out cases with AUC 0.73 ($n = 270$; Brier 0.118, calibration intercept 0.12, slope 0.93) (Fig. 3a,b).

Expressed per standard deviation of each predictor, higher feature correlation sharply reduced the log-odds of agreement (-2.68 , $p < 1 \times 10^{-4}$); temporal concentration (-1.89 , $p = 0.38$) and prediction margin ($+1.02$, $p = 0.14$) did not reach significance and should be treated as exploratory rather than relied upon individually. Feature correlation is therefore the dominant, reliable predictor of disagreement in this fit. The calibration slope of 0.93 indicates predicted probabilities are close to nominal, with modest over-confidence (intercept 0.12).

Stratifying by prediction margin and temporal concentration to search for a trust regime — a stratum reaching our pre-registered substantial-agreement threshold ($\tau \geq 0.70$, $n \geq 30$) — found none: the best of nine strata (low margin, temporally spiky) reached mean $\tau = 0.30$ ($n = 531$), and the worst reached $\tau = 0.14$ ($n = 677$) (Fig. 2c). Per our pre-registration, we report this as a negative result rather than lowering the threshold post hoc.

Restated as an operating rule, the honest one given these results is a negative one: no combination of the three covariates we tested identifies a stratum where post-hoc explanations can currently be certified as reliable at our pre-registered threshold. Feature correlation remains a usable continuous signal for ranking cases by expected disagreement (that is what the AUC 0.73 model captures) but it does not cash out into a bright-line rule for withholding versus trusting a given explanation. Quantifying when explanations should not be trusted therefore remains only partly solved by this study: we show it is possible to rank cases by risk of disagreement, not yet to certify any subset of cases as safe, and the boundary has not been tested prospectively or at sites beyond those in Table 1.

Disagreement does not predict a wrong prediction

Predictability and a trust boundary are not the only operational questions. A tempting further use of a disagreement score is as a proxy for predictive risk: flag a case for review when its explanations disagree, on the assumption that disagreement co-occurs with the model being wrong. We tested this directly rather than assuming it. For each case we computed a disagreement score (1 – the mean pairwise Kendall τ across the three gradient methods) and a misclassification indicator (predicted class, at $p = 0.5$, against the true label), and regressed the latter on the former with prediction confidence ($|\text{margin}|$, rescaled 0–1) as a covariate, using the same robust-variance GEE approach as the disagreement-prediction model above; each case is its own cluster here, since this analysis uses one prediction per subject rather than repeated method-pair rows.

On real data pooled across the 13 cohorts ($n = 3,147$; misclassification rate 16.9%), disagreement significantly predicted misclassification (coefficient – 1.36, 95% CI – 2.01 to – 0.72, $p < 0.001$) – in the opposite direction to the one a reader would likely expect: higher disagreement was associated with lower, not higher, odds of misclassification, net of confidence. Confidence was the stronger predictor, in the expected direction (coefficient – 2.20, 95% CI – 2.56 to – 1.85, $p < 0.001$): more confident predictions were less often wrong. The same sign held, without reaching significance, on the smaller synthetic demo cohort ($n = 180$; disagreement coefficient – 1.26, $p = 0.175$).

We probed this finding before reporting it. Disagreement and confidence were only weakly correlated (Pearson $r = 0.07$), so the reversed sign is not an artefact of collinearity: the same negative, significant coefficient (– 1.55, $p < 0.001$) appears when confidence is dropped from the model entirely. The direction was consistent in 9 of the 13 real cohorts, individually significant in 2 (eICU-CRD aneurysmal subarachnoid haemorrhage, MIMIC-III traumatic brain injury craniotomy); the remaining 4 cohorts showed a positive, non-significant coefficient, so we characterise the effect as directionally consistent rather than universal. An interaction model suggested the effect is not uniform across confidence levels: the disagreement–misclassification association, and the fall in raw misclassification rate across disagreement deciles, was markedly cleaner within the high-confidence tercile (misclassification fell from 11.1% to 1.1% between the lowest and highest disagreement decile) than within the low-confidence tercile, where the same direction held but was noisier (disagreement $\times$ confidence interaction: –2.07, $p = 0.082$, not conventionally significant). Finally, a distribution-free check: shuffling the disagreement score

within cohort and refitting 200 times gave a null distribution centred at -0.46 (SD 0.28), against an observed coefficient of -1.55 — more extreme than all 200 shuffles (permutation $p < 0.005$) so the association does not depend on the GEE's parametric variance assumptions.

This analysis was not pre-registered, and we report both it and the checks above as exploratory. We do not have a confirmed mechanistic account of why disagreement and misclassification move in opposite directions rather than being unrelated; the confidence-dependent pattern above is consistent with both saturating together on the easiest, most confidently-classified cases (where gradients are small and similar in shape across methods), but this is a plausible reading of the diagnostics, not a tested mechanism. What the result does establish, cautiously, is a negative one that survives the checks we ran: this study finds no evidence that explanation disagreement is a usable proxy for predictive risk, and some evidence that treating it as one (particularly for confidently-made predictions) could be actively misleading.

What a confirmatory reader study would require

The decision-flip analysis above uses a structural action mapping. Establishing that flips matter clinically requires clinician readers, and we used the observed flip rates to size that study. With 12 readers, 80% power is reached only if the true flip rate is at least ~ 0.16 ; at a true rate of 0.12 even 16 readers achieve 29% power (Fig. 3c). A reader study of conventional size can therefore confirm consequential disagreement between distant method pairs, but is underpowered to resolve near-equivalent pairs a design constraint worth knowing before such a study is run rather than after.

A clinician vignette survey corroborates the disagreement structure

To obtain independent, method-independent evidence on whether the structural action mapping picks out points of genuine clinical judgement, we surveyed 35 practising clinicians (Neurosurgery, $n = 12$; Neurocritical Care, $n = 12$; General Critical Care, $n = 11$; experience 1–5 years, $n = 12$; 6–10 years, $n = 11$; >10 years, $n = 12$) using a 20-item instrument built from 8 underlying bedside vignettes, each presented one to four times within the instrument to assess response consistency. Test–retest consistency across repeated presentations of the same vignette was $244/245$ (99.6%).

Three of the eight vignettes drew near-unanimous responses (plurality 97.1–100%; pairwise inter-rater agreement 94.3–100%), and five drew substantial disagreement (plurality 57.1–68.6%; 31.4–42.9% of respondents chose an action other than the plurality choice; pairwise agreement 40.2–55.6%) (Supplementary Table 1). This bimodal structure (most vignettes either near-unanimous or genuinely contested, few in between) mirrors the far/close structure of the decision-flip analysis above and supports the face validity of the underlying action space, though it is not a formal test of any specific method pair's flip rate.

For four of the five contested vignettes, disagreement was structured by specialty rather than distributed as noise within groups: one specialty group chose a given action unanimously or near-unanimously while

the other two converged on a different one (e.g., for the vasopressor-versus-fluids vignette, 11/11 General Critical Care respondents chose intravenous fluids while 12/12 Neurocritical Care and 12/12 Neurosurgery respondents chose vasopressors). The remaining contested vignette showed a within-specialty split concentrated in Neurosurgery, which favoured obtaining a stat CT scan where the other two specialties favoured adjusting sedation or oxygenation. This is a patterned form of clinical disagreement rather than random variability, and it is consistent with (though does not itself establish) the possibility that some attribution disagreement reflects genuine differences in which features different clinical schools of thought weight, rather than instability in the attribution methods alone.

This survey is a vignette-based bedside-action consensus study, not the paired reader study sized above: respondents were not shown attribution outputs and asked whether an explanation changed their recommended action, so it does not itself measure a decision-flip rate against any specific method pair. It is best read as external, method-independent evidence that the action categories used in the structural mapping correspond to points of real and sometimes specialty-patterned clinical disagreement, which strengthens confidence in the mapping's face validity while leaving the clinically adjudicated, attribution-linked reader study as necessary future work.

Discussion

Post-hoc explanations are being asked to carry regulatory and clinical weight that our results suggest they cannot currently bear unconditionally. On real ICU data across three independent databases, attribution methods applied to the same model agreed on roughly half of the feature ordering and never reached conventional thresholds for substantial agreement. On synthetic benchmarks where the truth is known, the identity of the most faithful method changed with the condition, to the point of exact rank reversal between two adjacent difficulty rungs. And a substantial fraction of disagreements translated into different recommended actions — a pattern that reproduced when we repeated the decision-flip analysis on real patients rather than the synthetic ladder alone.

The natural response pick the best method and standardise on it, is not supported by these data. That response presumes a method ordering stable enough to standardise on, and the ordering we observe is not: gradient saliency was the single best method for recovering a transient signal and the single worst for recovering an early one. Any recommendation derived from an average over conditions would be wrong in some identifiable subset of clinical cases, and would carry the authority of a benchmark while being so.

We think the more defensible use of an explanation audit is the one our results point to: characterising *when* explanations are likely to disagree rather than *which* method to use for a given case. Disagreement was partly predictable in advance, chiefly from feature correlation, one of three quantities we tested that are computable at prediction time without ground truth, alongside temporal concentration and prediction margin, neither of which reached significance, with out-of-sample AUC 0.73. This is a genuine, if partial, operationally useful finding, because a continuous disagreement-risk score can be computed for a

patient before an explanation is shown, even without a hard trust boundary. One caveat and one negative result attach immediately: the calibration slope of 0.93 indicates predicted probabilities are close to nominal, but our search for a trust regime, a stratum meeting the pre-registered substantial-agreement threshold, found none across nine strata (best mean $\tau = 0.30$). We report this negative result rather than the positive one we had originally expected, per our pre-registration commitment not to lower the threshold after seeing the data.

A related and distinct question is whether explanation disagreement predicts that the underlying model's prediction is wrong, as opposed to predicting that explanations disagree with each other; we tested this directly rather than leaving it as an open question (see Results). It does not, or rather, it does, but not in the direction a reader would expect. Net of prediction confidence, higher disagreement was associated with significantly lower odds of misclassification on real data ($n = 3,147$, $p < 0.001$), the opposite of what a risk-flagging use of disagreement would assume. We read this as a caution rather than a mechanism: it argues against repurposing an explanation-disagreement score as a proxy for predictive uncertainty, which is a different quantity that should be assessed on its own terms (via calibration and confidence, as above), not inferred from whether two attribution methods agree.

The finding that agreement is governed by data source rather than clinical population has a specific implication for how explanation evidence should be submitted and reviewed. It means an explanation audit is not a property of a model that travels with it. A vendor demonstrating explanation stability on one database has demonstrated very little about the behaviour of the same model and the same explanation method at a site whose data was collected differently, which is precisely the setting in which regulatory assurance is sought. Existing reporting frameworks for clinical AI ask for transparency and for external validation of predictive performance;^{1,2} our results argue that explanation behaviour requires the same external-validation discipline, and that it does not come for free with predictive external validation.

These findings extend to physiological time series a concern established in imaging, where saliency methods benchmarked against expert annotation underperformed clinicians, and in the machine-learning literature, where the disagreement problem has been documented but its clinical consequence left unquantified.⁸ The contribution here is not the observation that methods disagree, that is by now expected, but the calibration of that disagreement against action, the assessment of faithfulness against constructed ground truth under an admissibility gate, and the demonstration that disagreement is predictable rather than capricious.

Our results also carry a methodological warning that generalises past this study. In constructing the faithfulness benchmark we found that zero-ablation is not a null intervention where the injected effect is multiplicative: setting causal cells to zero leaves residual signal, and a model can score above chance on the "ablated" input because of the ablation operator itself. Across our 14 benchmark cells zero-ablation left more residual discrimination than counterfactual ablation in 10, and in one cell it crossed the admissibility boundary that the counterfactual operator does not (0.62 versus 0.53). The same objection

applies to zero-baseline Integrated Gradients and to occlusion-based attribution, both of which are in wide use. We gate on the counterfactual and report both.

The decision-flip finding is also not confined to the synthetic ladder: repeating it on the largest real cohort reproduced the same far/close pattern under the real 6-channel vital-sign panel (67–69% versus 7%, against 62–66% versus 11% on the synthetic benchmark). This strengthens confidence that the qualitative structure of the finding – some method pairs disagree consequentially, others do not, is a property of the methods rather than of the synthetic construction. It does not resolve whether the underlying action mapping is clinically correct, which is a separate question from whether the disagreement it detects is real; that question is unchanged by moving from synthetic to real data and remains the province of the clinician reader study sized above, which has not been run. A separate, method-independent vignette survey of 35 clinicians (see Results) found that the action categories used in the mapping correspond to points of real clinical disagreement (often patterned by specialty) for a majority of tested vignettes, which supports the mapping's face validity without substituting for the attribution-linked reader study itself.

Three independent lines of evidence bear on clinical consequence: the structural action mapping on synthetic data (up to 66% flip rate for the far pairs), its replication on a real cohort using the real clinical feature panel (67–69% versus 7%), and the clinician vignette survey's demonstration that the underlying action categories correspond to genuine, often specialty-patterned clinical disagreement. None of these three, individually or combined, establishes that a specific attribution method pair's disagreement changes what a clinician would actually do when shown that disagreement for a specific patient, only the attribution-linked reader study sized above can establish that. What they do establish, cumulatively, is that the disagreement is not a numerical curiosity: it lands on an action space that real clinicians treat as consequential and, in a majority of tested vignettes, contested.

Several further constraints bound what can be concluded, and we state them explicitly because some of them limit the headline numbers. Real-cohort replication of the decision-flip finding used a single database and population, the largest available real cohort, chosen for statistical power rather than clinical distinctiveness, so it demonstrates that the effect is not synthetic-only without establishing that it holds across every population in Table 1. Faithfulness evidence remains synthetic by necessity: real clinical data affords no ground truth about true causal drivers, so all faithfulness results, including the rank reversal, come from constructed cohorts, and their realism is bounded by the injection model. The attribution panel is narrower than the clinical field: we report three gradient-based methods, and TimeSHAP, DeepSHAP, LIME and attention weights, the last two among the most widely deployed in clinical work – are not included in the real-data panel; because our central finding is that method ranking is condition-dependent, a wider panel could only strengthen the claim of instability, but it could also reveal that some excluded method behaves more consistently, which our data cannot rule out. Cohorts are neurocritical and neurosurgical, and the outcome is in-hospital mortality, a narrower clinical scope than "ICU deterioration" generally, and generalisation to sepsis onset under consensus definitions¹⁷ or to benchmark deterioration tasks¹⁸ is untested; outcome definitions differ between the MIMIC and eICU

database families, which is one plausible contributor to the source effect we report. The trust model substitutes an estimating-equations fit for the registered mixed-effects analysis; its calibration slope of 0.93 is reasonably close to nominal, but its own trust-regime search returned a negative result (see Results), which is a more fundamental limitation than calibration. One of the 14 loaded cohort–database combinations (eICU-CRD spinal fusion and decompression) was excluded by the leakage audit, and a fifteenth (eICU-CRD brain tumour resection) never loaded at all; we report both rather than silently dropping them, but it means the eICU population is represented by fewer cohorts than the other two databases. The clinician vignette survey strengthens face validity for the action mapping but has its own scope limits: 35 respondents and 8 underlying vignettes drawn from the same intervention taxonomy as the structural mapping, not an independent clinical ontology, and it does not present attribution outputs, so it cannot speak to whether any specific method pair's flips are clinically consequential, only the attribution-linked reader study sized above can do that. Separately, this study establishes that explanations disagree, that disagreement is predictable, that it is often consequential for the recommended action, and that it does not positively predict misclassification, if anything the reverse. That last analysis was exploratory and not pre-registered, used the model's own predicted class rather than a clinician-adjudicated error, and its GEE clusters at the case level rather than the subject level (see Methods); it should be treated as a caution against one specific downstream use of a disagreement score, not as a validated characterisation of the disagreement–error relationship.

An explanation that changes with the method used to produce it is not a description of the model. Across 13 real ICU cohorts, three databases and a gated synthetic benchmark, we find that post-hoc attributions on clinical time series disagree substantially, that the disagreement is often consequential and reproduces on real patient data, that no method is reliably most faithful, and that the disagreement is nonetheless predictable in advance. We therefore argue against standardising on a single explanation method and in favour of requiring, as part of the evidence accompanying a clinical model, a consequence-calibrated audit of the kind reported here, externally validated across data sources, because explanation behaviour does not transfer between them.

Methods

Data sources and cohort construction

We used three independently collected, de-identified critical-care databases: MIMIC-III,¹⁹ MIMIC-IV²⁰ and the eICU Collaborative Research Database.²¹ Five neurocritical and neurosurgical populations were defined by diagnosis and procedure codes: aneurysmal subarachnoid haemorrhage, brain tumour resection, spinal fusion and decompression, spontaneous intracerebral haemorrhage, and traumatic brain injury craniotomy. Crossing five populations with three databases gave 15 candidate cohort–database combinations, one of which (eICU-CRD brain tumour resection) could not be extracted because its source file contained zero records; the remaining 14 loaded and were carried to the leakage audit (see Results for the corrected accounting of how these 14 became the 13 cohorts reported in Table 1).

Features were routinely recorded vital signs resampled onto a fixed hourly grid over the observation window, forming an (N, T, F) tensor per cohort. The outcome was in-hospital mortality. Outcome ascertainment differs between the MIMIC and eICU database families, reflecting how each records discharge disposition; this difference is a candidate explanation for the source effect reported in Results and is stated as a limitation rather than adjusted away.

One stay was retained per subject, so all splits are genuinely at the subject level and no patient appears in more than one fold. Cohorts with fewer than 15 positive outcomes were excluded from model fitting as underpowered for evaluation.

Within-window missingness in the six resampled vital-sign channels was handled by zero-imputation paired with a per-feature missingness-flag channel ($F = 12$: six values plus six flags), so the model can distinguish an absent reading from a true zero. Complete absence of vitals across the entire window — a stricter, subject-level measure — was rare and varied by cohort: 0.05–1.59% of subjects (mean 0.78%) across the 13 reported cohorts. We did not apply a correction for multiple comparisons across the 13-cohort agreement summary in Table 1, because it is reported descriptively rather than as a family of hypothesis tests; the study's inferential analyses — three pre-registered decision-flip non-inferiority tests per dataset and one disagreement-prediction model — were each pre-specified with their own margin or threshold rather than drawn from a larger family of comparisons that would require correction.

Leakage audit

Every cohort passed a pre-registered leakage audit before any model was fitted or any explanation computed. The audit tests for outcome information reachable from the feature window that would let the model exploit a shortcut — in which case attributions would faithfully describe the shortcut rather than clinical physiology. One of 14 cohorts (eICU spinal fusion and decompression) failed and was excluded from all downstream analysis. Failures are reported rather than silently dropped.

Predictive model

For each cohort we fitted a multilayer perceptron on the flattened observation window. Splits were by subject; feature standardisation used development-fold statistics only, so no test-fold information entered preprocessing. Discrimination (AUROC), precision-recall performance (AUPRC) and calibration (Brier score) are reported per cohort in Table 1. On the synthetic benchmarks a gradient-boosted model was fitted alongside the perceptron, giving $7 \text{ rungs} \times 2 \text{ model families} = 14 \text{ benchmark cells}$.

Attribution methods

The real-data panel comprises three gradient-based attribution methods — gradient saliency, Gradient×Input, and Integrated Gradients⁴ — together with two floor controls: randomly permuted attributions and constant attributions. The floors calibrate the agreement scale: any similarity statistic must be interpreted against what unrelated attributions produce on the same data, which we measured rather than assumed.

Attributions are computed per instance as (T, F) maps. Feature-level importance per instance is the sum of absolute attribution over time, a pre-registered aggregation; temporal importance is the corresponding sum over features. TimeSHAP,²² DeepSHAP, LIME⁵ and attention weights are specified in the pre-registration but are not implemented in the present panel; their absence is a stated limitation.

Agreement, stability and faithfulness metrics

Cross-method agreement is Kendall's τ between per-instance feature-importance rankings, computed for every method pair on every patient and averaged. Confidence intervals resample the patient axis. We use $\tau = 0.70$ as the conventional threshold for substantial agreement and report the empirical chance floor alongside every estimate.

Stability is the agreement of a method with itself under input perturbation.

Faithfulness is average precision (AP) of the attribution map against the known ground-truth mask, and is computed only on synthetic cohorts. On real data no ground truth exists, so no faithfulness number is reported there — a constraint we regard as fundamental rather than incidental: real clinical data can support claims about whether methods agree, but not about whether any of them is right.

Semi-synthetic ground truth and the admissibility gate

Seven rungs of graded difficulty were constructed by injecting known causal structure into realistic physiological trajectories: an easy case, a near-threshold case, and cases exercising transient signals, early signals, multiplicative feature interactions, redundant features and distractor features. The injected mask defines ground truth.

Each rung-by-model cell must pass a manipulation check before its attributions are scored. The model must discriminate on the cohort, and that discrimination must disappear when the injected causal cells are ablated. Two ablation operators are computed and the distinction is material. Zero-ablation sets causal cells to zero; counterfactual ablation restores the clean pre-injection values. Where the injected effect is multiplicative, zeroing is not a null intervention — with deviations a and b from the clean trajectory, zeroing gives $a \cdot b = x_{\text{clean}}^2 > 0$, so the "ablated" input still carries signal. Across the 14 benchmark cells, zero-ablation left the model closer to discriminating than counterfactual ablation did in 10 of 14 cells (mean ablated AUROC 0.516 versus 0.495), and in one cell it crossed the admissibility boundary: the gradient-boosted model on the easy rung reached 0.619 under zero-ablation against 0.530 under the counterfactual. We gate on the counterfactual operator and report both. All 14 cells passed the gate under the counterfactual operator, with ablated AUROC between 0.36 and 0.55.

This has implications beyond our benchmark: zero-baseline Integrated Gradients and occlusion attribution rest on the same assumption that a zeroed input is a signal-absent input, and that assumption fails wherever effects are multiplicative.

Decision-flip analysis

Each attribution is mapped to the intervention it most implicates, and a flip is recorded when two methods on the same patient implicate different interventions. The flip rate is the proportion of cases flipping, with 95% confidence intervals from a subject-clustered bootstrap. A pre-registered non-inferiority margin of $\delta = 0.10$ defines consequential disagreement: a pair exceeds the margin only when the lower bound of the interval exceeds δ . The demonstration cell is selected programmatically as the hardest admissible cell rather than hardcoded, so that no analysis runs on a benchmark that failed its own manipulation check.

The feature-to-action mapping is structural and has not been adjudicated by intensivists. It is a stated limitation, and the reader study sized below is the design that would replace it with clinically validated flips.

Real-data replication of the decision-flip analysis

To test whether the decision-flip finding depends on the synthetic construction, we repeated it on real attributions from the largest real cohort passing the leakage audit (MIMIC-IV spinal fusion and decompression; $n = 723$ test cases), selected by test-set size rather than by clinical distinctiveness. This cohort's model uses a 6-channel real vital-sign panel (heart rate, systolic blood pressure, mean arterial pressure, respiratory rate, oxygen saturation, temperature) rather than the synthetic cohort's generic feature slots; missingness-indicator channels paired with each vital were excluded from the ranking and the action mapping, since a missingness flag is not a physiological signal a clinician acts on. A separate structural action mapping was defined for this real panel, following the same taxonomy of interventions (fluids, vasopressor support, antibiotic escalation, respiratory support, monitoring) as the synthetic-cohort mapping; it is, like the synthetic mapping, a placeholder that has not been clinically adjudicated. Flip rates and bootstrap confidence intervals were computed identically to the synthetic-cohort analysis above.

Predicting disagreement

Per-case disagreement was modelled from three pre-registered covariates computable at prediction time without ground truth: the correlation among candidate features, the temporal concentration of the signal, and the model's prediction margin. The pre-registration specified a mixed-effects model; we fitted generalised estimating equations with an exchangeable working correlation and subject-level clustering, and record the substitution as a deviation. Coefficients are reported per standard deviation of the predictor, because the raw scale of temporal concentration is narrow enough that its unstandardised coefficient (121) is a scaling artifact rather than a large effect. Validation used held-out cases ($n = 270$) and reports AUC, Brier score, and calibration intercept and slope.

Disagreement and prediction correctness

For each case we defined a disagreement score ($1 -$ the mean pairwise Kendall τ across the three gradient methods) and a misclassification indicator (predicted class at $p = 0.5$ against the true label),

and fit a GEE logistic model of misclassification on disagreement with prediction confidence ($|\text{margin}|$, rescaled 0–1) as a covariate. This is an exploratory analysis, not part of the pre-registration. Because it uses one prediction per subject rather than repeated method-pair rows, each case constitutes its own GEE cluster; the estimator is therefore equivalent to a robust-variance logistic regression rather than one that pools information across repeated measurements on the same subject, unlike the disagreement-prediction model above. We additionally checked this result for collinearity between covariates, sign stability with confidence removed, consistency across cohorts, a confidence-stratified nonparametric check, and a within-cohort permutation test (200 shuffles) as a distribution-free check on the GEE p-value; all are reported alongside the main result (see Results).

Reader-study power analysis

Power to detect consequential flips was computed across reader panel sizes (8, 10, 12, 16) and true flip rates (0.12–0.25), at the pre-registered $\delta = 0.10$ margin, to size the confirmatory clinician study.

Clinician vignette survey

To obtain independent evidence on the face validity of the structural action mapping, we administered a 20-item survey to 35 practising clinicians recruited across three specialties (Neurosurgery, Neurocritical Care, General Critical Care) and a range of experience (1–5, 6–10, > 10 years). The instrument comprised 8 underlying bedside vignettes constructed from the same taxonomy of interventions used in the decision-flip analysis (fluids, vasopressor support, antihypertensive therapy, respiratory support, sedation adjustment, and diagnostic imaging); each vignette was presented one to four times within the instrument, at non-adjacent positions, to assess within-respondent response consistency. For each presentation, respondents selected the single action they would take next. This is a bedside-action consensus survey rather than the paired, attribution-linked reader study sized in the previous subsection: respondents were not shown model outputs or attribution maps, and the survey does not measure a decision-flip rate against a specific method pair. Plurality, dissent, test–retest consistency and pairwise inter-rater agreement (the probability that two randomly chosen respondents selected the same action for a given vignette) were computed per vignette and are reported in Supplementary Table 1.

Reporting and reproducibility

The study follows TRIPOD + AI reporting guidance for clinical prediction models² and the FUTURE-AI principles for trustworthy clinical AI.¹ Analyses were pre-registered before explanations were generated; deviations from the pre-registration (the estimating-equations substitution, and the attribution-panel restriction) are recorded above and in the amendment log accompanying the instrument. The audit is distributed as a version-controlled Python package with executable notebooks and deterministic scripts for reproducing the reported analyses, figures and tables; the computational environment is pinned (Python 3.12.13, numpy 2.0.2, pandas 2.3.3, scipy 1.16.3, scikit-learn 1.6.1, statsmodels 0.14.6, xgboost 3.2.0, matplotlib 3.10.0), with a single global random seed (0) applied throughout data splitting, bootstrap resampling and model initialisation. Reproducibility of the real-data analyses is subject to

independent credentialled access to MIMIC-III, MIMIC-IV and eICU-CRD; we do not claim bit-for-bit reproduction beyond the validated environment and documented data-access conditions.

Declarations

Ethics approval and consent to participate

This study is a secondary analysis of de-identified data from MIMIC-III, MIMIC-IV and the eICU Collaborative Research Database. The collection and sharing of MIMIC-III and MIMIC-IV were approved by the Institutional Review Boards of Beth Israel Deaconess Medical Center and the Massachusetts Institute of Technology, which granted a waiver of informed consent for these research resources; eICU-CRD was independently certified as meeting HIPAA Safe Harbor de-identification standards, and secondary analysis of it is exempt from further institutional review. All authors completed the required human-subjects research training and accessed the data under PhysioNet data use agreements; no attempt was made to re-identify any patient, and no patient-level data are reported.

The clinician vignette survey reported in this manuscript involved prospectively collected responses from human participants and therefore required its own ethics review, separate from the data-use terms governing the retrospective cohorts above. Ethical approval for the clinician vignette survey was obtained from the Research Ethics Committee at the Faculty of Medicine, Port Said University, and the study was conducted in accordance with the Declaration of Helsinki. A waiver of informed written consent was granted on the basis that the questionnaire was anonymous; participants were briefed on the study's objectives and on the confidentiality of their responses before proceeding. Sampling was by digital snowball distribution and was fully anonymous throughout.

Consent for publication

Not applicable. This manuscript reports only aggregate, de-identified results from public critical-care databases and an anonymous clinician survey; no individual person's data, images, or identifiable case details are presented.

Competing interests

The author declares that he has no competing interests.

Funding

Self-funded

Authors' contributions

A.M.F. is the sole author of this manuscript and is responsible for study conception and design, data acquisition and cohort construction, development of the audit instrument, statistical analysis,

interpretation of results, and drafting and revising the manuscript. A.M.F. read and approved the final manuscript.

Acknowledgements

Not applicable.

Data availability

MIMIC-III, MIMIC-IV and eICU-CRD are available to credentialed researchers via PhysioNet under their respective data use agreements. No patient-level data are redistributed with this work. Cohort definitions, extraction code and analysis configuration are included in the released repository (see Code availability). De-identified, respondent-level results of the clinician vignette survey (summarised in Supplementary Table 1; the item-level response file contains no patient data) are included with the released instrument.

Code availability

The source code implementing the consequence-calibrated explanation audit – cohort processing, leakage auditing, predictive modelling, attribution analysis, semi-synthetic benchmark generation and admissibility checks, agreement and stability metrics, decision-flip analysis, the disagreement-prediction model, reader-study power analysis, and manuscript figure/table reproduction – is publicly available at <https://github.com/ahmedfawaz879/explanation-audit>. Patient-level MIMIC-III, MIMIC-IV and eICU-CRD data are not redistributed and must be obtained independently under their respective data-use agreements.

References

1. Lekadir K, Frangi AF, Porras AR, et al. FUTURE-AI: international consensus guideline for trustworthy and deployable artificial intelligence in healthcare. *BMJ*. 2025. 10.1136/bmj-2024-081554. . PMID 39909534.
2. Collins GS, Moons KGM, Dhiman P et al. TRIPOD + AI statement: updated guidance for reporting clinical prediction models that use regression or machine learning methods. *BMJ* 2024. 10.1136/bmj-2023-078378. PMID 38626948.
3. Lundberg SM, Lee S-I. A unified approach to interpreting model predictions. *Adv Neural Inf Process Syst* 2017; arXiv:1705.07874.
4. Sundararajan M, Taly A, Yan Q. Axiomatic attribution for deep networks. *International Conference on Machine Learning* 2017; arXiv:1703.01365.
5. Ribeiro MT, Singh S, Guestrin C. Why should I trust you? explaining the predictions of any classifier. *KDD* 2016; arXiv:1602.04938.

6. Lundberg SM, Nair B, Vavilala MS, et al. Explainable machine-learning predictions for the prevention of hypoxaemia during surgery. *Nat Biomed Eng*. 2018. 10.1038/s41551-018-0304-0. . PMID 31001455.
7. Rudin C. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nat Mach Intell*. 2019. 10.1038/s42256-019-0048-x. . PMID 35603010.
8. Krishna S, Han T, Gu A et al. The disagreement problem in explainable machine learning: a practitioner's perspective. *arXiv:2202.01602* 2022.
9. Adebayo J, Gilmer J, Muelly M et al. Sanity checks for saliency maps. *Adv Neural Inf Process Syst* 2018; *arXiv:1810.03292*.
10. Jain S, Wallace BC. Attention is not explanation. *NAACL-HLT 2019*; *arXiv:1902.10186*.
11. Amann J, Blasimme A, Vayena E, et al. Explainability for artificial intelligence in healthcare: a multidisciplinary perspective. *BMC Med Inf Decis Mak*. 2020. 10.1186/s12911-020-01332-6. . PMID 33256715.
12. Ghassemi M, Oakden-Rayner L, Beam AL. The false hope of current approaches to explainable artificial intelligence in health care. *Lancet Digit Health*. 2021. 10.1016/S2589-7500(21)00208-9. . PMID 34711379.
13. Tomašev N, Glorot X, Rae JW, et al. A clinically applicable approach to continuous prediction of future acute kidney injury. *Nature*. 2019. 10.1038/s41586-019-1390-1. . PMID 31367026.
14. Churpek MM, Carey KA, Snyder A, et al. Multicenter development and prospective validation of eCARTv5: a gradient-boosted machine-learning early warning score. *Crit Care Explor*. 2025. 10.1097/CCE.0000000000001232. . PMID 40138535.
15. Wong A, Otles E, Donnelly JP et al. External validation of a widely implemented proprietary sepsis prediction model in hospitalized patients. *JAMA Intern Med* 2021. 10.1001/jamainternmed.2021.2626. PMID 34152373.
16. Rajkomar A, Oren E, Chen K, et al. Scalable and accurate deep learning with electronic health records. *NPJ Digit Med*. 2018. 10.1038/s41746-018-0029-1. . PMID 31304302.
17. Singer M, Deutschman CS, Seymour CW et al. The Third International Consensus Definitions for Sepsis and Septic Shock (Sepsis-3). *JAMA* 2016. 10.1001/jama.2016.0287. PMID 26903338.
18. Reyna MA, Josef CS, Jeter R, et al. Early prediction of sepsis from clinical data: the PhysioNet/Computing in Cardiology Challenge 2019. *Crit Care Med*. 2020. 10.1097/CCM.0000000000004145. . PMID 31939789.
19. Johnson AEW, Pollard TJ, Shen L, et al. MIMIC-III, a freely accessible critical care database. *Sci Data*. 2016;3:160035. 10.1038/sdata.2016.35.
20. Johnson AEW, Bulgarelli L, Shen L, et al. MIMIC-IV, a freely accessible electronic health record dataset. *Sci Data*. 2023. 10.1038/s41597-022-01899-x. . PMID 36596836.
21. Pollard TJ, Johnson AEW, Raffa JD et al. The eICU Collaborative Research Database, a freely available multi-center database for critical care research. *Sci Data* 2018. 10.1038/sdata.2018.178.

22. Bento J, Saleiro P, Cruz AF, Figueiredo MAT, Bizarro P. TimeSHAP: explaining recurrent models through sequence perturbations. *KDD 2021*. 10.1145/3447548.3467166. arXiv:2012.00073.

Figures

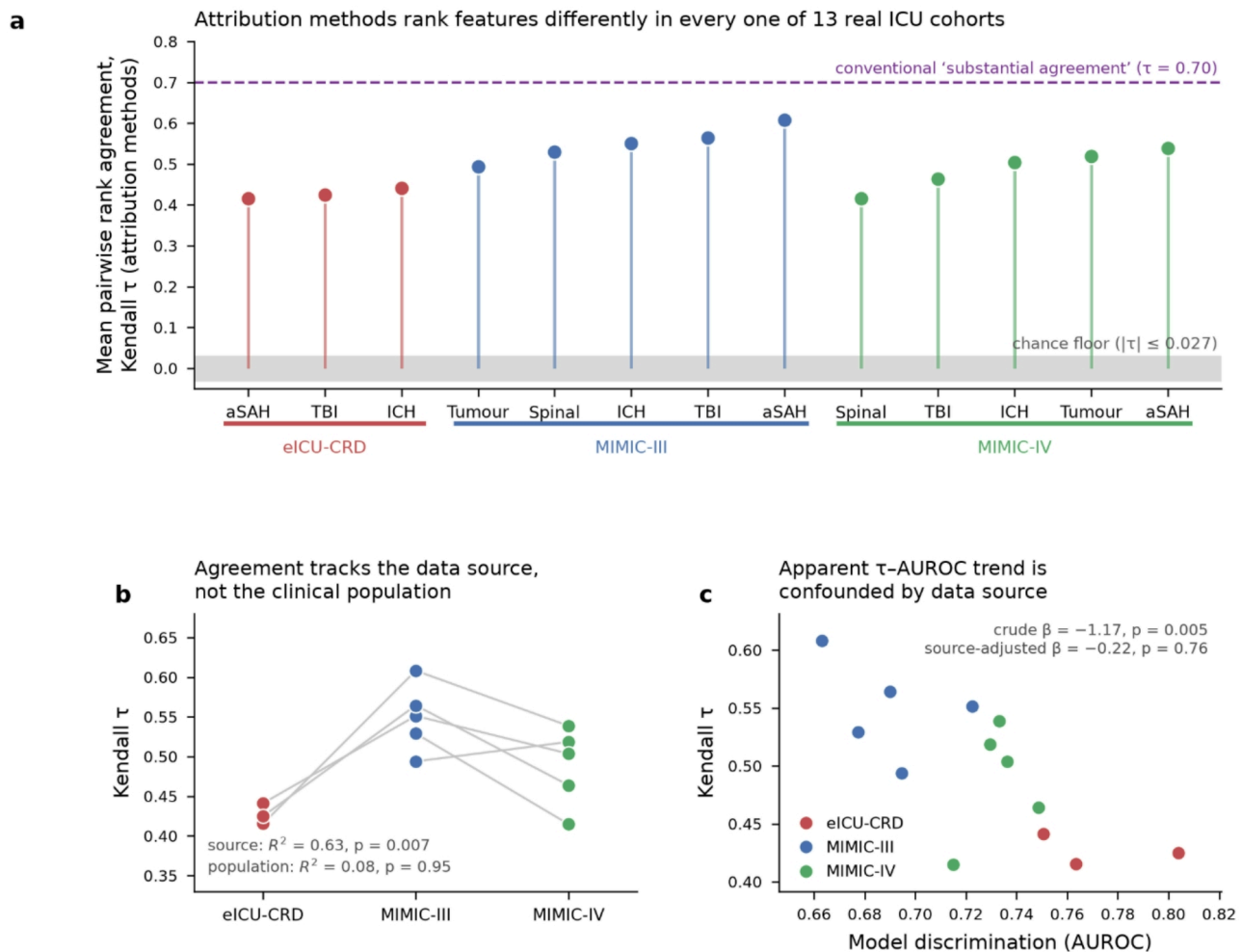


Figure 1

Attribution methods agree only moderately on real ICU data, and agreement tracks the data source rather than the clinical population.

a, Mean pairwise cross-method rank agreement (Kendall τ) for each of 13 cohorts, grouped by source database (colour bands beneath the axis). The dashed line marks the conventional threshold for substantial agreement ($\tau = 0.70$); no cohort reaches it. The grey band marks the empirical chance floor established by randomised control attributions ($|\tau| \leq 0.027$). **b**, The same cohorts arranged by database,

with grey lines connecting the same clinical population across sources. Source database explains 63% of between-cohort variance in agreement ($p = 0.007$); clinical population explains 8% ($p = 0.95$). **c**, Agreement against model discrimination. The crude association is significant ($\beta = -1.17$, $p = 0.005$) but collapses after adjustment for source database ($\beta = -0.22$, $p = 0.76$) and does not reproduce within any single database.

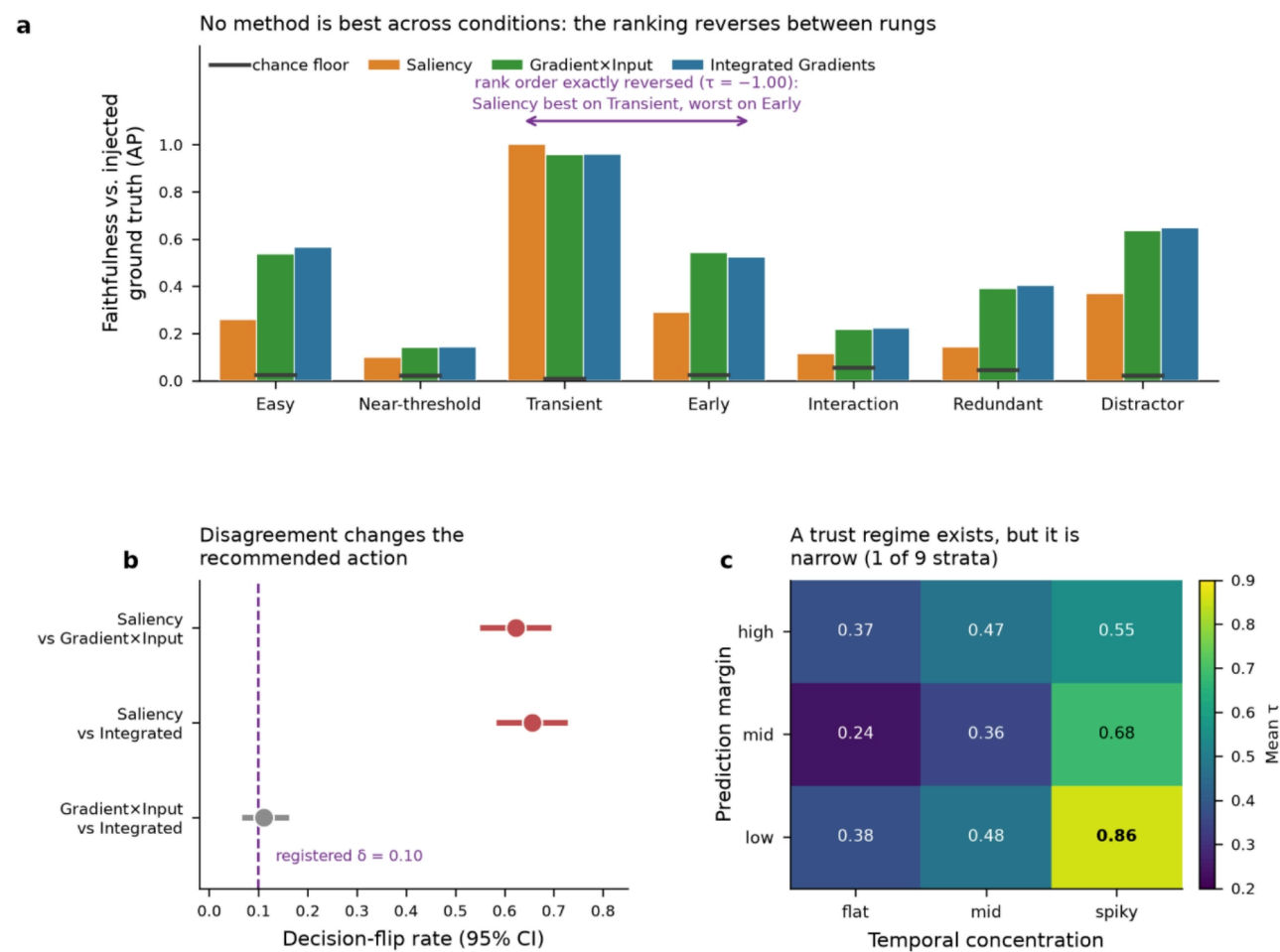


Figure 2

On synthetic benchmarks with known ground truth, no method is reliably most faithful, and disagreement changes the recommended action.

a, Faithfulness (average precision against the injected ground-truth mask) for three attribution methods across seven rungs of graded difficulty, all of which passed the counterfactual-ablation admissibility gate. Method ranking reverses exactly between the transient-signal and early-signal rungs ($\tau = -1.00$): gradient saliency is best on one and worst on the other. Floor and base-rate references are shown. **b**, Decision-flip rates between method pairs with 95% subject-clustered bootstrap intervals, against the pre-

registered non-inferiority margin $\delta = 0.10$. Two of three pairs exceed the margin; the same pairs exceed it when the analysis is repeated on real patient data (Results). **c**, Mean agreement stratified by prediction margin and temporal concentration: no stratum reached the pre-registered trust threshold ($\tau \geq 0.70$); the best stratum reached $\tau = 0.30$ ($n = 531$) against $\tau = 0.14$ ($n = 677$) in the worst. Panels report synthetic-ground-truth results and are not real-data estimates.

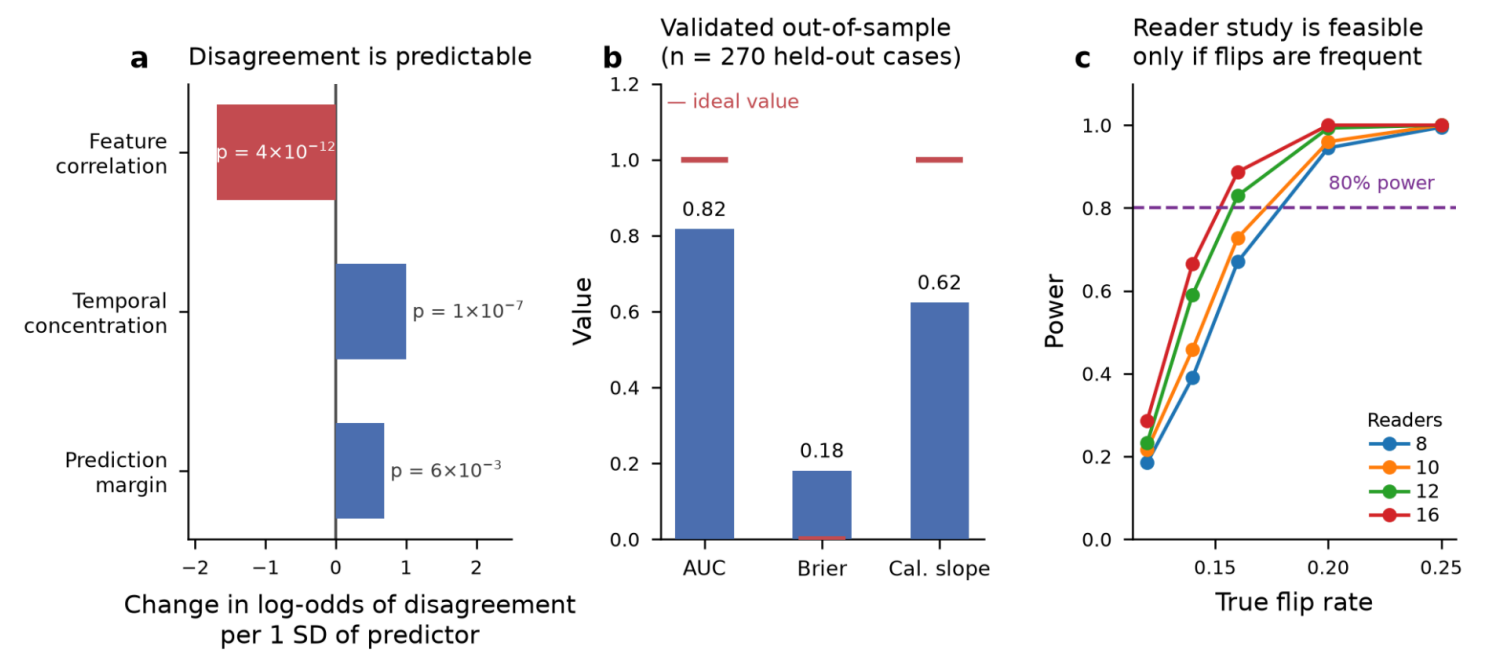


Figure 3

Disagreement is partly predictable in advance, but no case stratum meets the pre-registered trust threshold, and a confirmatory reader study is feasible only for distant method pairs.

a, Standardised effects on the log-odds of explanation disagreement from a generalised estimating equations model with subject-level clustering; effects are per 1 SD of each predictor. **b**, Out-of-sample validation on held-out cases ($n = 270$); red marks indicate the ideal value of each metric. The calibration slope of 0.93 indicates predicted probabilities close to nominal, with a modest positive intercept (0.12) indicating slight overprediction. **c**, Statistical power to detect consequential decision flips at the $\delta = 0.10$ margin, by reader panel size and true flip rate. With 12 readers, 80% power requires a true flip rate of approximately 0.16 or higher.

Supplementary Files

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

- [SupplementaryTable1ClinicianVignetteSurvey.docx](#)