DETECTING CARE: TREATMENT POLICY AND RETROSPECTIVE VALIDATION IN CLINICAL RISK MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Clinical risk models are trained on outcomes recorded under a treatment policy, so a model can learn to detect care rather than risk. On a published ICU benchmark, masking one clinician-set feature reduces AUROC more than masking any of six non-manipulable physiological features. In this paper, we prove that a risk score with no intrinsic predictive value that triggers effective treatment has retrospective AUROC strictly below 0.5, and we derive a closed-form solution for the deficit as a function of treatment policy and treatment effect. We also show that this deficit does not vanish as intrinsic quality increases, and as a result, retrospective AUROC is not comparable across sites with differing treatment policies. For empirical evidence, we split 116 eICU hospitals by severity-adjusted vasopressor practice and find that a clinician-set feature which improves same-practice external validation degrades by a similar amount on transfer to a different practice. Moreover, the model's calibration error is concentrated on untreated patients, which is exactly the group these models exist to identify. The failure we characterize is invisible to discrimination, calibration, and same-practice external validation and our results can be extended to any acted-upon score, clinical or otherwise.

1 INTRODUCTION

Machine learning models that predict deterioration in hospitalised and critically ill patients are now routinely published, benchmarked and deployed (Henry et al., 2015; Tomašev et al., 2019; Hyland et al., 2020; Escobar et al., 2020; Moor et al., 2023; van de Water et al., 2024). Every such model is fitted to data that clinicians generated while treating the very patients it describes, so both its labels and its features carry the imprint of care (Paxton et al., 2013; Agniel et al., 2018; Beaulieu-Jones et al., 2021; Zhang et al., 2025). A patient on 80% inspired oxygen has a recorded FiO_2 of 0.8 because a clinician set it, not because the patient's lungs produced it.

Throughout, we distinguish three things. A *treatment policy* is the function that maps a patient's state to the probability of being treated. A hospital's *treatment practice* is the specific treatment policy it follows. A *treatment rate* is the fraction of patients treated, and two hospitals can have the same treatment rate while following different treatment policies.

A clinical risk model trained under one treatment policy is often accurate on held-out data from its own institution. What it estimates, though, is risk under that institution's policy (Dyagilev & Saria, 2016; Sperrin et al., 2018; van Geloven et al., 2020), so it generalises only where the same policy is followed (van Geloven et al., 2020). Across 68 eICU hospitals with reliable vasopressor charting (Pollard et al., 2018), we find that the severity-adjusted odds of receiving a vasopressor differ between hospitals by a median odds ratio (MOR) of 3.21 (Larsen & Merlo, 2005). Held-out testing and calibration checks draw on data from the same treatment practice, so neither can reveal this dependence. Failures of deployed models on new populations are well documented (Futoma et al., 2020; Wong et al., 2021; Finlayson et al., 2021; Singh et al., 2022), but they are rarely attributed to a difference in treatment practice.

What is missing is a direct measurement of *policy sensitivity*. It quantifies how far a model's performance moves when the treatment policy it was trained under changes. Measuring it requires three properties that, to our knowledge, no prior work combines. First, treatment practice must be measured rather than simulated or hypothesised. So far, feedback and performative effects have been studied mainly in simulation or theory (Perdomo et al., 2020; Lenert et al., 2019; Liley et al., 2021;

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

Boeken et al., 2024; van Amsterdam et al., 2025; Butler et al., 2026). Second, the measurement must isolate the effect of treatment policy from that of measurement policy, which recent work has linked to cross-site shift (Agniel et al., 2018; Futoma et al., 2021; Yamamoto et al., 2026; Zou et al., 2026). Third, the split must be defined by measured treatment practice rather than by database or region, as in existing generalisation studies (Zhang et al., 2021; Singh et al., 2022; Spathis & Hyland, 2022; Moor et al., 2023; Rockenschaub et al., 2024).

We construct a design with all three properties and our contributions are as follows.

1. **A closed form for treatment-displaced AUROC.** We show that a score with no predictive value that triggers effective treatment has retrospective AUROC strictly below 0.5, with the deficit given exactly by treatment policy and treatment effect. The deficit does not vanish as the score improves, so retrospective AUROC cannot be compared across sites with different treatment policies.
2. **Treatment dependence in a published benchmark.** The preprocessing of GARLIC (Wang et al., 2026) misclassifies 13 of its 103 input channels by manipulability. Masking one clinician-set channel, $\mathrm{FiO_2}$, lowers AUROC by 1.08 pp which is comparable to the margin between the top two models on the benchmark.
3. **Policy sensitivity across real practice variation.** We split eICU hospitals at the median of severity-adjusted vasopressor use, into a high-use and a low-use stratum, and compare models trained with and without a vasopressor feature. For a model trained on the high-use hospitals, adding the feature raises AUROC on held-out high-use hospitals and lowers it by a similar amount on the low-use ones. After that transfer, observed deaths among untreated patients exceed the model's predictions by 23%, a discrepancy that calibration-in-the-large does not reveal.

All code, the full manipulability classification and per-run summaries are released as supplementary material.

2 Related Work

Treatment as a source of bias. The dependence of a risk model on the treatment policy that generated its labels appears in the literature under several names. These include censoring due to interventions (Dyagilev & Saria, 2016), treatment drop-in (Sperrin et al., 2018), performative prediction (Perdomo et al., 2020), feedback loops (Balcarcel et al., 2025) and self-fulfilling prophecy (Wilkinson, 2009; van Amsterdam et al., 2025; McCradden et al., 2025). A pneumonia risk model learned that asthmatic patients were at low risk because the treatment policy routed them directly to intensive care (Cooper et al., 1997; Caruana et al., 2015). A model that ignores treatment begun after baseline is estimating risk under the treatment policy of its training data, and so generalises only where that policy holds (van Geloven et al., 2020; Luijken et al., 2024). The proposed alternatives have causal blind spots of their own (van Geloven et al., 2025), and published models rarely respect the distinction (Prosepe et al., 2022). Schulam & Saria (2017) showed that risk models are sensitive to the policy that generated their training data, and Coston et al. (2020) show that the same dependence makes their evaluation misleading. Once deployed, a model can appear to degrade because the actions it triggers change the outcomes it predicts (Lenert et al., 2019; Sperrin et al., 2019; Balcarcel et al., 2025), and refitting on post-deployment data introduces further bias (Liley et al., 2021). In all the prior work, the evidence for policy dependence is simulated or analytical.

Formal results. For a binary feature, a binary treatment, a threshold policy and a constant, deterministic historical policy, van Amsterdam et al. (2025) give conditions under which AUROC does not fall after deployment even though the new policy harms a subgroup. Because their historical policy treats everyone or no one, their training data contain no treatment that depends on the features, which is the case our closed form addresses. Boeken et al. (2024) model deployment as causal domain shift and show that risk under the baseline policy is not identifiable from pre-deployment data without an auxiliary variable measured in the target domain. Our claim is complementary, since policy sensitivity can be measured through its consequences without being identified.

Treatment and measurement features. Features that record clinician actions carry information about care as well as physiology (Paxton et al., 2013; Agniel et al., 2018; Beaulieu-Jones et al., 2021; Sisk et al., 2021). Features that encode measurement policy are the better-studied half of this problem. An indicator of whether a variable was measured improves accuracy at the training site but can harm transfer (Futoma et al., 2021). A change in record-keeping alone degraded a deployed mortality model (Nestor et al., 2019), and features derived from the observation process are associated with larger shift between databases (Yamamoto et al., 2026; Zou et al., 2026). All three are clinical instances of shortcut learning (Geirhos et al., 2020). Treatment features have received less scrutiny. Moor et al. (2023) exclude treatment features from their inputs, yet their Sepsis-3 label still encodes vasopressor use through the SOFA score. Butler et al. (2026) describe the mechanism we study, a treatment feature whose policy differs between clinics, and recommend dropping the feature or reweighting it but their case is hypothetical and uses no data. To our knowledge, no study has measured what a treatment feature does to transfer when practice differs, or whether reweighting recovers it.

Metric stability under shift. Shift-aware metrics assume label shift (Flores et al., 2025), an assumption that a change of treatment policy violates, because it alters the relation between features and outcome. Causal taxonomies of metric stability (Pfohl et al., 2025) model no action on the outcome at all.

Absence of direct measurement. Logan Ellis et al. (2025) give a recent articulation of treatment masking, but their evidence is a single comparison of two features in isolation. Rockenschaub et al. (2024) attribute poor transfer to models that learn local investigation pathways rather than the underlying biology, but the explanation is never tested.

Cross-site generalisation. Cross-site studies split data by database (Moor et al., 2023), region (Zhang et al., 2021; Duan et al., 2026) or hospital (Singh et al., 2022; Spathis & Hyland, 2022; Lu & Nemati, 2025), and several name clinical or measurement practice among the likely causes of degradation (Singh et al., 2022; Lu & Nemati, 2025; Duan et al., 2026; Ji et al., 2026), but none splits by treatment practice. Duan et al. (2026) report mechanical ventilation rates from 24% in South Korea to 79% in China without attributing performance to them. Subbaswamy & Saria (2020) do isolate a practice variable, but they simulate it, and the variable is a measurement policy rather than a treatment policy.

3 METHODOLOGY

In this paper, we study a single cohort under a single treatment policy, evaluated once, with no deployment and no before-and-after comparison. Proofs are in Appendix A.

3.1 SETUP

A patient has state z, summarising everything that determines their prognosis. A score S has continuous distribution F, and $U = F(S)$ is its rank, uniform on $[0, 1]$. The *intrinsic quality* of the score is $q = \mathrm{corr}(S, z) \in [0, 1]$, its correlation with latent severity. A score with $q = 0$ carries no information about the patient. Absent treatment, the patient's risk is $p_0(z) \in [0, 1]$. Treatment $T \in \{0, 1\}$ reduces it to $p_0(z)\,(1 - e(z))$, where $e(z) \in [0, 1]$ is the treatment effect. The *treatment policy* is the function $t(s, z) = \mathbb{P}(T = 1 \mid S = s, z)$, and the *treatment rate* is the scalar $\bar{t} = \mathbb{E}[t]$. The observed outcome is Y, and the retrospective AUROC is $R = \mathbb{P}(S_1 > S_2 \mid Y_1 = 1, Y_2 = 0)$. The rate says how much treatment is given and the policy says where it is aimed. Only the aiming moves R.

3.2 AN IDENTITY

Lemma 1. *Let* $r(u) = \mathbb{P}(Y = 1 \mid U = u)$ *and* $P = \mathbb{E}[r(U)] \in (0, 1)$. *Then*

$$R \;=\; \tfrac{1}{2} \;+\; \frac{\mathrm{Cov}\big(U,\, r(U)\big)}{P\,(1 - P)}. \tag{1}$$

The identity is exact and classical. It requires no link function, no treatment model and no distributional form for S, and it holds for anti-predictive scores as well. For a binary outcome it is the covariance form of Somers' $D = 2R - 1$ (Bamber, 1975; Newson, 2002), equivalently the rank-biserial correlation (Cureton, 1956). We restate it because every result below is Lemma 1 evaluated at a particular r, and the choice of risk model affects only whether r has a closed form.

3.3 A SCORE WITH NO INTRINSIC PREDICTIVE VALUE

A score has *zero intrinsic quality* if $q = 0$, equivalently $S \perp\!\!\!\perp z$. Then $r(u) = \mu - \kappa(u)$, with $\mu = \mathbb{E}[p_0(z)]$ and

$$\kappa(u) \;=\; \mathbb{E}_z\big[\, p_0(z)\, e(z)\, t(F^{-1}(u), z)\,\big], \tag{2}$$

the expected absolute risk reduction delivered at rank u.

Proposition 1. *For a score of zero intrinsic quality,*

$$\tfrac{1}{2} - R \;=\; \frac{\operatorname{Cov}\big(U,\, \kappa(U)\big)}{P\,(1-P)}, \qquad P = \mu - \mathbb{E}[\kappa(U)]. \tag{3}$$

Hence $R < \frac{1}{2}$ exactly when $\operatorname{Cov}(U, \kappa(U)) > 0$, which holds whenever the delivered benefit κ is non-decreasing and non-constant in rank. $R = \frac{1}{2}$ when κ is constant.

Suppose the policy is separable, $t(s, z) = t(s)\,\tau(z)$ with $\mathbb{E}[\tau(z)] = 1$. Clinicians may then use their own reading of severity, provided it does not change how treatment varies with the score. Writing $\kappa = \mathbb{E}[p_0 e \tau]$ gives $\kappa(u) = \kappa\, t(F^{-1}(u))$, a treatment rate $\bar{t} = \mathbb{E}[t(S)]$, and

$$R \;=\; \tfrac{1}{2} \;-\; \frac{\kappa \operatorname{Cov}\big(U,\, t(S)\big)}{P(1-P)}, \qquad P = \mu - \kappa\,\bar{t}. \tag{4}$$

A score is therefore evaluated as worse than chance whenever the probability of treatment rises with it and the treatment works. No constraint is placed on the treatment rate.

3.4 A DETERMINISTIC THRESHOLD IS THE EXTREMAL CASE

Lemma 2. *Among all policies with rate $\bar{t}$, $\operatorname{Cov}(U, t(S))$ is maximised by the deterministic top-rank threshold $t = \mathbf{1}\{U > 1 - \bar{t}\}$, at the value $\bar{t}(1-\bar{t})/2$.*

An alert threshold therefore displaces AUROC further than any stochastic policy that treats the same fraction of patients. In the separable case this gives a bound on the displacement and, inverting it, a lower bound on the delivered benefit,

$$\tfrac{1}{2} - R \;\le\; \frac{\kappa\,\bar{t}(1-\bar{t})}{2P(1-P)}, \tag{5}$$

$$\kappa \;\ge\; \frac{2P(1-P)\big(\tfrac{1}{2} - R\big)}{\bar{t}(1-\bar{t})}. \tag{6}$$

Every quantity on the right of (6) is observed. These are the event rate, the treatment rate and the AUROC. The quantity bounded is the one that is not observed, so the bound can be applied to a published result without access to its data.

A published illustration. Logan Ellis et al. (2025) report that haemoglobin reaches an AUROC of only 0.49 for 24-hour mortality while age reaches 0.65, and attribute the gap to treatment masking. Reproducing the comparison in emergency admissions to MIMIC-IV (Johnson et al., 2023) and in eICU, we find that the relationship is U-shaped and that a spline recovers the discrimination the linear fit discards. Simulating masking adversarially, it could erase that non-monotonicity only if transfused patients had a counterfactual 24-hour mortality of 70.9%, against 3.15% observed. Appendix G gives the full analysis.

3.5 CLOSED FORMS

Appendix C gives closed forms for probit propensities, which approach the threshold value as the propensity sharpens, and extends them to scores of any quality.

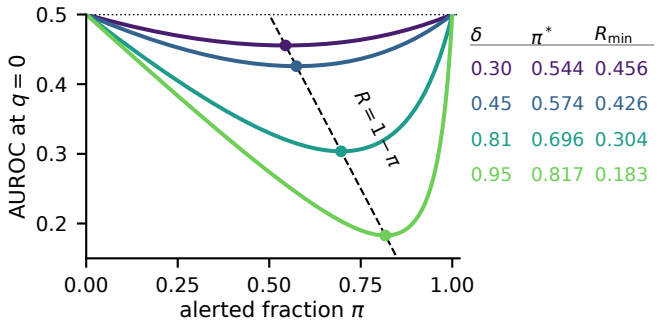


Figure 1: Retrospective AUROC of a score with no predictive value that triggers treatment, against the fraction alerted, for four treatment effects δ in the rare-outcome limit. Dots mark the minima, which lie on the dashed line $R = 1 - \pi$, and their coordinates are tabulated beside the plot.

The worst alert rate. The threshold policy fires when $S > \tau_0$, so the alerted fraction is $\pi = 1 - \Phi(\tau_0)$, independent of q. Let δ denote the product of treatment effect and adherence. For a rare outcome, minimising R over π gives

$$\pi^\star = \frac{1}{1 + \sqrt{1 - \delta}}, \qquad R_{\min} = 1 - \pi^\star. \tag{7}$$

In this limit the floor equals the fraction not alerted. The worst rate is not 50%. It moves toward treating everyone as the treatment improves (Figure 1). In the limit every treated patient survives, so the only events lie below the threshold and the ranking inverts exactly.

3.6 The displacement does not vanish

Let $\Delta(q)$ be the AUROC of the score with no effective treatment minus its AUROC when the treatment it triggers works.

Proposition 2. $\Delta(q) \geq \underline{\Delta} > 0$ *for every* $q \in [0, 1]$*, with* $\underline{\Delta}$ *explicit.*

The bound is explicit, and in clinical configurations it is loose by a factor of 3.8 to 7.0 against the exact minimum. The proof and the bound are in Appendix B.

3.7 Consequences and scope

Retrospective AUROC is not comparable across treatment policies. Two sites that evaluate the same score on the same case mix, under policies that differ only in how treatment tracks the score, report different values of R. By Proposition 1 the gap is exactly the difference in $\mathrm{Cov}(U, \kappa(U))/(P(1 - P))$, and by Proposition 2 it does not close as the score improves. Two hospitals that each treat 20% of patients, one the top 20% by score and the other 20% at random, have covariances of 0.08 and 0, so the rate alone does not determine the displacement. A difference in retrospective AUROC between sites is therefore not, on its own, evidence about the score. The displacement is absent or reversed in four identifiable cases (Appendix C.1).

Relation to the self-fulfilling prophecy setting. In the setting of van Amsterdam et al. (2025), the historical policy is constant, so $\mathrm{Cov}(U, t(S)) = 0$ and retrospective AUROC is undisplaced. Their post-deployment policy, a deterministic threshold on a binary feature, is the extremal case of Lemma 2. Removing the binary feature, the deterministic threshold and the constant historical policy lets the result describe two sites, rather than one site before and after deployment.

4 Experiments

We report two studies. The first audits the feature set of a published ICU benchmark. The second measures policy sensitivity across real variation in vasopressor practice.

	full	masked	Δ AUROC (pp), 95% CI
GARLIC	0.9161	0.9046	$+1.143$ $[+0.413, +1.872]$
Warpformer	0.8995	0.8893	$+1.020$ $[+0.525, +1.515]$
pooled, naive ($n = 10$)			$+1.081$ $[+0.739, +1.423]$
pooled, seed-clustered ($n = 5$)			$+1.081$ $[+0.694, +1.472]$
pooled, combined			$+1.081$ $[+0.526, +1.636]$

Table 1: Masking a single clinician-set channel. Deltas are computed from per-run predictions. The three pooled estimators agree on the point estimate to three decimals and differ only in the width of the interval. Naive pooling treats the ten runs as independent, seed clustering resamples the five seeds the two models share, and the combined row adds test-set resampling to the clustered initialisation variance. The combined row is the one we report throughout, so every interval we quote is seed-clustered. A homogeneity test between models justifies pooling (Appendix I). AUPRC moves with AUROC for both models ($+0.033$, $p = 0.022$ and $+0.026$, $p = 0.003$).

4.1 Treatment dependence in a published benchmark

Feature audit. GARLIC (Wang et al., 2026) inherits its preprocessing from Warpformer (Zhang et al., 2023). That pipeline reduces 122 raw channels to 103 and labels 42 of them "interventions" and 61 "biomarkers". Thirteen of the 103 channels are misclassified. Seven labelled interventions are not clinician-set quantities (carrier volumes, nursing flushes and oral intake), and six labelled biomarkers are set by clinicians or conditioned on treatment, among them $\mathrm{FiO_2}$ and intubation status. GARLIC's default configuration also drops vasopressin, epinephrine and phenylephrine, so any ablation of the remaining channels understates treatment dependence. We release the full classification (Appendix H).

Ablation protocol. We run the ablation on GARLIC's P19 cohort, whose 34 channels include $\mathrm{FiO_2}$. We mask $\mathrm{FiO_2}$ by zeroing its value and observation-mask columns, which leaves all parameter shapes intact, and retrain each model at its published configuration with five paired seeds. We mask rather than remove the channel because removal changes the input width and breaks seed pairing (Appendix I). On a split verified to be identical to the published one, GARLIC reaches an AUROC of 0.9161 against a published 0.9096 ± 0.0084, 0.77 standard deviations high, and Warpformer 0.8995 against a published 0.8995 ± 0.0057, within 0.01 standard deviations.

Uncertainty. Random initialisation is a known source of instability in clinical deep learning (Lopez-Martinez et al., 2022), so the combined interval adds initialisation and test-set variance, with 60.5 effective degrees of freedom (Bouthillier et al., 2021). Appendix I gives the decomposition.

The effect is comparable to the published benchmark margins. On P19, GARLIC's published lead is 0.95 pp over the next model, L-ODE-ODE, and 1.01 pp over Warpformer, our second model (Wang et al., 2026). Every interval in Table 1 contains both, while the combined interval lies below our measured margin of 1.66 pp between the two models. We therefore claim that a single clinician-set channel, one of the 34 in this cohort, accounts for an AUROC loss comparable to the published margins, not one that exceeds them.

Placebo null. The ablation shows that the model uses the channel. It does not show that the channel encodes care rather than severity. Of the 34 channels in this cohort, only six are cleanly non-manipulable. These are assays that a clinician orders but whose value a clinician cannot set. $\mathrm{FiO_2}$, by contrast, is set directly. We masked each of the seven channels identically, in GARLIC only, with three seeds per arm, and compare arms on the three seeds they share (Figure 2, Appendix I). Masking $\mathrm{FiO_2}$ lowers AUROC by 1.44 pp ($[0.31, 2.57]$, $p = 0.032$), at $1.46\times$ the most extreme placebo. It is the only arm whose interval excludes zero.

4.2 Policy sensitivity across real practice variation

Practice variation. We measure vasopressor practice in the eICU Collaborative Research Database (Pollard et al., 2018). The cohort is adult first ICU stays with an APACHE physiology

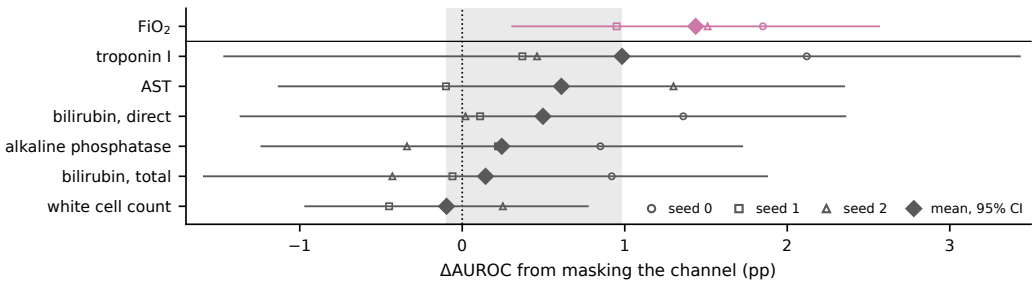


Figure 2: AUROC loss from masking each channel in GARLIC, on the three seeds shared across arms. Small markers are single seeds, and large markers are means with 95% intervals. The grey band spans the six placebo means. $\mathrm{FiO_2}$ is positive on every seed and its mean lies outside the band. Per-seed variation exceeds the between-feature effect, which is why the arms are compared on matched seeds. The three-seed mean differs from the five-seed estimate in Table 1.

record. From it we keep the 117 hospitals that record at least one dose-convertible vasopressor infusion rate and contribute at least 40 of these stays, which leaves 99,629 stays in all. We define the *charting quality* of a hospital as the fraction of its vasopressor rows with a dose-convertible infusion rate. Across all 117 hospitals, charting quality correlates with the apparent vasopressor rate (Spearman $+0.419$, $p = 2.6 \times 10^{-6}$). Among the 68 hospitals whose charting quality is at least 0.20 it does not (-0.040, $p = 0.75$), and those 68 are the ones we retain. Earlier work in this database likewise retains only hospitals that record infusions (Bosch et al., 2021). A random-intercept logistic model on these 68 hospitals, adjusted for APACHE predicted mortality, the acute physiology score, age and sex, gives a median odds ratio (Larsen & Merlo, 2005; Merlo et al., 2006) of 3.21 (profile 95% CI $[2.63, 4.17]$), which case-mix adjustment raises rather than lowers (Appendix J). Two published figures bracket this estimate. In the same database, the hospital MOR for choosing a second vasopressor over corticosteroids is 3.28 $[2.81, 3.83]$ (Bosch et al., 2021), close to ours. For corticosteroid prescribing the ICU-level MOR is 1.48 $[1.36, 1.61]$ (Soni et al., 2025), well below it. Vasopressor dosing also varies between ICUs after adjustment for severity (Hill et al., 2026).

Design. We rank hospitals by their severity-adjusted vasopressor propensity, defined as observed receipt divided by expected receipt, where the expectation comes from a model with the acute physiology score, age, sex and unit type. Splitting at the median gives the HIGH and LOW practice strata, that is, the high-use and low-use hospitals. The transfer experiment moves both filters, in opposite directions. It raises the stay threshold to 50, leaving 116 hospitals and 99,580 stays (mortality 0.091), and it drops the charting-quality threshold altogether. Because that second change readmits the low-charting hospitals, we also repeat the experiment on the 67 hospitals that pass the quality filter. Each stratum has 58 hospitals, with vasopressor rates of 0.191 and 0.041, a 4.7-fold contrast at matched severity.

Practice partly follows geography. Census region is associated with practice. Western hospitals have a median vasopressor rate of 0.024, against 0.113 to 0.157 elsewhere, and regional differences in vasopressor choice have been reported before (Fawzy et al., 2015). The charting filter retains only 2 of the 27 Western hospitals. Charting and prescribing are therefore confounded in that region, and the median odds ratio above is computed on a cohort from which the West is nearly absent. In the 67-hospital repeat, eleven hospitals change group, all from HIGH to LOW and none of them Western, so the repeat differs from the primary analysis by attrition rather than by reassignment (Appendix K).

PHYS denotes the 19 acute-physiology variables, and PHYS+VASO adds the single feature whose policy differs between strata. We train on one stratum and evaluate within that stratum and on the other, in both directions. We report two quantities. The *validation gain* is the AUROC gain from adding the feature, measured within the training practice. The *transfer cost* is the difference-in-differences (DiD) against PHYS on the other stratum.

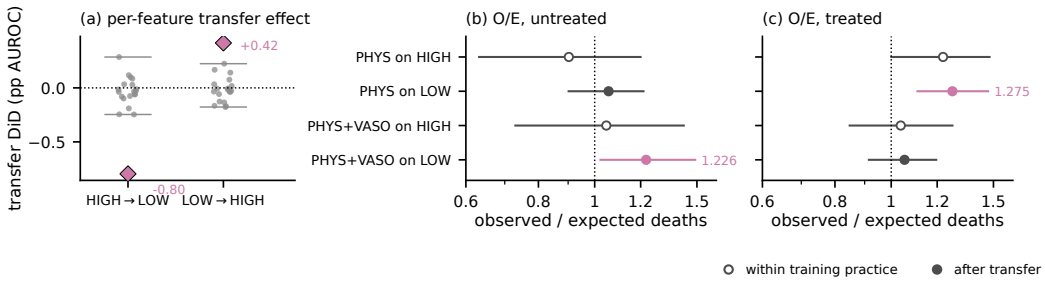


Figure 3: (a) Transfer DiD against PHYS for adding each feature, in percentage points of AUROC, from HIGH to LOW practice and back. The vasopressor feature (diamonds) lies outside the range of all 19 physiological features (grey) in both directions. Split-level point estimates. (b, c) Observed to expected ratio for the HIGH-trained models among untreated (b) and treated (c) patients, within the training practice (open) and after transfer to LOW practice (filled), with hospital-clustered 95% intervals. Values above 1 mean risk is under-predicted. Only the two highlighted cells exclude 1.

Uncertainty. The main estimate, the 67-hospital repeat, the stratified calibration and the weighting analysis use a hospital-clustered bootstrap with 200 replicates. The remaining analyses use ten splits of one fixed hospital pool, whose intervals are too narrow, so we report their point estimates and withhold any split-level interval that excludes zero.

The transfer cost cancels the validation gain. Within HIGH practice, adding the vasopressor feature raises validation AUROC by $+0.0084$ $[+0.0020, +0.0158]$. On transfer to LOW practice it costs -0.0082 $[-0.0159, -0.0009]$, with $\mathbb{P}(\text{DiD} < 0) = 98.5\%$. The cost is therefore $0.98\times$ the gain $[0.33\times, 1.60\times]$, and $\mathbb{P}(\text{ratio} > 1) = 44\%$. Hence, we claim that the cost cancels the gain.

The effect persists without the low-charting hospitals. On the 67 quality-filtered hospitals the practice contrast narrows to 2.4-fold. The transfer cost nonetheless grows, to -0.0122 $[-0.0274, +0.0010]$, against a validation gain of $+0.0136$ $[+0.0014, +0.0257]$, with $\mathbb{P}(\text{DiD} < 0) = 96\%$. This is consistent with the charting artefact diluting the effect rather than producing it. We keep the full cohort as primary for power and treat the repeat as a sensitivity analysis that does not itself reach conventional significance.

The effect depends on direction. Transferring from HIGH to LOW practice, adding the feature costs -0.0076 relative to PHYS. Transferring from LOW to HIGH, it gains $+0.0042$. Averaged over both directions the effect is -0.0017 ($p = 0.24$), because two effects of opposite sign cancel. An evaluation that pools directions would report no effect.

Placebo null over 19 physiological features. To test whether any feature degrades on transfer, or whether residual case mix differs between strata, we re-ran the DiD for each of the 19 physiological features, adding it back to PHYS without that feature. The vasopressor feature lies below every placebo from HIGH to LOW ($z = -6.06$) and above every placebo from LOW to HIGH (Figure 3a). With 19 placebos the rank-based floor is $p = 1/20$. The placebos differ between strata through case mix and show no effect, which argues against residual case-mix confounding. The c-statistic can vary by 0.1 or more across settings (Van Calster et al., 2026), but such heterogeneity moves all twenty features together and cancels within a matched split.

Transfer under-predicts risk among the untreated. Calibration-in-the-large shows no effect (-0.0040 $[-0.0143, +0.0062]$, $p = 0.42$), because the difference in base rate between strata dominates it. We therefore report it separately by treatment stratum (Figure 3b,c). Only two cells exclude 1, both after transfer to LOW practice. The model without the treatment feature under-predicts risk among treated patients (1.275 $[1.110, 1.468]$). Among untreated patients, observed deaths exceed the predictions of the model with the feature by 23% (1.226 $[1.023, 1.490]$). We do not claim that adding the feature causes this shift: the corresponding contrasts between models and between strata contain zero (Appendix L), and we do not infer a difference from one interval excluding 1 while

another does not. Within the training practice the model with the feature appears better calibrated than the model without it, so a same-site calibration check favours the feature that fails on transfer.

Robustness. The sign pattern, the placebo separation and the sign of the calibration contrast replicate in a multilayer perceptron and a logistic model (Appendix M). Adding all six treatment features gives a smaller transfer cost than the vasopressor feature alone, -0.0050 against -0.0076, because the other treatment rates barely differ between strata.

Inverse-probability weighting does not repair the cost. We fit stabilised inverse-probability weights for vasopressor receipt on the training hospitals only, truncated at the 1st and 99th percentiles. The weights are well behaved (Appendix N). We re-ran the comparison in a separately seeded hospital-clustered bootstrap, whose unweighted arm reproduces the headline cost (-0.00799 against -0.0082). The weighted cost is also -0.00799, so the repair is -0.0000 $[-0.0053, +0.0054]$, with $\mathbb{P}(\text{repair} > 0) = 55\%$. Weighting retains 92% of the validation gain ($+0.00732$ against $+0.00797$) and all of the transfer cost. The weighted gain, $[-0.0004, +0.0155]$, is not itself distinguishable from zero, and we infer no difference from one gain interval excluding zero and the other not. Under split-level resampling weighting appeared to buy a small repair, which disappears once hospitals are the resampling unit.

5 Discussion

The vasopressor feature passes every same-site diagnostic. It raises AUROC, it appears well calibrated, and it holds up on held-out hospitals drawn from the same practice. However, it fails when the model moves to hospitals that use vasopressors less. This results in three practical consequences. First, a retrospective AUROC is not a property of a score alone. It describes a score together with the treatment policy under which it was recorded, so a comparison across sites is a comparison of scores only when practice is matched or the displacement is corrected for. Our closed form gives the size of the term that is otherwise ignored. Second, calibration should be reported separately for treated and untreated patients. Third, a feature audit should test the feature whose policy differs between sites, since testing all six together weakened the cost we detected.

We also find that reweighting is not the remedy it appears to be. Inverse-probability weighting gave up part of the validation gain and recovered none of the transfer cost. To first order, a method that only reduces how much the model relies on the feature decreases the validation gain and the transfer cost in proportion, so their ratio is unchanged. Our result is consistent with that prediction, though it attenuates the feature too little to test it. The implication is that a remedy has to change what the feature *means* across sites, not how much of it the model is allowed to use.

The work in this paper is not specific to clinical practice. Proposition 1 applies to any score that triggers an outcome-changing action, and fraud review and early-alert systems in education have the same structure. Wherever a score triggers an effective response, the retrospective metric that evaluates it is displaced by an amount set by the response policy, so a comparison between settings with different policies is not, on its own, a comparison of the scores.

Our work has certain limitations and they bound each claim differently. The benchmark placebo comparison rests on three shared seeds, so we do not claim that treatment features differ categorically from physiological ones. The policy-sensitivity experiment uses one database, one exposure and one outcome, and geography and practice are not fully separable within it, because Western hospitals prescribe far less and are nearly absent from the quality-filtered cohort. The calibration findings rest on single cells, and the contrasts between them are not significant at this number of hospitals. Proposition 2 is proved for a threshold policy with normally distributed score and severity, and we have not shown it more generally. Finally, our reanalysis of the early warning paradox corrects only one illustration of treatment-induced label bias without addressing the wider evidence for the phenomenon.

Use of AI tools

LLMs were used for drafting and revising prose from author-supplied results, assembling the LaTeX source, and auditing the manuscript for internal inconsistency between sections.

REPRODUCIBILITY STATEMENT

All analyses are reproducible from the supplementary material, which contains code, the benchmark classification and aggregate results but no patient-level data. MIMIC-III, MIMIC-IV and the eICU Collaborative Research Database are available to credentialed users through PhysioNet, and P19 is derived from the openly available PhysioNet/CinC Challenge 2019 data. The manipulability classification of the benchmark is released keyed to tensor position rather than to feature index, together with the reconstruction of the benchmark index map. A single script, which needs no data, reproduces every theoretical constant, including the Monte Carlo checks in Appendix A, the uniform bound of Proposition 2, the floor inversion of Corollary 3 and the scope cases of Appendix C.1. The eICU cohort is built by an explicit filter chain that logs hospital and stay counts at each step, and the charting-quality metric is given as a query over raw rows. Per-run summaries for every ablation arm are included.

ETHICS STATEMENT

This work is a retrospective secondary analysis of three de-identified critical-care databases, MIMIC-III, MIMIC-IV and the eICU Collaborative Research Database, all distributed under PhysioNet credentialed access, and of the openly available P19 benchmark cohort. The authors completed the required human-subjects research training and are covered by signed data use agreements. No patient was contacted, no intervention was assigned, and no attempt was made to re-identify any individual.

Two aspects of the work warrant comment. First, the manipulability classification of benchmark features is a clinical judgement rather than a measurement. It was made by the authors, one of whom is a practising physician, and the full classification is released.

Second, our findings concern how deterioration models are *evaluated*, not how patients should be treated. We show that a feature reflecting treatment practice can improve retrospective discrimination while degrading under a change of practice, and that the resulting error, after transfer, is that the model with the feature under-predicts risk among untreated patients. We draw no conclusion about the effectiveness of any treatment, and nothing here should be read as a recommendation to alter vasopressor or transfusion practice.

REFERENCES

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

Daniel R. Balcarcel, Sanjiv D. Mehta, Celeste G. Dixon, Charlotte Z. Woods-Hill, Ewan C. Goligher, Wouter A. C. van Amsterdam, and Nadir Yehya. Feedback loops in intensive care unit prognostic models: an under-recognised threat to clinical validity. *The Lancet Digital Health*, 7:100880, 2025. doi: 10.1016/j.landig.2025.100880.

Donald Bamber. The area above the ordinal dominance graph and the area below the receiver operating characteristic graph. *Journal of Mathematical Psychology*, 12(4):387–415, 1975. doi: 10.1016/0022-2496(75)90001-2.

Brett K. Beaulieu-Jones, William Yuan, Gabriel A. Brat, Andrew L. Beam, Griffin M. Weber, Marshall Ruffin, and Isaac S. Kohane. Machine learning for patient risk stratification: standing on, or looking over, the shoulders of clinicians? *npj Digital Medicine*, 4(1):62, 2021. doi: 10.1038/s41746-021-00426-3.

Philip Boeken, Onno Zoeter, and Joris M. Mooij. Evaluating and correcting performative effects of decision support systems via causal domain shift. In *Proceedings of the Third Conference on Causal Learning and Reasoning*, volume 236 of *Proceedings of Machine Learning Research*, pp. 551–569, 2024.

Nicholas A. Bosch, Bijan Teja, Hannah Wunsch, and Allan J. Walkey. Practice patterns in the initiation of secondary vasopressors and adjunctive corticosteroids during septic shock in

the United States. *Annals of the American Thoracic Society*, 18(12):2049–2057, 2021. doi: 10.1513/AnnalsATS.202102-196OC.

Xavier Bouthillier, Pierre Delaunay, Mirko Bronzi, Assya Trofimov, Brennan Nichyporuk, Justin Szeto, Nazanin Mohammadi Sepahvand, Edward Raff, Kanika Madan, Vikram Voleti, Samira Ebrahimi Kahou, Vincent Michalski, Tal Arbel, Chris Pal, Gaël Varoquaux, and Pascal Vincent. Accounting for variance in machine learning benchmarks. In *Proceedings of Machine Learning and Systems*, volume 3, pp. 747–769, 2021.

Kurt Butler, Stephanie Riley, Damian Machlanski, Edward Moroshko, Panagiotis Dimitrakopoulos, Thomas Melistas, Akchunya Chanchal, Konstantinos Vilouras, Zhihua Liu, Steven McDonagh, Hana Chockler, Ben Glocker, Niccolò Tempini, Matthew Sperrin, Sotirios A. Tsaftaris, and Ricardo Silva. A causal framework for mitigating data shifts in healthcare, 2026. arXiv:2603.13595.

Rich Caruana, Yin Lou, Johannes Gehrke, Paul Koch, Marc Sturm, and Noemie Elhadad. Intelligible models for HealthCare: Predicting pneumonia risk and hospital 30-day readmission. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 1721–1730, 2015. doi: 10.1145/2783258.2788613.

Gregory F. Cooper, Constantin F. Aliferis, Richard Ambrosino, John Aronis, Bruce G. Buchanan, Richard Caruana, Michael J. Fine, Clark Glymour, Geoffrey Gordon, Barbara H. Hanusa, Janine E. Janosky, et al. An evaluation of machine-learning methods for predicting pneumonia mortality. *Artificial Intelligence in Medicine*, 9(2):107–138, 1997.

Amanda Coston, Alan Mishler, Edward H. Kennedy, and Alexandra Chouldechova. Counterfactual risk assessments, evaluation, and fairness. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, pp. 582–593, 2020. doi: 10.1145/3351095.3372851.

Edward E. Cureton. Rank-biserial correlation. *Psychometrika*, 21(3):287–290, 1956. doi: 10.1007/BF02289138.

Minjie Duan, Xiaoli Liu, Wesley Yeung, Sicheng Hao, Jesse Raffa, Pan Hu, Chao Liu, Lin Chen, Zitao Li, Zhenyue Gao, Tao Li, Desen Cao, Feihu Zhou, Zhongheng Zhang, Zhengbo Zhang, and Leo Anthony Celi. Multicenter validation and updating of the ELDER-ICU model for severity assessment in elderly critical illness. *npj Digital Medicine*, 2026. doi: 10.1038/s41746-026-02472-1. In press.

Kirill Dyagilev and Suchi Saria. Learning (predictive) risk scores in the presence of censoring due to interventions. *Machine Learning*, 102(3):323–348, 2016. doi: 10.1007/s10994-015-5527-7.

Gabriel J. Escobar, Vincent X. Liu, Alejandro Schuler, Brian Lawson, John D. Greene, and Patricia Kipnis. Automated identification of adults at risk for in-hospital clinical deterioration. *New England Journal of Medicine*, 383(20):1951–1960, 2020. doi: 10.1056/NEJMsa2001090.

Ashraf Fawzy, Stephen R. Evans, and Allan J. Walkey. Practice patterns and outcomes associated with choice of initial vasopressor therapy for septic shock. *Critical Care Medicine*, 43(10):2141–2146, 2015. doi: 10.1097/CCM.0000000000001149.

Samuel G. Finlayson, Adarsh Subbaswamy, Karandeep Singh, John Bowers, Annabel Kupke, Jonathan Zittrain, Isaac S. Kohane, and Suchi Saria. The clinician and dataset shift in artificial intelligence. *New England Journal of Medicine*, 385(3):283–286, 2021. doi: 10.1056/NEJMc2104626.

Gerardo A. Flores, Alyssa H. Smith, Julia Fukuyama, and Ashia Wilson. Aligning evaluation with clinical priorities: Calibration, label shift, and error costs. In *Advances in Neural Information Processing Systems*, 2025. arXiv:2506.14540.

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

Joseph Futoma, Morgan Simons, Finale Doshi-Velez, and Rishikesan Kamaleswaran. Generalization in clinical prediction models: the blessing and curse of measurement indicator variables. *Critical Care Explorations*, 3(7):e0453, 2021.

Robert Geirhos, Jörn-Henrik Jacobsen, Claudio Michaelis, Richard Zemel, Wieland Brendel, Matthias Bethge, and Felix A. Wichmann. Shortcut learning in deep neural networks. *Nature Machine Intelligence*, 2(11):665–673, 2020. doi: 10.1038/s42256-020-00257-z.

Paul C. Hébert, George Wells, Morris A. Blajchman, John Marshall, Claudio Martin, et al. A multicenter, randomized, controlled clinical trial of transfusion requirements in critical care. *New England Journal of Medicine*, 340(6):409–417, 1999. doi: 10.1056/NEJM199902113400601.

Katharine E. Henry, David N. Hager, Peter J. Pronovost, and Suchi Saria. A targeted real-time early warning score (TREWScore) for septic shock. *Science Translational Medicine*, 7(299):299ra122, 2015. doi: 10.1126/scitranslmed.aab3719.

Andrea D. Hill, Bijan Teja, Lisa Burry, Ruxandra Pinto, Nicholas A. Bosch, Allan Walkey, Henry T. Stelfox, Martin Chapman, Damon C. Scales, and Hannah Wunsch. Variation in maximal doses of vasopressors used for treatment of shock in intensive care units in Alberta Canada: A retrospective cohort study. *Anesthesia & Analgesia*, 142(6):1088–1097, 2026. doi: 10.1213/ANE.0000000000007821.

Lars B. Holst, Nicolai Haase, Jørn Wetterslev, Jan Wernerman, Anne B. Guttormsen, Sari Karlsson, Pär I. Johansson, Anders Åneman, Marianne L. Vang, Robert Winding, et al. Lower versus higher hemoglobin threshold for transfusion in septic shock. *New England Journal of Medicine*, 371(15): 1381–1391, 2014. doi: 10.1056/NEJMoa1406617.

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

Zongliang Ji, Yifei Sun, Andre Amaral, Anna Goldenberg, and Rahul G. Krishnan. Can we generate portable representations for clinical time series data using LLMs? In *International Conference on Learning Representations*, 2026. arXiv:2603.23987.

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

Klaus Larsen and Juan Merlo. Appropriate assessment of neighborhood effects on individual health: integrating random and fixed effects in multilevel logistic regression. *American Journal of Epidemiology*, 161(1):81–88, 2005. doi: 10.1093/aje/kwi017.

Gyeongsil Lee, Seulggie Choi, Kyuwoong Kim, Jae-Moon Yun, Joung Sik Son, Su-Min Jeong, Sung Min Kim, and Sang Min Park. Association of hemoglobin concentration and its change with cardiovascular and all-cause mortality. *Journal of the American Heart Association*, 7(3): e007723, 2018. doi: 10.1161/JAHA.117.007723.

Matthew C. Lenert, Michael E. Matheny, and Colin G. Walsh. Prognostic models will be victims of their own success, unless.... *Journal of the American Medical Informatics Association*, 26(12): 1645–1650, 2019. doi: 10.1093/jamia/ocz145.

James Liley, Samuel Emerson, Bilal Mateen, Catalina Vallejos, Louis Aslett, and Sebastian Vollmer. Model updating after interventions paradoxically introduces bias. In *Proceedings of the 24th International Conference on Artificial Intelligence and Statistics*, volume 130 of *Proceedings of Machine Learning Research*, pp. 3916–3924, 2021.

Hugh Logan Ellis, Edward Palmer, James T. Teo, Martin Whyte, Kenneth Rockwood, and Zina Ibrahim. The early warning paradox. *npj Digital Medicine*, 8(1):81, 2025. doi: 10.1038/s41746-024-01408-x.

Daniel Lopez-Martinez, Alex Yakubovich, Martin Seneviratne, Adam D. Lelkes, Akshit Tyagi, Jonas Kemp, Ethan Steinberg, N. Lance Downing, Ron C. Li, Keith E. Morse, Nigam H. Shah,

and Ming-Jun Chen. Instability in clinical risk stratification models using deep learning. In *Proceedings of the 2nd Machine Learning for Health Symposium*, volume 193 of *Proceedings of Machine Learning Research*, pp. 552–565, 2022.

Xiaolei Lu and Shamim Nemati. Adaptive test-time training for predicting need for invasive mechanical ventilation in multi-center cohorts, 2025. arXiv:2512.06652.

Kim Luijken, Paweł Morzywołek, Wouter van Amsterdam, Giovanni Cinà, Jeroen Hoogland, Ruth Keogh, Jesse H. Krijthe, Sara Magliacane, Thijs van Ommen, Niels Peek, Hein Putter, et al. Risk-based decision making: Estimands for sequential prediction under interventions. *Biometrical Journal*, 66(8):e70011, 2024.

Melissa D. McCradden, Mjaye L. Mazwi, and Lauren Oakden-Rayner. Can an accurate model be bad? *Patterns*, 6:101205, 2025. doi: 10.1016/j.patter.2025.101205.

Juan Merlo, Basile Chaix, Henrik Ohlsson, Anders Beckman, Kristina Johnell, Per Hjerpe, Lennart Råstam, and Klaus Larsen. A brief conceptual tutorial of multilevel analysis in social epidemiology: using measures of clustering in multilevel logistic regression to investigate contextual phenomena. *Journal of Epidemiology and Community Health*, 60(4):290–297, 2006. doi: 10.1136/jech.2004.029454.

Michael Moor, Nicolas Bennett, Drago Plečko, Max Horn, Bastian Rieck, Nicolai Meinshausen, Peter Bühlmann, and Karsten Borgwardt. Predicting sepsis using deep learning across international sites: a retrospective development and validation study. *eClinicalMedicine*, 62:102124, 2023. doi: 10.1016/j.eclinm.2023.102124.

Bret Nestor, Matthew B. A. McDermott, Willie Boag, Gabriela Berner, Tristan Naumann, Michael C. Hughes, Anna Goldenberg, and Marzyeh Ghassemi. Feature robustness in non-stationary health records: Caveats to deployable model performance in common clinical machine learning tasks. In *Proceedings of the 4th Machine Learning for Healthcare Conference*, volume 106 of *Proceedings of Machine Learning Research*, pp. 381–405, 2019.

Roger Newson. Parameters behind "nonparametric" statistics: Kendall's tau, Somers' D and median differences. *The Stata Journal*, 2(1):45–64, 2002. doi: 10.1177/1536867X0200200103.

Chris Paxton, Alexandru Niculescu-Mizil, and Suchi Saria. Developing predictive models using electronic medical records: challenges and pitfalls. *AMIA Annual Symposium Proceedings*, 2013: 1109–1115, 2013.

Juan Perdomo, Tijana Zrnic, Celestine Mendler-Dünner, and Moritz Hardt. Performative prediction. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 7599–7609, 2020.

Stephen R. Pfohl, Natalie Harris, Chirag Nagpal, David Madras, Vishwali Mhasawade, Olawale Salaudeen, Awa Dieng, Shannon Sequeira, Santiago Arciniegas, Lillian Sung, Nnamdi Ezeanochie, Heather Cole-Lewis, Katherine Heller, Sanmi Koyejo, and Alexander D'Amour. Understanding challenges to the interpretation of disaggregated evaluations of algorithmic fairness. In *Advances in Neural Information Processing Systems*, 2025. arXiv:2506.04193.

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

Ilaria Prosepe, Rolf H. H. Groenwold, Rachel Knevel, Romin Pajouheshnia, and Nan van Geloven. The disconnect between development and intended use of clinical prediction models for Covid-19: a systematic review and real-world data illustration. *Frontiers in Epidemiology*, 2:899589, 2022.

Patrick Rockenschaub, Adam Hilbert, Tabea Kossen, Paul Elbers, Falk von Dincklage, Vince Istvan Madai, and Dietmar Frey. The impact of multi-institution datasets on the generalizability of machine learning prediction models in the ICU. *Critical Care Medicine*, 52(11):1710–1721, 2024. doi: 10.1097/CCM.0000000000006359.

Peter Schulam and Suchi Saria. Reliable decision support using counterfactual models. In *Advances in Neural Information Processing Systems 30*, pp. 1697–1708, 2017.

Harvineet Singh, Vishwali Mhasawade, and Rumi Chunara. Generalizability challenges of mortality risk prediction models: A retrospective analysis on a multi-center database. *PLOS Digital Health*, 1(4):e0000023, 2022. doi: 10.1371/journal.pdig.0000023.

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

Kanupriya Soni, John S. Minturn, Billie S. Davis, Leigh A. Bukowski, Jeremy M. Kahn, and Ian J. Barbash. Variation in corticosteroid prescribing practices for patients with septic shock. *Critical Care Explorations*, 7(3):e1196, 2025. doi: 10.1097/CCE.0000000000001196.

Dimitris Spathis and Stephanie L. Hyland. Looking for out-of-distribution environments in multi-center critical care data, 2022. arXiv:2205.13398.

Matthew Sperrin, Glen P. Martin, Alexander Pate, Tjeerd Van Staa, Niels Peek, and Iain Buchan. Using marginal structural models to adjust for treatment drop-in when developing clinical prediction models. *Statistics in Medicine*, 37(28):4142–4154, 2018. doi: 10.1002/sim.7913.

Matthew Sperrin, David Jenkins, Glen P. Martin, and Niels Peek. Explicit causal reasoning is needed to prevent prognostic models being victims of their own success. *Journal of the American Medical Informatics Association*, 26(12):1675–1676, 2019. doi: 10.1093/jamia/ocz197.

Adarsh Subbaswamy and Suchi Saria. I-SPEC: An end-to-end framework for learning transportable, shift-stable models, 2020. arXiv:2002.08948.

Nenad Tomašev, Xavier Glorot, Jack W. Rae, Michal Zielinski, Harry Askham, Andre Saraiva, Anne Mottram, Clemens Meyer, Suman Ravuri, Ivan Protsyuk, et al. A clinically applicable approach to continuous prediction of future acute kidney injury. *Nature*, 572(7767):116–119, 2019. doi: 10.1038/s41586-019-1390-1.

Wouter A. C. van Amsterdam, Nan van Geloven, Jesse H. Krijthe, Rajesh Ranganath, and Giovanni Cinà. When accurate prediction models yield harmful self-fulfilling prophecies. *Patterns*, 6(4): 101229, 2025. doi: 10.1016/j.patter.2025.101229.

Ben Van Calster, Maarten van Smeden, Wouter van Amsterdam, Maarten Coemans, Laure Wynants, and Ewout W. Steyerberg. The enemies of reliable and useful clinical prediction models: A review of statistical and scientific challenges. *Annual Review of Statistics and Its Application*, 13 (1):465–492, 2026. doi: 10.1146/annurev-statistics-042324-123749.

Robin van de Water, Hendrik Schmidt, Paul Elbers, Patrick Thoral, Bert Arnrich, and Patrick Rockenschaub. Yet another ICU benchmark: A flexible multi-center framework for clinical ML. In *International Conference on Learning Representations*, 2024.

Nan van Geloven, Sonja A. Swanson, Chava L. Ramspek, Kim Luijken, Merel van Diepen, Tim P. Morris, Rolf H. H. Groenwold, Hans C. van Houwelingen, Hein Putter, and Saskia le Cessie. Prediction meets causal inference: the role of treatment in clinical prediction models. *European Journal of Epidemiology*, 35(7):619–630, 2020. doi: 10.1007/s10654-020-00636-1.

Nan van Geloven, Ruth H. Keogh, Wouter van Amsterdam, Giovanni Cinà, Jesse H. Krijthe, Niels Peek, Kim Luijken, Sara Magliacane, Paweł Morzywołek, Thijs van Ommen, Hein Putter, Matthew Sperrin, Junfeng Wang, Daniala L. Weir, and Vanessa Didelez. The risks of risk assessment: causal blind spots when using prediction models for treatment decisions. *Annals of Internal Medicine*, 178:1326–1333, 2025.

Ruirui Wang, Yanke Li, Manuel Günther, and Diego Paez-Granados. GARLIC: Graph attention-based relational learning of multivariate time series in intensive care. In *International Conference on Learning Representations*, 2026. arXiv:2608.10969. Wang and Li contributed equally.

Dominic Wilkinson. The self-fulfilling prophecy in intensive care. *Theoretical Medicine and Bioethics*, 30(6):401–410, 2009.

Andrew Wong, Erkin Otles, John P. Donnelly, Andrew Krumm, Jeffrey McCullough, Olivia DeTroyer-Cooley, Justin Pestrue, Marie Phillips, Judy Konye, Carleen Penoza, Muhammad Ghous, and Karandeep Singh. External validation of a widely implemented proprietary sepsis prediction model in hospitalized patients. *JAMA Internal Medicine*, 181(8):1065–1070, 2021. doi: 10.1001/jamainternmed.2021.2626.

R. Yamamoto, F. Wu, L. K. Sprehe, A. Abeer, Leo Anthony Celi, and T. Tohyama. Observation-process features are associated with larger domain shift in sepsis mortality prediction: a cross-database evaluation using MIMIC-IV and eICU-CRD. *medRxiv*, 2026. doi: 10.64898/2026.04.05.26350209. Preprint.

Haoran Zhang, Natalie Dullerud, Laleh Seyyed-Kalantari, Quaid Morris, Shalmali Joshi, and Marzyeh Ghassemi. An empirical framework for domain generalization in clinical settings. In *Proceedings of the Conference on Health, Inference, and Learning (ACM CHIL)*, pp. 279–290, 2021. doi: 10.1145/3450439.3451878.

Jiawen Zhang, Shun Zheng, Wei Cao, Jiang Bian, and Jia Li. Warpformer: A multi-scale modeling approach for irregular clinical time series. In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 2023. doi: 10.1145/3580305.3599543.

Kevin Zhang, Yonghan Jung, Divyat Mahajan, Karthikeyan Shanmugam, and Shalmali Joshi. Path-specific effects for pulse-oximetry guided decisions in critical care. In *Advances in Neural Information Processing Systems*, 2025. arXiv:2506.12371.

Jing Zou, Haozengran Wang, Zile Xin, Qian Zhou, Jingxi Chen, and Junlei Huang. From predictive signal to deployable value: EHR measurement-process features in ICU AKI prediction. *International Journal of Medical Informatics*, 221:106668, 2026. doi: 10.1016/j.ijmedinf.2026.106668.

A THE IDENTITY AND PROPOSITION 1

Every displayed constant below is reproduced by `analysis/props.py`. The threshold parameter is written τ_0 throughout, reserving $\tau(\cdot)$ for the severity factor of a separable policy. Lowercase s denotes the score S of the main text.

A.1 SETUP AND NOTATION

Let $z \sim \mathcal{N}(0,1)$ be latent severity and $p_0 : \mathbb{R} \to [0,1]$ the untreated risk, nondecreasing, with $\mu := \mathbb{E}[p_0(z)] \in (0,1)$. A score of quality $q = \operatorname{corr}(s,z) \in [0,1]$ is $s = qz + \sqrt{1-q^2}\,\varepsilon$ with $\varepsilon \sim \mathcal{N}(0,1) \perp\!\!\!\perp z$, so $s \sim \mathcal{N}(0,1)$ for every q. In particular, the treated fraction at a fixed boundary does not depend on q. Treatment is Bernoulli with $t(s,z) := \Pr(T = 1 \mid s, z)$ and, conditionally on (s,z), independent of everything else. It removes a fraction $e(z) \in [-1,1]$ of risk, so that $Y \mid z, s, T \sim \text{Bernoulli}\big(p_0(z)(1 - e(z)T)\big)$, and a negative e is harm.

Write $\rho(s) := \Pr(Y = 1 \mid s)$, which is the $r(U)$ of Lemma 1 on the score scale. Let φ and Φ be the standard normal density and CDF, and $U := \Phi(s) \sim \text{Unif}(0,1)$. The threshold policy is $t(s,z) = a\mathbf{1}[s > \tau_0]$ with constant e. For it we write $\delta := ea$, $\pi := \Pr(s > \tau_0) = 1 - \Phi(\tau_0)$, $W(s) := 1 - \delta\mathbf{1}[s > \tau_0]$ and $m_q(s) := \mathbb{E}[p_0(z) \mid s]$, so $\rho = m_q W$. The retrospective AUROC is $R := \Pr(s^+ > s^-)$ for independent draws among $Y = 1$ and $Y = 0$, with outcomes recorded under the treatment policy t. Because s is continuous, there are no ties.

A.2 THE MASTER IDENTITY

Lemma 1 is not new, and the proof below is included only to keep the appendix self-contained. Bamber (1975) established that the area equals $\Pr(S_1 > S_2)$ and is the Mann–Whitney statistic up to normalisation, which is the load-bearing step; the covariance display is the population form of Somers' D (Newson, 2002), and for a binary outcome the rank-biserial correlation (Cureton,

1956). The first form below, which substitutes $s \sim \mathcal{N}(0,1)$, is bookkeeping that makes the identity computable here rather than a separate result.

Proof of Lemma 1. Let $u := \varphi\rho$, $P := \int u = \Pr(Y = 1)$ and $I := \int u\Phi$. The unnormalised density of s among cases is u and among controls $\varphi - u$. Hence

$$R\,P(1-P) = \iint_{\{s>t\}} u(s)\big(\varphi(t) - u(t)\big)\,\mathrm{d}t\,\mathrm{d}s = \int u\Phi - \iint_{\{s>t\}} u(s)u(t)\,\mathrm{d}t\,\mathrm{d}s.$$

The first term is I. In the second the integrand is symmetric in (s,t) and the diagonal is null, so the half-plane contributes exactly half of $(\int u)^2 = P^2$, giving $R = (I - \frac{1}{2}P^2)/(P(1-P))$. For the rank form, $\int \varphi\Phi = \frac{1}{2}$, so $I - \frac{1}{2}P = \int(\Phi - \frac{1}{2})\varphi\rho = \mathrm{Cov}(U, \rho(s))$ since $\mathbb{E}[U] = \frac{1}{2}$, and therefore $R - \frac{1}{2} = (I - \frac{1}{2}P)/(P(1-P))$. □

Every theoretical AUROC in this paper is computed from (1), which reduces it to two one-dimensional integrals of ρ. The second form says that retrospective AUROC is, up to normalisation, the covariance between score rank and treated risk. It therefore depends on the treatment policy under which outcomes were recorded, not on z alone.

A.3 Proposition 1

Proof of Proposition 1. At $q = 0$, $s \perp\!\!\!\perp z$, so the conditional law of z given s is the marginal law of z. Conditioning on (s,z) and then on s,

$$\rho(s) = \mathbb{E}\big[p_0(z)(1 - e(z)T) \mid s\big] = \mathbb{E}_z[p_0(z)] - \mathbb{E}_z[p_0(z)e(z)t(s,z)] = \mu - \kappa(s),$$

using $\mathbb{E}[T \mid s, z] = t(s,z)$ together with $z \mid s \overset{d}{=} z$. No conditional independence between T and z given s is required, because (2) integrates out the severity dependence of the policy. Substituting into the covariance form of (1) and using $\mathrm{Cov}(U, \mu) = 0$ gives (3). Since $P \in (0,1)$, $R(0) < \frac{1}{2}$ if and only if $\mathrm{Cov}(U, \kappa(U)) > 0$.

For the sufficient condition, suppose $t(\cdot, z)$ is nondecreasing for each z and $p_0 e \geq 0$. Then κ is an integral of nondecreasing functions against a nonnegative measure, hence nondecreasing. Both $u \mapsto u$ and $u \mapsto \kappa(\Phi^{-1}(u))$ are nondecreasing on $(0,1)$, so Chebyshev's association inequality gives $\mathrm{Cov}(U, \kappa(U)) \geq 0$, strictly unless κ is a.e. constant. □

Remark 1. Proposition 1 requires neither a deterministic policy, nor a threshold, nor any particular treatment rate, nor a constant treatment effect, nor separability of t in (s,z). The hypotheses that do bind are collected in Appendix C.1.

Corollary 1 (Separable policies)**.** *If $t(s,z) = t(s)\tau(z)$ then $\kappa(s) = \kappa\, t(s)$ with $\kappa = \mathbb{E}[p_0(z)e(z)\tau(z)]$, and*

$$R(0) = \tfrac{1}{2} - \frac{\kappa\,\mathrm{Cov}(U, t(U))}{P(1-P)}, \qquad P = \mu - \kappa\bar{t}, \quad \bar{t} = \mathbb{E}[t(s)].$$

If $\kappa > 0$ and t is nondecreasing and non-constant then $R(0) < \frac{1}{2}$. The case $\tau \equiv 1$ is a policy that ignores severity.

Corollary 2 (Threshold policy)**.** *For $t = a\mathbf{1}[s > \tau_0]$ with constant e, writing $\delta = ea$ and $\pi = 1 - \Phi(\tau_0)$,*

$$R(0) = \tfrac{1}{2} - \frac{\delta\pi(1-\pi)}{2(1-\delta\pi)(1-P)}, \qquad P = \mu(1-\delta\pi). \tag{8}$$

Proof. Here $\kappa(s) = e\mu a\mathbf{1}[s > \tau_0]$, so $\mathbb{E}[\kappa] = \mu\delta\pi$ and $P = \mu(1-\delta\pi)$. Also $\mathrm{Cov}(U, \mathbf{1}[U > 1-\pi]) = \int_{1-\pi}^{1}(u - \frac{1}{2})\,\mathrm{d}u = \frac{1}{2}\pi(1-\pi)$, so $\mathrm{Cov}(U, \kappa(U)) = \frac{1}{2}\mu\delta\pi(1-\pi)$. Substituting into (3) gives (8). □

The first form of (1) gives the same result. With $\rho = \mu W$, $I = \mu[\frac{1}{2} - \delta\int_{\tau_0}^{\infty}\varphi\Phi] = \frac{\mu}{2}[1 - \delta\pi(2-\pi)]$, using $\int_{\tau_0}^{\infty}\varphi\Phi = \frac{1}{2}(1 - \Phi(\tau_0)^2)$ and $1 - \Phi(\tau_0)^2 = \pi(2-\pi)$, so $I - \frac{1}{2}P = -\frac{\mu}{2}\delta\pi(1-\pi)$. Table 2 checks these closed forms against simulation.

policy	rate	closed form	Monte Carlo
deterministic threshold, $e = 0.9$	0.16	0.4187	0.4187
stochastic probit propensity	0.17	0.4173	0.4173
weak propensity $0.3 + 0.1U$	0.35	0.4878	0.4882
threshold, 90% treated	0.90	0.2804	0.2796
probit propensity, effect rising with severity	0.17	0.4598	0.4597
two eras pooled (threshold + probit)	0.30	0.3739	0.3742

Table 2: Verification of Proposition 1, $n = 6 \times 10^6$. Equation (8) matches quadrature to 10^{-15} across six (δ, τ_0) settings.

Proof of Lemma 2. Maximise $\int_0^1 (u - \frac{1}{2}) t(u)\,\mathrm{d}u$ over $0 \le t \le 1$ subject to $\int_0^1 t = \bar{t}$. The objective is linear in t, so by the bathtub principle an optimum places the mass where the weight $u - \frac{1}{2}$ is largest, that is on $\{u > 1 - \bar{t}\}$, with value $\int_{1-\bar{t}}^1 (u - \frac{1}{2})\,\mathrm{d}u = \frac{1}{2}\bar{t}(1 - \bar{t})$. Uniqueness up to null sets follows because $u - \frac{1}{2}$ is strictly increasing. A deterministic threshold therefore displaces AUROC furthest below $\frac{1}{2}$. Any stochastic policy at the same rate displaces it strictly less, but still below $\frac{1}{2}$. □

Proof of the worst alert rate (7). In the rare-outcome limit $P \to 0$, minimising (8) is maximising $g(\pi) = \delta\pi(1 - \pi)/(1 - \delta\pi)$ on $(0, 1)$. Differentiating, $g'(\pi) \propto (1 - 2\pi)(1 - \delta\pi) + \delta\pi(1 - \pi) = 1 - 2\pi + \delta\pi^2$, whose roots are $\pi = (1 \pm r)/\delta$ with $r := \sqrt{1 - \delta}$. Only $\pi^\star = (1 - r)/\delta$ lies in $(0, 1]$, and $\delta = 1 - r^2$ gives $\pi^\star = 1/(1 + r)$ and $1 - \delta\pi^\star = r$. Then $\pi^\star(1 - \pi^\star) = r/(1 + r)^2$, so $g(\pi^\star) = \delta/(1 + r)^2 = (1 - r)/(1 + r)$ and $R_{\min} = \frac{1}{2} - \frac{1}{2} g(\pi^\star) = r/(1 + r)$. □

Since $R_{\min} \to 0$ as $\delta \to 1$, no positive constant bounds R from below, and a score with no predictive value can appear arbitrarily far below chance. Inverting the floor gives what any particular value requires.

Corollary 3 (Inverting the floor)**.** *In the rare-outcome limit, a retrospective AUROC of $R < \frac{1}{2}$ at zero quality is attainable only if*

$$\delta \;\ge\; 1 - \Big(\frac{R}{1 - R}\Big)^2,$$

and at the smallest such δ only at the alerted fraction $\pi^\star = 1 - R$.

Proof. Solving $R = r/(1 + r)$ for $r = \sqrt{1 - \delta}$ gives $r = R/(1 - R)$, so $\delta = 1 - (R/(1 - R))^2$, and $\pi^\star = 1/(1 + r) = 1 - R$. □

Under defensible settings, with a treatment effect of at most 0.5, adherence of at most 0.9 and at most 30% of patients alerted, $\delta \le 0.45$ and the deficit is at most 0.055 in the rare-outcome limit and 0.060 at a 9% event rate, both attained at $\delta = 0.45$ and $\pi = 0.3$. A reported sub-chance value should therefore be quoted with the (δ, π) that produces it.

B THE DISPLACEMENT DOES NOT VANISH

The proof of Proposition 2 has three steps. *Direction.* Effective treatment moves case scores down and control scores up for any score-rising propensity, so $\Delta \ge 0$ with no condition on q. *Size.* For the threshold policy, the shift in the case-score distribution has an exact product form. *Uniformity.* Each q-dependent factor is replaced by its endpoint value, using the fact that the factors are expectations of increasing functions of a correlated normal pair. The untreated-risk factor is bounded via $\sigma(x) \le e^x$ and the normal moment generating function, maximised over q in closed form. The bound is two one-dimensional integrals of fixed functions, with no grid over q.

Write $R_0(q)$ for the AUROC of the same score under the same treatment policy when treatment has no effect ($e \equiv 0$), and $R_\delta(q)$ for its AUROC when treatment has effect e. The displacement is $\Delta(q) := R_0(q) - R_\delta(q)$. The proof neither claims nor uses monotonicity of R_δ in q (Remark 3).

Lemma 3 (Dominance). *Let $q \in [0,1]$, $e \in [0,1]$ constant, $t(s)$ nondecreasing, and $W = 1 - e\,t$. Let X and X' be the score among cases without and with a treatment effect, and Y and Y' the score among controls. Then $X' \preceq_{\mathrm{st}} X$ and $Y' \succeq_{\mathrm{st}} Y$, and consequently $\Delta(q) \geq 0$.*

Proof. $m_q(s) = \mathbb{E}[p_0(qs + \sqrt{1-q^2}X)]$ with $X \sim \mathcal{N}(0,1)$ is nondecreasing in s, since p_0 is nondecreasing and $q \geq 0$. Case densities are $\propto \varphi m_q$ and $\propto \varphi m_q W$. Their ratio W is nonincreasing, so X' is smaller than X in the likelihood-ratio order, hence in the usual stochastic order. Control densities are $\propto \varphi(1-m_q)$ and $\propto \varphi(1-m_qW)$, with ratio $(1-m_qW)/(1-m_q) = 1 + e\,t\,m_q/(1-m_q)$, a product of nonnegative nondecreasing factors, hence nondecreasing, so $Y' \succeq_{\mathrm{lr}} Y$. Finally $R = \Pr(X > Y)$ with $X \perp\!\!\!\perp Y$ is nondecreasing in the law of X and nonincreasing in that of Y for the usual stochastic order, because $y \mapsto \Pr(X > y)$ is nonincreasing and $x \mapsto \Pr(x > Y)$ nondecreasing. Hence $R_\delta \leq R_0$. □

Splitting the difference,

$$\Delta = \underbrace{\int (F_{X'} - F_X)\,\mathrm{d}F_Y}_{\text{case side}} + \underbrace{\big[\Pr(X' > Y) - \Pr(X' > Y')\big]}_{\geq 0 \text{ by Lemma 3}}, \tag{9}$$

so Δ is bounded below by the case-side term alone. For the threshold policy write $C_q(y) := \mathbb{E}[p_0(z)\mathbf{1}[s \leq y]]$ and $\widehat{C}_q(y) := \mu - C_q(y)$, so $P_q = \mu - \delta\widehat{C}_q(\tau_0)$.

Lemma 4 (Exact case-CDF shift). *For all y, $F_{X'}(y) - F_X(y) = \dfrac{\delta}{\mu P_q}\, C_q(y \wedge \tau_0)\, \widehat{C}_q(y \vee \tau_0) \;\geq\; 0$.*

Proof. The unnormalised case densities are φm_q (mass μ) and $\varphi m_q W$ (mass P_q). For $y \leq \tau_0$, $W \equiv 1$ on $(-\infty, y]$, so $F_{X'}(y) - F_X(y) = C_q(y)(1/P_q - 1/\mu) = \frac{\delta}{\mu P_q} C_q(y)\widehat{C}_q(\tau_0)$. For $y > \tau_0$, $1 - F_{X'}(y) = (1-\delta)\widehat{C}_q(y)/P_q$ and $1 - F_X(y) = \widehat{C}_q(y)/\mu$, so the difference is $\widehat{C}_q(y)\big[P_q - (1-\delta)\mu\big]/(\mu P_q) = \frac{\delta}{\mu P_q}\widehat{C}_q(y)C_q(\tau_0)$, using $P_q - (1-\delta)\mu = \delta(\mu - \widehat{C}_q(\tau_0)) = \delta C_q(\tau_0)$. □

Lemma 5 (Endpoint ordering in q). *$\widehat{C}_q(y)$ is nondecreasing in q for every y. Hence for all $q \in [0,1]$, $\widehat{C}_q(y) \geq \widehat{C}_0(y) = \mu(1 - \Phi(y))$ and $C_q(y) \geq C_1(y) = \int_{-\infty}^{y} \varphi\, p_0$.*

Proof. $\widehat{C}_q(y) = \mathbb{E}[p_0(z)\mathbf{1}[s > y]]$ is the expectation of a product of two nondecreasing functions of a jointly Gaussian pair with correlation q, hence nondecreasing in q by Slepian's inequality. Equivalently, Price's theorem applied to a smooth approximation of the indicator gives $\partial_q \widehat{C}_q(y) = \varphi(y)\mathbb{E}[p_0'(z) \mid s = y] \geq 0$. The bounds are the values at $q = 0$, where $s \perp\!\!\!\perp z$, and at $q = 1$, where $s = z$. Since $C_q = \mu - \widehat{C}_q$, the ordering reverses for C_q. □

Lemma 6 (Risk envelope). *For the logistic link $p_0(z) = \sigma(b_0 + b_1 z)$ with $b_1 \geq 0$, $m_q(y) \leq \bar{m}(y) := \min\{1, \exp(b_0 + M(y))\}$ with $M(y) = b_1^2/2$ for $y < 0$, $y^2/2 + b_1^2/2$ for $0 \leq y \leq b_1$, and $b_1 y$ for $y > b_1$.*

Proof. From $\sigma(x) \leq e^x$ and the Gaussian moment generating function, $m_q(y) \leq \exp(b_0 + b_1 q y + \frac{1}{2}b_1^2(1-q^2))$. The exponent $f(q) = b_1 q y + \frac{1}{2}b_1^2(1-q^2)$ is concave with $f'(q) = b_1 y - b_1^2 q$, so on $[0,1]$ it is maximised at $q = 0$ if $y < 0$, at $q = y/b_1$ if $0 \leq y \leq b_1$, and at $q = 1$ if $y > b_1$. Any other link needs only some envelope $\bar{m}$. □

Proof of Proposition 2. By (9) and Lemma 4, with $\mathrm{d}F_Y = \varphi(1-m_q)/(1-\mu)$,

$$\Delta(q) \geq \frac{\delta}{\mu P_q(1-\mu)} \int_{-\infty}^{\infty} C_q(y \wedge \tau_0)\, \widehat{C}_q(y \vee \tau_0)\, \varphi(y)\big(1 - m_q(y)\big)\,\mathrm{d}y,$$

an integral of a pointwise nonnegative integrand. We bound each q-dependent factor by a q-free one. By Lemma 5, $P_q \leq \mu(1 - \delta\pi)$, so $1/P_q \geq 1/(\mu(1-\delta\pi))$. On $y \leq \tau_0$, $C_q(y) \geq C_1(y)$

δ	τ_0	bound $\underline{\Delta}$	$\min_q \Delta$	at q	$\Delta(q \to 1)$
0.81	$+1.0$	0.0190	0.0716	0.00	0.1387
0.48	$+1.0$	0.0106	0.0404	0.00	0.0702
0.15	$+1.0$	0.0031	0.0121	0.00	0.0194
1.00	0.0	0.0531	0.2708	0.00	0.4695
0.90	$+2.0$	0.0029	0.0120	0.00	0.0334
0.50	-1.0	0.0053	0.0366	**1.00**	0.0366
1.00	-2.0	0.0148	0.4903	0.00	0.7833
0.05	$+1.5$	0.0005	0.0018	0.00	0.0039

Table 3: Equation (10) against the exact displacement, minimised over 25 values of q and computed from (1), for the logistic model with $b_0 = -2.2$ and $b_1 = 1.3$. The guarantee holds in every row and is loose by a factor of 3.8 to 33. The slack comes from discarding the control-side term in (9) and from the crude envelope $\bar{m}$. Δ is usually smallest at $q = 0$ but not always (row 6), so a bound argued through $\Delta(0)$ alone would be false in general.

and $\widehat{C}_q(\tau_0) \geq \mu\pi$. On $y > \tau_0$, $C_q(\tau_0) \geq C_1(\tau_0)$ and $\widehat{C}_q(y) \geq \mu(1 - \Phi(y))$. Everywhere, $1 - m_q(y) \geq 1 - \bar{m}(y) \geq 0$ by Lemma 6. This gives

$$\Delta(q) \;\geq\; \underline{\Delta} \;:=\; \frac{\delta}{\mu^2(1-\delta\pi)(1-\mu)} \Bigg[\mu\pi \int_{-\infty}^{\tau_0} C_1(y)\varphi(y)\big(1-\bar{m}(y)\big)\,\mathrm{d}y + C_1(\tau_0) \int_{\tau_0}^{\infty} \mu\big(1-\Phi(y)\big)\varphi(y)\big(1-\bar{m}(y)\big)\,\mathrm{d}y \Bigg]. \quad (10)$$

The bound is positive because $\delta > 0$, $\pi \in (0,1)$, $C_1(\tau_0) > 0$ and $\bar{m} < 1$ on a set of positive measure. □

Against the exact minimum the bound is loose by a factor of 3.8 to 7.0 in the seven ordinary rows of Table 3, and by 33.1 in the extreme row $\delta = 1$, $\tau_0 = -2$. That row treats 98% of patients with certainty of cure, so the guarantee degrades only where the configuration stops being a clinical scenario.

Remark 2 (Qualitative version). If only $\inf_q \Delta > 0$ is needed, it follows from Lemma 3 with strict inequality together with continuity of Δ on the compact interval $[0,1]$. Δ is continuous because P_q is bounded away from 0 and 1 and the integrands depend continuously on q.

B.1 Monotonicity in q is not claimed

Remark 3. With $g_q(s) := \mathbb{E}[p_0'(z) \mid s] \geq 0$, Price's theorem applied to $P = \mathbb{E}[p_0 W]$ and $I = \mathbb{E}[p_0 \Phi W]$ through a smooth approximation of W gives $\mathrm{d}P/\mathrm{d}q = -\delta\varphi(\tau_0)g_q(\tau_0)$ and $\mathrm{d}I/\mathrm{d}q = \int \varphi^2 g_q W - \delta\Phi(\tau_0)\varphi(\tau_0)g_q(\tau_0)$, hence

$$\frac{\mathrm{d}R}{\mathrm{d}q} = \frac{\int \varphi^2 g_q W - \delta\varphi(\tau_0)g_q(\tau_0)\big[\Phi(\tau_0) - P - R(1-2P)\big]}{P(1-P)}. \quad (11)$$

The gain from quality is a bulk term over the whole score range. The targeting penalty is a boundary term supported at the alert threshold and weighted by the risk gradient there. At $\delta = 1$ the case density is supported on $(-\infty, \tau_0]$ and proportional to φm_q, so association under $\mathrm{d}\nu = \varphi \mathbf{1}_{(-\infty,\tau_0]}/\Phi(\tau_0)$ gives $I \geq P\Phi(\tau_0)/2$ and hence $R \geq (\Phi(\tau_0) - P)/(2(1-P))$. The bracket in (11) then obeys $\Phi(\tau_0) - P - R(1-2P) \leq (\Phi(\tau_0) - P)/(2(1-P)) \leq \Phi(\tau_0)/2$. For a flat risk gradient, the condition that the bulk term exceeds the boundary term reduces to $\sqrt{2}\,\Phi(\sqrt{2}\tau_0) \geq e^{-\tau_0^2/2}\Phi(\tau_0)$. This holds for all τ_0, and the two sides agree to leading order as $\tau_0 \to -\infty$, where by the Mills expansion their ratio is $1 + 1/(2\tau_0^2) + O(\tau_0^{-4})$.

In general, however, we observe that the bulk term exceeds the boundary term but do not prove it. On a dense grid the minimum ratio is 1.019, attained as everyone is treated, and monotonicity also held with effectiveness rising with severity and with an absolute-risk alert threshold. Proposition 2 does not use it, so we do not claim it. A proof would require a compact parameter box, a Lipschitz bound on (11) over that box, and a grid finer than the bound.

C CLOSED FORMS AT GENERAL QUALITY, AND SCOPE

We first treat a stochastic policy at zero quality, then extend the calculation to scores of any quality under a probit risk.

Corollary 4 (Probit propensity)**.** *For the separable stochastic policy $t(s) = \Phi(\alpha + \beta s)$ with $\beta \geq 0$, put $k := \alpha/\sqrt{1+\beta^2}$ and $\varrho := \beta/\sqrt{2(1+\beta^2)}$. Then $\bar{t} = \Phi(k)$ and $\mathrm{Cov}(U, t(U)) = \Phi_2(0, k; \varrho) - \frac{1}{2}\Phi(k)$.*

Proof. Let W_1, W_2 be standard normal, independent of each other and of s. Then $\mathbb{E}[\Phi(s)\Phi(\alpha + \beta s)] = \Pr(W_1 - s \leq 0,\ W_2 - \beta s \leq \alpha)$. The pair $(W_1 - s,\ W_2 - \beta s)$ is centred Gaussian with variances 2 and $1 + \beta^2$ and covariance β. Standardising gives correlation ϱ and upper limits 0 and k. Subtracting $\mathbb{E}[U]\bar{t} = \frac{1}{2}\Phi(k)$ gives the covariance. □

At a 50% rate, with $\alpha = 0$, the covariance reduces to $\arcsin(\varrho)/2\pi$, since $\Phi_2(0, 0; \varrho) = \frac{1}{4} + \arcsin(\varrho)/2\pi$. As $\beta \to \infty$, $\varrho \to 1/\sqrt{2}$ and the covariance tends to $1/8$, the threshold value at $\bar{t} = \frac{1}{2}$, as Lemma 2 requires. Under a probit p_0 the same calculation extends to scores of any quality, with both terms of (1) becoming Gaussian orthant probabilities in Φ_2 and Φ_3.

Lemma 7 (Probit reduction)**.** *If $p_0(z) = \Phi(b_0 + b_1 z)$ then, with $v := \sqrt{1 + b_1^2(1 - q^2)}$, $m_q(s) = \Phi(c_0 + c_1 s)$ with $c_0 = b_0/v$ and $c_1 = b_1 q/v$.*

Proof. $z \mid s \sim \mathcal{N}(qs, 1 - q^2)$ and $\mathbb{E}[\Phi(a + bZ)] = \Phi\big((a + b\,\mathbb{E}Z)/\sqrt{1 + b^2\,\mathrm{Var}\,Z}\big)$ for Gaussian Z. Apply with $a = b_0$, $b = b_1$. □

Proposition 3 (Closed form under a probit link)**.** *Let $p_0 = \Phi(b_0 + b_1 z)$ and let the policy be the threshold policy. Put $k := c_0/\sqrt{1 + c_1^2}$, $\varrho_{sB} := -c_1/\sqrt{1 + c_1^2}$, $\varrho_{sA} := -1/\sqrt{2}$ and $\varrho_{AB} := c_1/\sqrt{2(1 + c_1^2)}$. Then*

$$P = \Phi(k) - \delta\big[\Phi(k) - \Phi_2(\tau_0, k; \varrho_{sB})\big],$$
$$I = \Phi_2(0, k; \varrho_{AB}) - \delta\big[\Phi_2(0, k; \varrho_{AB}) - \Phi_3\big((\tau_0, 0, k); \varrho_{sA}, \varrho_{sB}, \varrho_{AB}\big)\big],$$

and R follows from (1)*.*

Proof. Let V, W_1 be standard normal, independent of each other and of s, and set $A := (W_1 - s)/\sqrt{2}$, $B := (V - c_1 s)/\sqrt{1 + c_1^2}$, so (s, A, B) is centred Gaussian with unit variances and the stated correlations. For P, Lemma 7 gives $\mathbb{E}[\Phi(c_0 + c_1 s)] = \Pr(B \leq k) = \Phi(k)$ and $\mathbb{E}[\Phi(c_0 + c_1 s)\mathbf{1}[s > \tau_0]] = \Phi(k) - \Phi_2(\tau_0, k; \varrho_{sB})$, and $P = \mathbb{E}[m_q W]$. For I, $\mathbb{E}[\Phi(s)\Phi(c_0 + c_1 s)] = \Pr(A \leq 0, B \leq k) = \Phi_2(0, k; \varrho_{AB})$ and $\mathbb{E}[\Phi(s)\Phi(c_0 + c_1 s)\mathbf{1}[s \leq \tau_0]] = \Phi_3$. Subtracting gives the $\mathbf{1}[s > \tau_0]$ term, and $I = \mathbb{E}[\Phi(s) m_q W]$. □

The expression is closed only up to the trivariate normal CDF, which has no elementary form. The logistic link admits no such reduction, since $\mathbb{E}[\sigma(a + bZ)]$ is not elementary. Under the standard surrogate $\sigma(x) \approx \Phi(x/1.702)$, Proposition 3 is an analytic stand-in for the simulator rather than an exact description of it.

C.1 SCOPE

The conclusion $R(0) < \frac{1}{2}$ requires $q = 0$ and $\mathrm{Cov}(U, \kappa(U)) > 0$. Four cases break it. In the first three the covariance is zero or negative, and in the fourth $q = 0$ fails. Numbers are at $q = 0$ with $p_0 = \sigma(-2.2 + 1.3z)$ ($\mu = 0.1535$), computed from (3) and confirmed by simulation at $n = 8 \times 10^6$.

Treatment unrelated to the score. If κ is a.e. constant, as when $t(s, z) = t(z)$ and clinicians act on severity while ignoring the score, then $\mathrm{Cov}(U, \kappa(U)) = 0$ and $R(0) = \frac{1}{2}$ exactly, however effective the treatment. Simulation with $t \equiv 0.2$ and $e = 0.9$ gives $R = 0.4992$. The displacement is a property of the coupling between score and treatment, not of treatment.

Harmful treatment. If $e < 0$ then $\kappa < 0$, and for a nondecreasing policy $\mathrm{Cov}(U, \kappa(U)) < 0$, so $R(0) > \frac{1}{2}$. A score with no predictive value then appears predictive because the action it triggers is harmful. Equation (8) holds with $\delta < 0$. At $\tau_0 = 1$ it gives $R(0) = 0.542$ for $e = -0.5715$, and 0.5371 for $e = -0.5$ (simulation 0.5361).

Policies that use the score inversely. Separability is not needed, but monotonicity in s is. Non-separable policies that are nondecreasing in s at every z behave as the proposition says. $t = \Phi(1.5s + z)$ gives $R(0) = 0.3076$ (simulation 0.3074), and $t = \mathbf{1}[s > 1]\mathbf{1}[z > 0.5]$ gives 0.4455 (simulation 0.4472). The sign reverses for a policy that decreases in s at fixed severity, which describes a score used as a reassurance signal, so that escalation happens when severity outruns the score. With $t = \mathbf{1}[z - s > c]$ and $e = 0.9$, $c = 2.595$ gives $R(0) = 0.5418$ (simulation 0.5420), and $t = \Phi(2(z - s - 2))$ gives 0.5768 (simulation 0.5759). Non-monotone policies need not reverse the sign, and $t = \mathbf{1}[|s - z| < 0.75]$ gives 0.4257.

Pooling strata whose case mix and score level co-move. This case violates the hypothesis rather than contradicting the result. Suppose two eras of equal size with $z \sim \mathcal{N}(\pm\lambda, 1)$ and $s \sim \mathcal{N}(\pm\lambda, 1)$, the score independent of severity within each era. Pooled, $\mathrm{corr}(s, z) = \lambda^2/(1 + \lambda^2) \neq 0$, so the score predicts risk through the era. At $\lambda = 1.3$ (pooled $q = 0.63$) the two within-era AUROCs are 0.21 and 0.50, while the pooled AUROC is 0.5385. The condition $q = 0$ must therefore be checked in the population being analysed, not stratum by stratum. A policy that varies across strata is otherwise harmless, since a mixture of policies nondecreasing in s is nondecreasing in s.

D Decision metrics

The AUROC loss from masking $\mathrm{FiO_2}$ in the benchmark ablation is a ranking statistic. Here we report what it costs at an operating point a deployment could use, a fixed alert budget of the top 5% of the 4,033-patient test set, which gives 202 alerts in every arm. Deltas are full minus masked, pooled across the two models and clustered by seed ($n = 5$).

	Δ	95% CI	p
sensitivity	$+0.0194$	$[+0.0018, +0.0370]$	0.038
positive predictive value	$+0.0282$	$[+0.0026, +0.0538]$	0.038
number needed to alert	-0.094	$[-0.180, -0.008]$	0.039

Table 4: Masking $\mathrm{FiO_2}$ at a fixed alert budget. The number needed to alert rises from 1.71 to 1.80 for GARLIC and from 1.84 to 1.94 for Warpformer. The masked model needs more alerts per true case, so Δ = full − masked is negative. Per model the effects are borderline ($p = 0.067$ to 0.092). Pooling and seed clustering follow the procedure of Table 1.

The gain in net benefit is positive at every threshold examined and reaches conventional significance at five of seven (Table 5).

threshold p_t	0.05	0.10	0.15	0.20	0.30	0.40	0.50
Δ NB per 1,000	$+1.33$	$+2.04$	$+2.24$	$+3.09$	$+1.96$	$+1.95$	$+1.54$
p	0.009	0.007	0.003	0.003	0.073	0.010	0.085

Table 5: Gain in net benefit from keeping $\mathrm{FiO_2}$, in net true positives per 1,000 patients. All 14 threshold by model cells are positive. Significance is not monotone in the threshold, since $p_t = 0.40$ reaches $p = 0.010$ while its neighbours at 0.30 and 0.50 do not. We therefore report the row descriptively. The cells share seeds and use nested thresholds, so they are not seven independent tests.

The yield does not establish bedside benefit. Expressed as cases, the full model identifies $+6.0$ more of 294 events $[-0.7, +12.7]$ for GARLIC and $+5.4$ $[-1.3, +12.1]$ for Warpformer. Per model the interval crosses zero. The case counts and the sensitivity delta are the same measurement at different aggregation ($0.0194 \times 294 = 5.70$), so they are not two pieces of evidence. We do not claim that keeping the treatment channel is clearly beneficial at the bedside.

Net benefit is compared at a fixed threshold. The threshold p_t encodes the harm ratio the decision-maker accepts, so a comparison at a fixed alert budget evaluates the two arms at different harm ratios. In our own testing this reversed the sign of the comparison.

E MARKING RATHER THAN MASKING

If treatment blunts a biomarker's signal, analytes a clinician can correct should be less predictive than those a clinician cannot. We tested this in MIMIC-IV across 60 analytes, correlating univariate predictivity with correctability rank.

The primary test is null. The Spearman correlation is $+0.128$ $[-0.138, +0.380]$, $p = 0.33$. The interval excludes any inversion stronger than $\rho = -0.14$, so the test rules out a strong inversion without showing that there is none.

Two problems had to be fixed before the panel was interpretable. The treatment metric as first specified, any treatment within 24 hours, ranged from 0.925 to 0.996 across all 60 analytes, so it measured ICU admission rather than treatment. Rebuilt as organ support within 4 hours, it ranges from 0.245 to 0.920. Measurement intensity is also a strong confounder. The correlation between excursion count and AUROC is $+0.52$ ($p = 2 \times 10^{-5}$), and $+0.465$ ($p = 1.8 \times 10^{-4}$) on the monotone AUROC variant.

The pre-registered kill criterion fired in the opposite direction to the hypothesis. The criterion was the untreated-subgroup control. If treatment contaminates the signal, predictivity should rise when treated patients are removed. It fell instead, by 0.0184 on average ($p = 0.0001$), rising for only 18 of 60 analytes, and it fell most for the most correctable analytes.

specification	ρ(correctability, untreated gain)
raw	-0.538
+ statistical severity-linkage control	-0.537
SOFA band 2–7	-0.441
SOFA band 3–6	-0.267
exact SOFA stratification	$\mathbf{-0.277}$ $[-0.523, -0.001]$, $p = 0.034$

Table 6: About half of the raw association is severity truncation. Untreated patients have a mean SOFA of 2.01 against 5.11 for treated patients, with a variance ratio of 0.62. The effect survives exact stratification at about half its raw size. The column holds two quantities. The first two rows compare the untreated subgroup with the full cohort, and the band and exact rows compare untreated with treated patients within severity strata. Exact stratification uses the 59 analytes that can be scored in both arms within a SOFA level, which excludes "Protein, Total". Every other row uses all 60.

Design guards. The cardiovascular and respiratory SOFA components are excluded from the severity score, since they encode vasopressors and ventilation. The three analytes that are themselves score components are scored leave-one-out. The prediction specific to truncation fails, since severity linkage does not predict the untreated gain ($\rho = -0.082$, $p = 0.54$), and correctability is independent of volatility ($\rho = -0.028$, $p = 0.83$). Leave-one-out scoring is applied to the severity linkage. Exact stratification uses the full SOFA score, so for platelet count, bilirubin and creatinine it stratifies partly on the analyte itself.

We report the result as marginal. The second treatment metric does not corroborate it (-0.104, $p = 0.43$), so the claim rests on the correction-rate metric alone. The two metrics measure different constructs, resolution of the abnormality and triggering of organ support. The difference is interpretable, but it is not the agreement the two-metric design was built to provide.

Correctable analytes appear to be predictive largely because they mark patients who go on to receive organ support, not because treatment erased their signal. That reading is the opposite of masking, and it is why Appendix G treats the haemoglobin result as misspecification rather than as masking.

	MIMIC-IV	eICU
n / 24-hour deaths	80,624 / 314	130,946 / 2,393
haemoglobin, linear	0.5502 $[0.519, 0.583]$	0.5785
haemoglobin, spline	0.6140 $[0.577, 0.641]$	0.5980
spline − linear	$+0.0638$ $[+0.027, +0.099]$	$+0.0194$ $[+0.0125, +0.0257]$
age, linear	0.5944 $[0.564, 0.622]$	0.5958
age, spline gain	-0.0076	-0.001
reported paradox (age − haemoglobin, linear)	0.0442	0.0173
haemoglobin (spline) − age	$+0.0189$, $P = 0.78$	$+0.0009$, $P = 0.51$

Table 7: A spline recovers discrimination for haemoglobin that the linear fit discards. The same flexibility gains nothing for age, which is monotone. Intervals and probabilities are from 300 paired resamples, and $P(\Delta > 0) = 1.000$ for the spline gain in both cohorts. The two contrast rows are estimated on the paired resamples and so do not equal the differences of the point estimates above them.

F SEED CHURN AND VARIANCE DECOMPOSITION

Alert-list membership at a constrained operating point changes by roughly 30% between seeds. For GARLIC, masking FiO_2 changes 32.7% $[31.8, 33.7]$ of the list, against a seed-only floor of 31.2% (ratio $1.05\times$, $p = 0.40$). For Warpformer the ablation churn is 23.6%, below its own floor of 29.8%. We therefore make no churn claim. The phenomenon itself is established, and Lopez-Martinez et al. (2022) show it for outpatient risk stratification. Our variance decomposition follows Bouthillier et al. (2021), whose warning that sources are non-additive is why we report a combined interval rather than an initialisation-only error bar.

G THE EARLY WARNING PARADOX

Logan Ellis et al. (2025) report that in an emergency department cohort of 49,011 admissions with 4.74% in-hospital mortality, haemoglobin reaches an AUROC of 0.49 for 24-hour mortality while age reaches 0.65. They attribute the 0.16 gap to treatment masking haemoglobin's signal. The analysis compares two single features in isolation, each fitted linearly, and the authors describe it as intentionally simplified. We reproduce the comparison in two independent health systems, emergency admissions to MIMIC-IV (Johnson et al., 2023) and the eICU Collaborative Research Database (Pollard et al., 2018). In MIMIC-IV the 24-hour window starts at admission, and in eICU at the haemoglobin measurement. We find an explanation that involves no treatment at all, and we then bound how much of the gap treatment masking could explain.

The relationship is U-shaped. Table 7 gives the discrimination in both cohorts. Mortality falls and then rises across the haemoglobin range in both cohorts, as it does in large population cohorts (Lee et al., 2018). Against the nadir decile, the low tail is elevated at $z = +4.42$ ($p = 1 \times 10^{-5}$) in MIMIC-IV and $z = +13.67$ in eICU. The high tail is elevated at $z = +2.98$ ($p = 0.003$) and $z = +5.77$ ($p = 8 \times 10^{-9}$) respectively. About three quarters of patients lie below the nadir in both cohorts (74.2% and 74.4%). A linear fit is therefore fitted through a turning point, not merely imprecise.

The flexibility is not a free parameter. The obvious objection is that a spline would help any feature. Fitting a spline to age lowers its AUROC in both cohorts. The asymmetry is a property of the two relationships, not of the analysis.

No age advantage of the reported size survives. Once haemoglobin is fitted flexibly, the largest age advantage compatible with the data is 0.032 in MIMIC-IV and 0.016 in eICU, against the 0.16 reported. These are five and ten times smaller respectively.

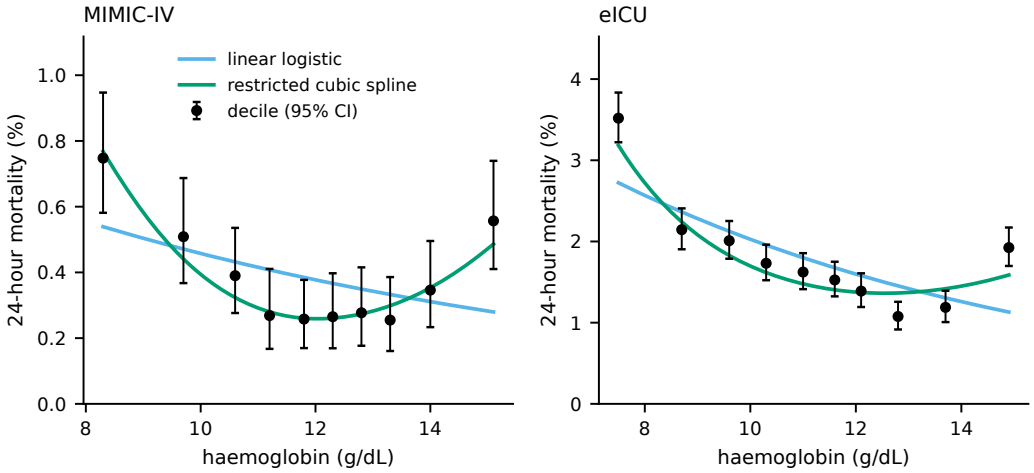


Figure 4: Observed 24-hour mortality by haemoglobin decile, with a linear-logistic and a restricted-cubic-spline fit, in MIMIC-IV emergency admissions and eICU. The relationship turns, so a linear fit is fitted through a turning point rather than merely imprecise.

G.1 Treatment masking explains little of the gap

Suppose the remaining deficit is nonetheless attributed to treatment masking. Proposition 1 fixes how far masking can push R given the observed treatment rate, and Lemma 2 makes that bound tight at a deterministic threshold. Inverting (6) at the observed rate gives the absolute risk reduction the treatment would have to deliver, and hence the counterfactual mortality the treated patients would need to have had.

Suppose the remaining deficit is nonetheless attributed to treatment masking. We simulate masking directly. Among the MIMIC-IV emergency admissions, 1,874 patients had a recorded transfusion within 24 hours and 59 of them died, an observed mortality of 3.15%. Transfusion is identified from ICU input events, so transfusions given only in the emergency department or on the ward are not counted. Masking with effectiveness e hides $k = 59e/(1-e)$ deaths among the 1,815 transfused survivors. We restore these deaths adversarially, lowest haemoglobin first, and refit both models. This recovers the most signal per death, so it approximates the minimum effectiveness that any real masking process would need. Values between simulated k are linearly interpolated (Table 8).

AUROC deficit attributed to masking	required e	counterfactual transfused mortality
Half the misspecification gain (0.032)	0.288	4.4%
Reported paradox (0.044)	0.365	5.0%
Misspecification gain (0.064)	0.467	5.9%
Reduce the spline gain below 0.01	**0.956**	**70.9%**

Table 8: Masking bound. Observed 24-hour mortality among transfused patients is 3.15%, and the implied counterfactual is $3.15/(1-e)$.

We report counterfactual mortality rather than e because no trial compares transfusion against no transfusion in the severely anaemic. Trials of restrictive against liberal thresholds (Hébert et al., 1999; Holst et al., 2014) bound the marginal effect of transfusing more, not the effect of transfusion against none in patients who were actually transfused, and equipoise does not exist for the latter. The effectiveness e therefore has no external anchor. A claim that 70.9% of these patients would have died within 24 hours is, by contrast, one that any clinician can judge.

Sensitivity. The conclusion does not depend on a tight prior over e. Even at $e = 0.50$, with transfusion halving mortality and far beyond any trial estimate, three quarters of the misspecification gain survives (Table 9).

G.2 FURTHER ANALYSES

The transfusion rate does not cap the displacement. We make no claim that the 2.32% transfusion rate caps the achievable displacement. Swept to its representable maximum, the displacement reaches $+0.32$, because transfused patients are concentrated at low haemoglobin, where the signal lies. What is bounded is the effectiveness required, not the displacement available (Table 9).

restored deaths k	$e = k/(59+k)$	displacement of the linear AUROC	misspecification gain, recomputed
3	0.048	$+0.0043$	0.0632 (99%)
5	0.078	$+0.0071$	0.0625 (98%)
13	0.181	$+0.0180$	0.0595 (93%)
59	0.500	$+0.0714$	0.0477 (75%)

Table 9: Each row restores k deaths, and $e = k/(59+k)$ is the effectiveness that hides exactly k. Displacement is the rise in the linear AUROC when those deaths are restored. The recomputed gain is the spline minus linear gap on the same restored data. Restoring deaths raises both curves, so the two columns must not be subtracted. At $k = 59$ the linear AUROC rises from 0.5502 to 0.6216 and the spline from 0.6140 to 0.6692, leaving a gap of 0.0477. Table 8 and this table therefore answer different questions. At $e = 0.467$, masking displaces the linear AUROC by the full 0.064, yet the gap recomputed on the restored data is still 0.0505.

The gain is larger where the turning point is visible. If misspecification is the mechanism, the spline gain should be larger in patients whose haemoglobin spans both arms of the U. Among never-transfused patients, with median haemoglobin 12.0, the gain is $+0.0789$. Among transfused patients, with median 8.8 and almost all on the low arm, it is $+0.0400$. The transfused stratum contains 59 events, so we report this as consistent with the mechanism rather than as a demonstration of it.

H CLASSIFICATION DETAILS FOR THE BENCHMARK AUDIT

The classification covers the 103 channels of the MIMIC-III pipeline. Of the six labelled biomarkers that we reclassify, $\mathrm{FiO_2}$ and intubation status are set by clinicians. The other four (`Ultrafiltrate`, `Chest Tube`, `Jackson-Pratt` and `Pre-admission Output`) are measurements conditioned on treatment rather than settings. Systolic and diastolic blood pressure are swapped at tensor positions 0 and 1 between the reconstructed map and the tensor actually loaded, so an index-based ablation would remove the wrong channel. We release the full classification keyed to tensor position, together with a check that reads the loaded tensor directly.

I PAIRING, POOLING AND SEEDS IN THE BENCHMARK ABLATION

Pairing. Under masking, the correlation between full and ablated runs is $+0.328$ for GARLIC and $+0.746$ for Warpformer, against -0.549 under removal. With ablated-to-full standard deviation ratios of 1.12 and 0.59, the ratio of unpaired to paired variance, $(\sigma_1^2 + \sigma_2^2)/(\sigma_1^2 + \sigma_2^2 - 2\rho\sigma_1\sigma_2)$, is 1.48 and 2.88. We validated masking against removal through this pairing correlation, but not through agreement of the point estimates.

Masking does not change the model ranking. The difference-in-differences between models is $+0.122$ pp $[-0.846, +1.090]$, $p = 0.74$. Reversing the ranking would require a difference above 1.66 pp, and the upper bound is 1.09, so reversal is excluded at the 95% level. The same statistic is the homogeneity test that justifies pooling in Table 1.

Seed alignment changes the result. The treatment arm has five seeds and the placebo arms three. Scored against the five-seed baseline, the $\mathrm{FiO_2}$ delta is 1.14 pp and its ratio to the largest placebo is $1.16\times$. On the three shared seeds, the same quantities are 1.44 pp and $1.46\times$. The five-seed mean is lower mainly because masking $\mathrm{FiO_2}$ hurt least on seed 3 (0.31 pp, against 0.95 to 1.85 pp on the other four). Comparing arms with different seed counts is not a conservative approximation. It targets a different estimand, and here it separates a result from a non-result.

Variance decomposition. Initialisation contributes a standard error of 0.140 (25% of variance) and test-set resampling 0.240 (75%), for a total of 0.278. The Satterthwaite effective degrees of freedom are 60.5, so we report the t interval with those degrees of freedom, $+1.081$ pp $[+0.526, +1.636]$. Treating all variance at 4 degrees of freedom would give $[+0.311, +1.852]$. That worst case penalises the 75% of variance that has many degrees of freedom.

Fixed partition. The benchmark runs on a single fixed data partition. This is a property of the benchmark rather than of our replication, since the split seed is fixed independently of the run seed, so the published error bars also reflect initialisation only.

J ESTIMATOR DETAILS FOR PRACTICE VARIATION

Of 146 hospitals with any vasopressor infusion record, 132 have at least one row with a dose-convertible infusion rate, and 117 of these have at least 40 adult first ICU stays with an APACHE physiology record. The charting-quality filter retains 68 of the 117 hospitals and 56,450 stays. The filter was reconstructed after the original computation, which was not saved. The reconstruction selects exactly the same 68 hospitals, but its Spearman correlations ($+0.419$ and -0.040) differ from those first recorded ($+0.390$ and -0.124), so some detail of the original denominator differed. The partition, and every result that depends on it, is unchanged.

The random-intercept model is fitted on the 68 hospitals and has ICC $= 0.31$. Adding adjustments in turn raises the median odds ratio from 3.00 unadjusted to 3.20 with the acute physiology score and 3.22 with age and sex, and adding APACHE predicted mortality gives the headline 3.21. Hospitals with sicker patients were therefore not the heavier prescribers. Three estimators agree, giving 3.208 under the Laplace approximation, 3.214 under adaptive Gauss–Hermite quadrature and 3.344 under NUTS (95% CrI $[2.68, 4.36]$, no divergences).

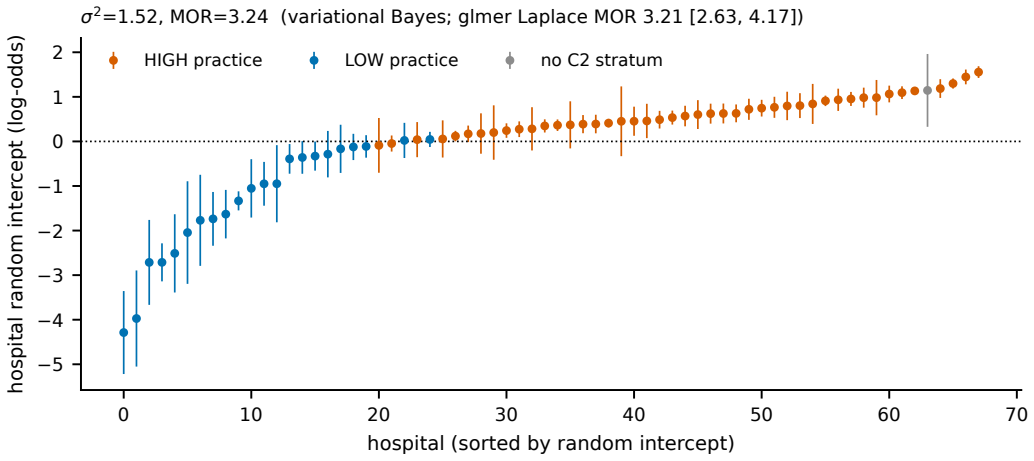


Figure 5: Hospital random intercepts for vasopressor receipt, sorted, with 95% intervals, for the 68 charting-filtered hospitals and coloured by practice stratum. The spread is continuous rather than driven by a few outliers. The variational-Bayes fit shown gives $\sigma^2 = 1.52$ and a median odds ratio of 3.24, consistent with the three estimators above. The grey hospital has fewer than 50 stays and so is absent from the 67-hospital repeat.

K PRACTICE AND GEOGRAPHY

Hospital bed size ($p = 0.65$) and teaching status ($p = 1.0$) do not predict whether a hospital falls in HIGH or LOW practice. Region is blank for 8 of the 116 hospitals, 4.8% of stays, so both region tests use the remaining 108, and no stays are dropped from any model. The eight are unremarkable, with a median vasopressor rate of 0.104 against 0.091 for the rest, which is consistent with the field being missing for administrative rather than clinical reasons. Across the 108, region predicts the

hospital vasopressor rate at $p = 0.031$ and stratum membership at $p = 0.079$. The second test is the same association with the rate dichotomised, so we attribute its larger p-value to lower power and do not read it as balance.

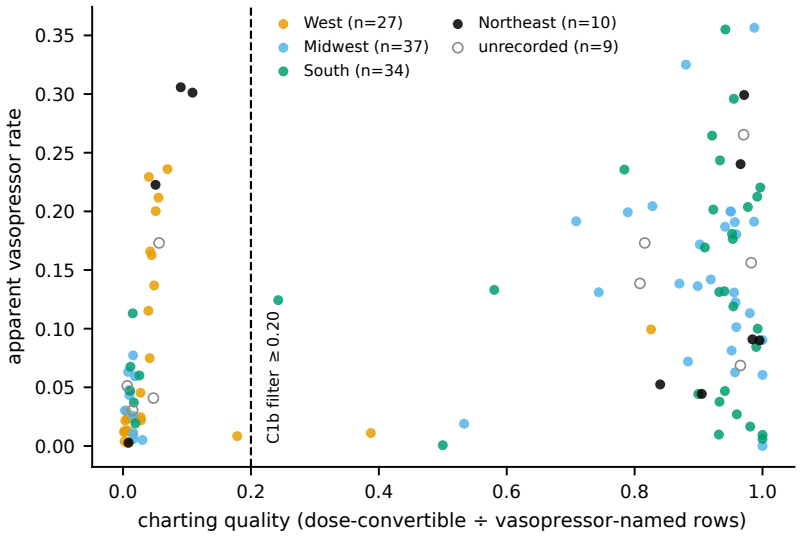


Figure 6: Charting quality against the apparent vasopressor rate for the 117 hospitals, by census region. The filter at 0.20 retains 2 of the 27 Western hospitals, against 73% to 82% elsewhere, so charting and prescribing cannot be separated in that region. Region is unrecorded for 9 of the 117, of which one is also dropped by the 50-stay filter, leaving the 108 hospitals used in both region tests.

The charting-quality filter retains 2 of the 27 Western hospitals, against 73% to 82% of hospitals in each other region (χ^2, $p < 0.0001$). Under the median split of the 67-hospital repeat, the West contributes no HIGH and two LOW hospitals, and region against stratum gives $p = 0.293$ on the 63 hospitals with a recorded region. Neither p-value is evidence of balance. The first comes from 108 hospitals and the second from a table with an empty cell among 63, and both are consistent with one association measured on too few hospitals to resolve it. Eleven of the 67 hospitals change group on re-ranking, all from HIGH to LOW, because the median propensity rises once the low-charting hospitals are removed. None of the eleven is Western, and both remaining Western hospitals are LOW under either ranking.

L Calibration contrasts

Drift is the observed to expected ratio on LOW minus the ratio on HIGH for the same model and stratum, for example $1.226 - 1.047 = 0.179$ for PHYS+VASO among the untreated. For PHYS+VASO, untreated minus treated drift is $+0.163$ $[-0.080, +0.418]$. The change in untreated drift on adding the feature is $+0.0250$ $[-0.0341, +0.0890]$. Under split-level resampling the second contrast has a point estimate of $+0.0209$ and an interval about six times narrower. Its apparent precision came from resampling splits rather than hospitals.

M Replication across model classes

Table 10 gives the replication. The logistic model is not degenerate, with within-practice AUROC of 0.816 in HIGH and 0.811 in LOW practice, against 0.860 and 0.852 for the gradient-boosted model and 0.840 and 0.830 for the multilayer perceptron. The gradient-boosted model handles missing values natively, while the other two require median imputation, so the three arms are not identical. The effects are larger in the two imputed models, and we do not have an explanation for the ordering of magnitudes across model classes.

	GBM	MLP (64, 32)	logistic
HIGH $\rightarrow$ LOW DiD	-0.0080	-0.0103	-0.0156
LOW $\rightarrow$ HIGH DiD	$+0.0042$	$+0.0076$	$+0.0113$
placebo null	$z = -6.06$, 0/19	$z = -7.30$, 0/19	$z = -7.10$, 0/19
calibration contrast, split-level	$+0.0209$	$+0.0320$	$+0.0730$

Table 10: Replication across model classes. These are split-level point estimates, so we attach no intervals. The calibration contrast is the change in the untreated ratio on adding the feature, the same contrast whose hospital-clustered interval contains zero. It is reported only to show that its sign is the same across three model classes, not as three independent confirmations, since the arms share a design and could share a bias. All three columns use the splits of the placebo run.

N WEIGHTING AS AN INSTANCE OF THE ATTENUATION REMARK

Remark 4. Any intervention that acts by attenuating a feature's contribution, including weighting, shrinkage and partial removal, moves along the same ray to first order and so preserves the ratio of validation gain to transfer cost. A remedy must change what the feature means across sites, not how much of it is used.

The stabilised weights have a propensity AUROC of 0.878, lie in $[0.23, 3.00]$ after truncation and give an effective sample size of 84.4%. On the clustered replicates, the division-free contrast $\mathrm{cost}_{\mathrm{IPW}}\,\mathrm{gain}_{\mathrm{naive}} - \mathrm{cost}_{\mathrm{naive}}\,\mathrm{gain}_{\mathrm{IPW}}$ is $+4.7\times10^{-6}$ $[-2.7\times10^{-5}, +3.8\times10^{-5}]$, so the weighting result does not contradict Remark 4. It does not confirm it either. Weighting attenuates the feature so little, 8% of the gain and none of the cost, that the test has almost no power. On the 173 of 200 replicates in which both gains exceed 0.002, so that both ratios are defined, the ratio contrast is $+0.10$ $[-0.51, +0.81]$ against a predicted 0, and on all 200 it is $+0.46$ $[-2.70, +2.60]$. The remark is therefore a first-order result with one consistent instance, not an empirically tested claim. The split-level version of this test looked decisive only because its interval was too narrow.