INTERVAL-WIDTH DRIFT: A CONTROLLED STUDY OF LOCALIZATION, DETECTION, AND ERROR RANKING

Anonymous authors
Paper under double-blind review

ABSTRACT

Can interval widths from a frozen imputer support variable localization, independent-window detection, and individual error ranking interchangeably? We study the evidence needed for each use. With $S = 10$ draws from one coronary-trained flow imputer, raw-width localization AUROC on 40 paired PhysioNet 2012 encounters is 0.987 with shared imputation randomness, but 0.386 with independent randomness; disjoint groups of 40 give 0.447. The latter two estimates are compatible with chance. Monte Carlo-only controls locate a dependence on numerical noise cancellation before patient pairing is removed; paired localization alone is therefore insufficient monitoring evidence. Across window sizes 20–100, width-statistic responses remain small while a direct timing diagnostic increasingly detects the intervention. Source-validation pseudo-masking reveals worse reconstruction than simple linear interpolation and small, measurable widening after context removal. A separate encounter–feature-matched design changes the hidden positions; harder selections increase error with weak average response in both local width and the deployed representation, including a 50-draw sensitivity check. In a separate supervised error-ranking analysis, restoring probability information to the baseline nearly eliminates the apparent benefit of widths. We provide an evaluation protocol that tests these uses separately, varying patient pairing, random-stream coupling and baseline information. The findings delimit this imputer's behavior rather than establishing a universal failure of width monitoring.

1 INTRODUCTION

A frozen clinical prediction pipeline can receive measurements collected at different times or with different feature availability. Its imputation module must then operate under a changed observation process. Raw values, masks, model outputs, and imputation summaries offer different views of this change. The practical question is whether a particular representation adds useful diagnostic information under the same monitoring resources.

We study the marginal distributions of imputation-interval widths. An additive sum of univariate KL divergences, denoted D_2, produces a cohort discrepancy and a feature ledger. Raw numerical width is the primary representation; a nonlinear footprint-of-uncertainty (FOU) projection is an optional alternative. FOU may respond to interval position as well as length and does not inherit numerical interval coverage. The monitored object is conditioned on a frozen imputer, normalization and feature scope. It does not require a prediction head, which is present only in the downstream evaluations that explicitly use one.

Three appealing observations can otherwise be conflated. A ledger can locate variables in a known intervention without detecting that intervention in an independent patient window. An imputer can emit a width response without tracking actual reconstruction difficulty. An auxiliary risk model can benefit from widths relative to a restricted confidence baseline without improving on a more informative use of the same prediction probabilities. These are distinct empirical questions rather than automatic consequences of a two-sample statistic.

Our contribution is a reusable evaluation protocol, instantiated as a controlled case study (Table 1). Timing controls vary random-stream coupling with patients fixed, then patient pairing and window size. A source-validation check establishes reconstruction quality and conditional-width response. An

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.

algebraic analysis identifies information lost by a restricted error-ranking baseline. The transferable lesson is to vary evaluation units, numerical coupling and baseline information: the same width representation can receive different validity assessments under these choices. Timing and difficulty findings concern one coronary-trained flow imputer.

2 Related Work

Shift detection and model monitoring. Dataset-shift taxonomies, covariate-shift correction and importance weighting provide established foundations (Shimodaira, 2000; Sugiyama et al., 2007; Quiñonero-Candela et al., 2009; Moreno-Torres et al., 2012). Kernel MMD, classifier two-sample tests and output-space methods monitor different representations (Gretton et al., 2012; Lopez-Paz & Oquab, 2017; Lipton et al., 2018; Rabanser et al., 2019). Conditional feature-shift tests also address localization (Kulinski et al., 2020); an additive ledger is therefore not the only way to identify changed variables. Concept-drift and clinical-monitoring studies distinguish measurement protocols, coding and patient-mix changes (Gama et al., 2014; Ginart et al., 2022; Subbaswamy et al., 2019; Finlayson et al., 2021). Our monitored representation is the frozen imputer's width vector; the two-sample statistics are standard.

Uncertainty and downstream interpretation. Classical missing-data analysis and multiple imputation distinguish missingness mechanisms from uncertainty about missing values (Rubin, 1976; Little & Rubin, 2019; van Buuren & Groothuis-Oudshoorn, 2011). Modern recurrent, adversarial, variational and score-based imputers provide useful modeling tools (Che et al., 2018; Cao et al., 2018; Yoon et al., 2018; Ivanov et al., 2019; Mattei & Frellsen, 2019; Fortuin et al., 2020; Tashiro et al., 2021). Ensembles and conformal methods study predictive uncertainty and calibration under their respective assumptions (Lakshminarayanan et al., 2017; Tibshirani et al., 2019; Zaffran et al., 2023). We neither train a new imputer nor claim a new coverage guarantee. Uncertainty under shift, unlabeled performance estimation and external benchmarks already show why detecting change does not automatically predict degradation (Ovadia et al., 2019; Garg et al., 2022; Koh et al., 2021). We test this distinction for interval-width monitoring with explicit competing representations and evaluation budgets.

Explanation-space monitoring. SHAP and Explanation Shift monitor model-related representations (Lundberg & Lee, 2017; Mougan et al., 2025). Explanation Shift evaluates changes in how a fixed model's predictions are explained. Width monitoring instead diagnoses an imputation module, including settings without a clinical head. Our historical width-surrogate SHAP adaptation and later actual-head column-occlusion controls are disclosed separately; neither is presented as a faithful replacement for the full Explanation Shift method.

3 Method

3.1 Frozen Pipeline and Width Construction

Let x contain a sample's input sequence, observed mask, availability mask, and valid length L_x. The source-trained imputation module, source normalizer, target-feature scope, and membership parameters remain fixed when scoring a target cohort. Target labels are unavailable to the monitor. A prediction head may consume the resulting imputed features; the width monitor itself depends on the frozen imputer and does not require its gradients or prediction labels.

The PN2012 case study uses a coronary-trained conditional flow imputer, conditioned on values, masks, time and valid lengths. Gaussian draws are integrated at missing in-scope positions while observed values stay fixed. The training loss and conditional inputs are specified in Appendix A.1; the model is not trained to predict clinical errors.

Write $\widehat{g}_S(x;\omega)$ for the finite-draw representation: S is the number of imputations and ω the random stream under a fixed device, encounter order and batching schedule. For feature j at valid time t, let

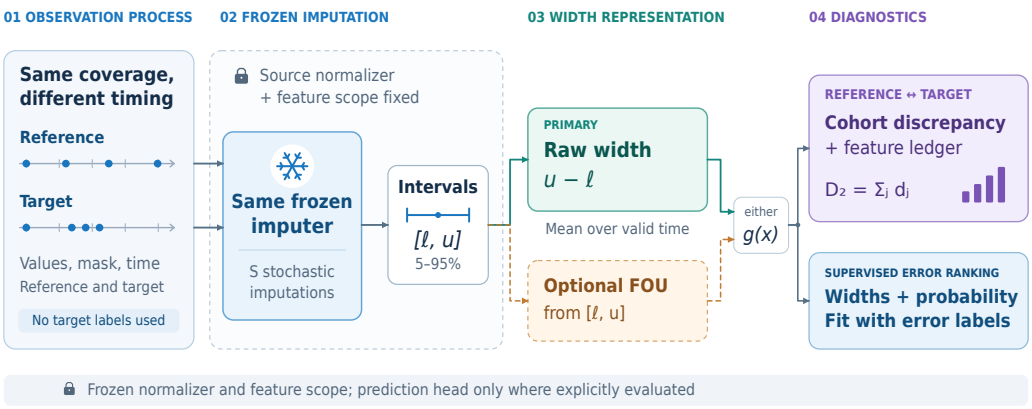


Figure 1: Frozen interval-width pipeline. Numerical intervals yield raw widths or optional FOU summaries. The upper branch compares unlabeled cohorts. The lower branch evaluates supervised error ranking and additionally requires prediction probabilities and error labels during risk-model fitting; it is not the reference-width deviation score $a_j(x)$. Timing patterns are schematic.

$\tilde{x}_{tj}^{(s)}(x;\omega)$ be draw s in source-normalized units. With $\alpha = 0.05$, define

$$\ell_{tj} = Q_\alpha(\tilde{x}_{tj}^{(1:S)}), \quad u_{tj} = Q_{1-\alpha}(\tilde{x}_{tj}^{(1:S)}), \qquad \widehat{g}_{S,j}^{\mathrm{raw}}(x;\omega) = \frac{1}{L_x}\sum_{t=1}^{L_x}(u_{tj} - \ell_{tj}).$$

Quantiles use linear interpolation between ordered draws. We abbreviate the realized representation as $g_j(x)$ below; freezing model parameters does not remove its Monte Carlo variability. Observed cells are preserved as singleton intervals. Binary features also have singleton intervals at the sample mean; they therefore do not supply numerical interval-width coordinates. Padding is excluded from time aggregation.

Let $c_j(x)$ denote observation coverage over valid times and $\bar{u}_j^{\mathrm{miss}}(x)$ the mean numerical width at missing positions. Whenever missing positions exist,

$$g_j^{\mathrm{raw}}(x) = (1 - c_j(x))\,\bar{u}_j^{\mathrm{miss}}(x).$$

For complete observation, $g_j^{\mathrm{raw}}(x) = 0$. Thus raw width mixes missingness frequency with the imputer's conditional response. Holding observation coverage fixed removes the first factor's variation, but timing and context may still change the second. A width alarm alone does not identify a clinical population change.

The optional FOU representation averages the ranges of five frozen Gaussian membership functions over each interval and valid time. It can depend on interval location as well as length and carries no numerical coverage guarantee; Appendix A gives its formula and parameters. Unless stated otherwise, $S = 10$ without validation-derived interval expansion.

3.2 Marginal Statistic and Diagnostics

For $g \in \{g^{\mathrm{raw}}, g^{\mathrm{FOU}}\}$, viewed as a map of the augmented input (x, ω) under the stipulated sampling law, the marginal statistic is

$$D_g(P, Q) = \sum_{j \in \mathcal{A}} \mathrm{KL}(P_{g_j} \| Q_{g_j}).$$

Standard marginal-KL properties are collected in Appendix B. They do not guarantee finite-sample detection or sensitivity to dependence changes that preserve every width marginal.

The implementation sums feature contributions $D_2 = \sum_j d_j$, using regularized Gaussian KDEs on a shared 200-point grid. Appendix A specifies the estimator, bandwidth and boundary rules.

Coordinates deterministic at the same value in both cohorts contribute zero. One-sided degeneracy is retained with a bandwidth floor and flagged in the coverage report. The implementation also skips coordinates with fewer than five finite samples on either side; such coordinates cannot support a monitoring claim. The d_j form the cohort feature ledger. The separate reference-width deviation diagnostic is defined in Appendix A.12; it is not the supervised error-risk model evaluated in Section 5.3.

3.3 Calibration and Comparison Contract

Randomization respects participants or encounters. Natural-cohort audits compare source-validation and target-test widths under the same source stack, using 1999 whole-cluster permutations; Appendix A.2 separates this protocol from window tests. Permutation validity requires exchangeability of the stated units. Bootstrap units and the uncertainty they omit are specified per experiment.

Comparators include observed/imputed summaries, temporal masks, and actual prediction outputs when a clinical head exists. Raw width is primary and FOU optional. Complete suites share the permitted data and are calibrated jointly to the same nominal alarm budget; an unadjusted OR of component alarms would not establish incremental value. Fixed-budget triage is evaluated separately.

4 Experimental Design

Table 1 gives the evaluation protocol: each proposed use requires a direct comparison at its own unit of inference. PN2012 timing and difficulty analyses share one coronary-trained flow imputer and eight numerical width coordinates. NIH-DEP supplies a separate supervised error-ranking study. These retrospective cohorts have been used during development; follow-up endpoints and seeds were fixed before new computations, not prospectively registered on unseen clinical data. Favorable and unfavorable earlier results remain in the appendix.

Table 1: Evidence requirements for three uses of width. These checks can be applied to another frozen pipeline; their outcomes need not match this case study. Repeated seeds and windows do not create independent patients.

Intended use	Evidence to test directly	Insufficient substitute
Independent-window monitoring	Matched unperturbed windows, common alarm budget, continuous responses and transitions; vary patient pairing, random streams and n	Paired feature-localization AUROC or separated representation fixtures
Difficulty-related width diagnosis	State whether query cells are fixed or hidden positions change within matched encounter–features; compare error, coverage, local and aggregate width	Mean width on selected hidden cells without error or aggregation checks
Incremental error ranking	Information-preserving probability baseline; same risk learner and participant-grouped folds	Gain over a margin baseline that discards probability direction

5 Results

5.1 Localization and Detection on the Same Representation Bank

A controlled change in encounter pairing. Matched-coverage timing bursts preserve each affected encounter–feature's observation count and observed-value multiset while rearranging measurement times. Our primary bridge uses ten perturbation seeds, two affected variables and three severities, giving 30 conditions on the same 223 medical encounters. These comparisons use $S = 10$, resetting the same seed before each full export with fixed encounter order and batches of 32. Full noise tensors align by draw, encounter, time and feature before masking (Appendix A.8); changed masks do not change random-number consumption. We evaluate all 223 paired encounters, the 40 paired target encounters in each window, and two disjoint groups of 40 drawn from the test pool. Three fresh

window pairs per condition give 90 comparisons. Clinical encounters can recur across windows; no window contains repeated or overlapping reference/target encounters.

Shared-stream raw-width AUROC/Recall@2 are 0.997/0.983 on all 223 paired encounters and 0.987/0.972 on paired subsets of 40. Disjoint groups of 40 give 0.447/0.217 (Figure 2); the AUROC interval $[0.338, 0.556]$ includes chance. Sample size alone does not explain the separation, but patient pairing and random-stream coupling still change together. Appendix A.6 retains the archived broader benchmark.

Patient pairing without shared random numbers. We export three independent unperturbed $S = 10$ banks, retaining the same 223 patients, model, batch size and all 30 existing timing interventions. On the same 40-patient subsets, mean raw AUROC falls to 0.386 ($[0.224, 0.563]$), close to the MC-only control's 0.384 when neither input is altered. Median KL rises from 0.00115 with shared randomness to 0.222, versus 0.221 for MC-only comparisons. FOU similarly falls from 0.960 to 0.391. Full-cohort raw AUROC also falls to 0.382. Thus cancelling Monte Carlo variation is already essential to the excellent paired ranking in this $S = 10$ implementation; patient composition is not required to destroy it. Intervals resample perturbation seeds, conditioning on the three independent banks and clinical pool.

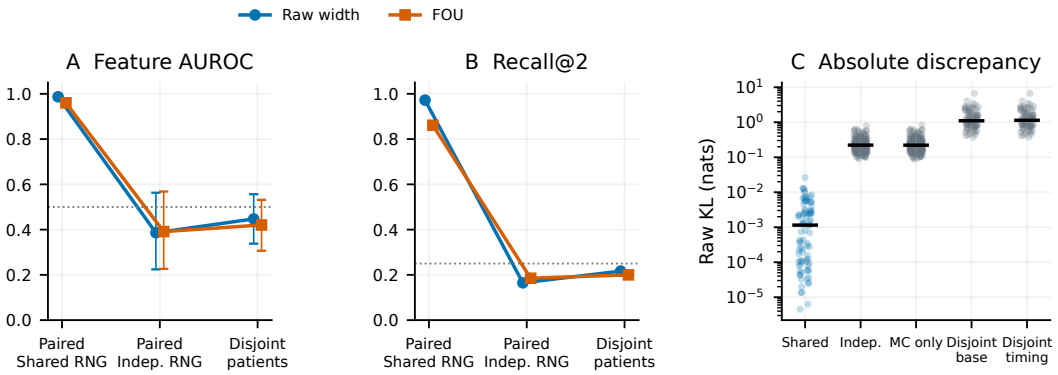


Figure 2: Coupling boundary at the same sample size, $n = 40$, with $S = 10$ imputation draws throughout. A–B: localization with the same patients/shared randomness, the same patients/independent randomness, and disjoint patients. Dotted lines indicate chance. Independent-stream values average three fixed baseline banks. C: raw KL for these comparisons and unperturbed controls; black lines are medians. MC-only comparisons alter no input. AUROC bars are 95% perturbation-seed intervals conditional on the reused patients.

Absolute response relative to between-window variation. Perfect feature ranking requires no minimum discrepancy magnitude. Median disjoint-window raw KL is 1.103 unperturbed and 1.138 after intervention. MC-only comparisons isolate numerical variation on fixed patients; disjoint comparisons additionally involve patient composition. This is not an additive variance decomposition or an estimate of training variability.

For method a and matched window b, we compute

$$z^{(a)}_{\Delta,b} = \frac{T_a(R_b, T_b^{\text{timing}}) - T_a(R_b, T_b^{\text{unperturbed}})}{\text{SD}_\pi\{T_a(\pi[R_b, T_b^{\text{unperturbed}}])\}}.$$

The denominator is the sample SD of 199 unperturbed encounter permutations. Each statistic has its own scale; z_Δ is a descriptive standardized change, not a normal-theory test. Mean changes are 0.0010 for raw KL, 0.0008 for raw MMD and 0.0050 for raw C2ST. Their 5th–95th percentile ranges are respectively $[-0.024, 0.050]$, $[-0.082, 0.074]$ and $[-0.083, 0.107]$. Temporal-mask C2ST has a larger mean change of 0.101. Centering by each condition's permutation mean gives the same qualitative result (Appendix A.6). The unchanged width alarms therefore accompany small continuous responses relative to null variability, rather than concealing a large shift just below threshold.

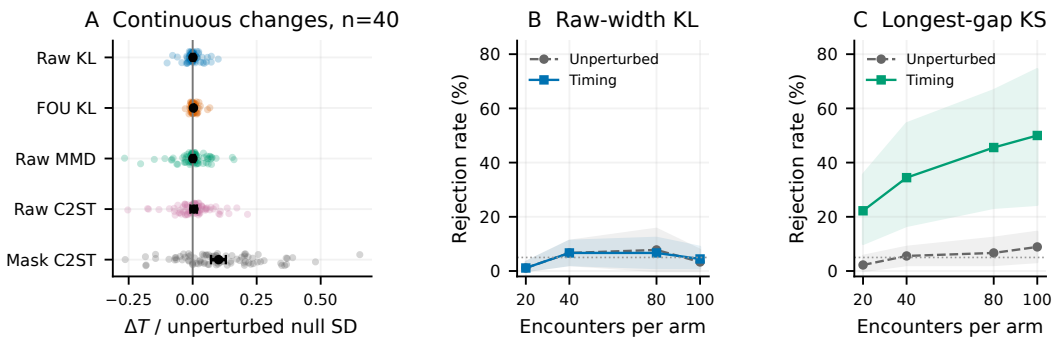


Figure 3: Continuous response and window-size boundary. A: standardized changes at $n = 40$, with 95% perturbation-seed-block intervals. B–C: raw-KL and longest-gap rejection rates for 90 matched pairs per size; bands resample the ten perturbation seeds. The $n = 40$ windows are unchanged, smaller/larger windows are nested subsets/extensions, and each arm is disjoint within a comparison. Patients recur across windows. All curves condition on the same $S = 10$ representation bank.

Is the timing signal detectable in these windows? A direct manipulation check summarizes each scope variable by its longest consecutive missing run divided by valid length, including boundary runs. We fixed this summary and a maximum marginal-KS test before computing the follow-up scores. On the same encounter pairs and 199 permutation assignments, its rejection count rises from 5/90 to 31/90, with 27 new alarms and one lost alarm (Table 2). This demonstrates a detectable temporal signal for a targeted mask summary at $n = 40$; it does not imply that an arbitrary mask learner must detect it or that widths intrinsically discard all timing information.

Table 2: Matched alarm transitions, $n = 40$, $S = 10$, nominal 5%. Every row uses the same 90 window pairs; $0 \to 1$ and $1 \to 0$ count both transition directions. Zero observed transitions do not establish zero population effect.

Method	Unperturbed	Timing	$0 \to 1$	$1 \to 0$
Temporal-mask C2ST	8	7	0	1
Raw-width KL	6	6	0	0
FOU KL	7	7	0	0
Raw-width MMD	4	4	0	0
Raw-width C2ST	2	2	0	0
Input/mask/output suite	4	4	0	0
Suite + raw KL	3	4	1	0
Suite + FOU KL	3	4	1	0
Longest-gap KS check	5	31	27	1

All tests share encounter permutations; C2STs refit fold-local scalers/classifiers for every assignment and complete suites are jointly recalibrated. Appendix A.3 retains pooled-randomization and separated-representation checks, which assess the statistical implementation rather than the real input intervention.

Does a larger window recover a response? With only raw KL and the fixed gap check, we extend the same comparisons to $n = 20, 40, 80, 100$ per arm, using 199 permutations. Raw-KL unperturbed/intervened counts are 1/1, 6/6, 7/6 and 3/4 out of 90. Gap-check counts are 2/20, 5/31, 6/41 and 8/45. Mean raw standardized changes remain between -0.002 and 0.007; the few alarm transitions at larger n are retained in Appendix A.8. Thus width detection does not recover over this tested range, whereas the targeted timing check benefits from larger windows. This does not establish a large-sample impossibility result.

5.2 Difficulty Response in Local Widths and the Monitor Input

Establishing the imputer's working state. We replay two existing 20% random pseudo-masks on all 86 source-validation encounters, then remove 50% of the remaining scope observations while keeping the original hidden queries fixed. Table 3 reports encounter–feature macro means. The imputer only modestly improves on source-mean imputation and has larger MAE than within-encounter linear interpolation; this limits the model scope of the negative findings. Nominal 90% intervals cover 85.6% of random queries. Context removal produces small but measurable widening on the same queries: $+0.0117$ ($[0.0095, 0.0139]$), versus a baseline 2.621. Mean missing-position width changes by $+0.0096$, while g_j increases by $+0.459$ as missingness grows. The model is not wholly unresponsive, but the aggregate increase alone would substantially overstate conditional widening.

Table 3: Fixed-query context removal on source validation, $S = 10$. Flow MAE uses the arithmetic sample mean $\widehat{x}_{tj} = S^{-1} \sum_s \tilde{x}_{tj}^{(s)}$ at the same hidden queries in both rows. Cells are averaged within encounter–feature and mask, then across masks and 532 eligible pairs from 86 encounters. Linear interpolation uses retained observations, nearest endpoints and source-mean fallback. Widths/MAE use source-normalized units; coverage is percent.

Context	Flow MAE	Mean MAE	Linear MAE	Cover.	Query width	Missing width	g^{raw}
Random 20% mask	0.680	0.700	0.328	85.6	2.621	2.644	1.741
Less context	0.696	0.700	0.378	84.8	2.633	2.654	2.200

Matched selections with different hidden positions. In a separate design, we fix encounter, variable, deletion count and fraction, but change which observed cells are hidden: easier, random or harder selections. Unlike Table 3, these arms do not fix the query cells; their contrast combines query selection and context changes. A source-only pseudo-masking difficulty scorer uses neither clinical labels nor width inputs. Actual reconstruction errors evaluate the selected contexts. Ten seeds, one/two variables and three severities yield 360 cached interventions across coronary and medical test cohorts. Only artificially hidden observations supply ground truth. The clinical heads in secondary analyses share this imputer and do not constitute architecture replication.

We pair harder-minus-easier effects within encounter–feature and condition, average repeated conditions within encounter–feature, then macro average across eligible pairs. The 584 source and 1380 medical pairs come from 87 and 223 encounters. Hidden-cell MAE increases by 0.195 and 0.192, while mean hidden-cell width changes by only $+0.008$ and -0.004 and coverage decreases by about eight percentage points (Figure 4). Cell-weighted summaries and feature heterogeneity remain in Appendix A.5.

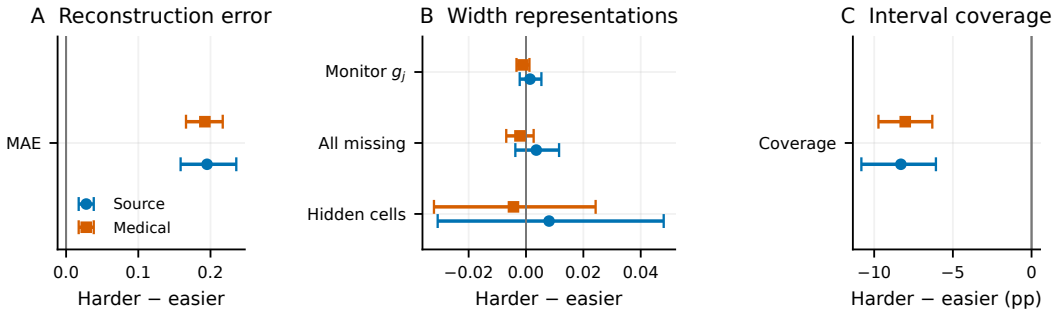


Figure 4: Harder-minus-easier macro effects, matched by encounter–feature, with equal deletion counts but different hidden positions. Error and coverage refer to each arm's selected hidden cells; width is shown for hidden cells, all missing positions, and the actual valid-time-averaged monitor input. Bars are 95% encounter-cluster bootstrap intervals after averaging conditions within encounter–feature. Width heterogeneity is preserved in the supplement; near-zero means do not establish identical distributions.

Connecting the local interval to the deployed summary. The same exports give $\Delta g_j^{\mathrm{raw}} = +0.0014$ in source and -0.0011 in medical, with intervals spanning zero (Table 4). We verify the cached monitor input and the identity $g_j^{\mathrm{raw}} = (1 - c_j)\bar{u}_j^{\mathrm{miss}}$ for every condition. Paired missing-position width changes are $+0.0036$ and -0.0021. Their Spearman associations with paired MAE changes are -0.015 and -0.016; for hidden-cell width the corresponding correlations are 0.002 and 0.009. Thus both local width and the actual monitor summary have weak average response and weak monotone association with increased error in these comparisons. This is a measured connection, not an inference from the hidden-cell mean alone.

Table 4: Harder-minus-easier effects for matched encounter–features with different hidden positions; 95% encounter-cluster intervals. Width and MAE use source-normalized units; coverage uses percentage points.

Endpoint	Source	Medical
Hidden-cell MAE	+0.195 [+0.159, +0.236]	+0.192 [+0.166, +0.217]
Hidden-cell width	+0.008 [-0.031, +0.048]	-0.004 [-0.032, +0.024]
Missing-position width	+0.004 [-0.004, +0.012]	-0.002 [-0.007, +0.003]
Monitor g_j^{raw}	+0.0014 [-0.0022, +0.0054]	-0.0011 [-0.0033, +0.0012]
Coverage (pp)	-8.304 [-10.814, -6.077]	-8.026 [-9.729, -6.307]

Monte Carlo sensitivity on a fixed subset. We rerun $S = 50$ on two affected variables, severity 0.4, all ten perturbation seeds and both cohorts, retaining exact easier/harder masks. The comparison uses the matching $S = 10$ subset, not the full grid (Table 5). This checks quantile-estimation sensitivity within the same model; it is not a second imputation architecture. At $S = 50$, paired MAE increases remain $+0.206$ and $+0.202$, and hidden-cell width changes remain small ($+0.008$ and -0.004). The actual monitor changes are -0.0003 and -0.0019; the medical interval $[-0.0036, -0.0003]$ indicates a small decrease, not exact invariance. Source coverage loss is smaller than at $S = 10$ (-4.21 versus -8.95 percentage points), while medical loss remains about -7.85 points. Thus the weak average width response persists, but coverage magnitudes depend on Monte Carlo size.

Table 5: Fixed-subset harder-minus-easier effects: two features, severity 0.4, all ten seeds. Each arm's mask is identical across S, while easier/harder masks differ; 578/1366 encounter–feature pairs from 87/223 source/medical encounters.

Cohort	S	ΔMAE	Δhidden width	Δcoverage (pp)	Δg^{raw}
Source	10	+0.207	+0.002	-8.95	+0.0007
Source	50	+0.206	+0.008	-4.21	-0.0003
Medical	10	+0.188	-0.009	-7.79	-0.0017
Medical	50	+0.202	-0.004	-7.85	-0.0019

Appendix A.7 gives the exact-mask replay checks and the non-nested relationship between the $S = 10$ and $S = 50$ Monte Carlo banks.

5.3 Incremental Error Ranking Depends on the Probability Baseline

The NIH-DEP probe predicts classification error of five frozen fuzzy-attention Transformers: $e_i = \mathbf{1}\{\mathbf{1}[p_i \geq \tau] \neq y_i\}$. The endpoint is recorded binary depression status at the target wave, using earlier waves as input. Here p_i is the saved uncalibrated probability and τ the existing validation-selected F1 threshold (0.18–0.24). Original clinical-head training, checkpoint selection and threshold selection exclude all test participants: train/validation/test contain 107,583/23,053/23,054 distinct people with zero pairwise overlap. The subsequent risk-model analysis remains retrospective on previously analyzed test data.

All 52,185 test windows are aligned by participant and wave and verified by replaying the clinical logits. The source fixes the single CES-D width scope. Auxiliary risk models use five participant-grouped folds, with fold-local standardization and $C = 1$ logistic regression. This available-cell width summary differs from the multi-target valid-time cohort representation.

The original baseline uses margin $m = -|p - \tau|$ and missingness. Our intermediate baseline adds predicted class $c = \mathbf{1}[p \geq \tau]$ and mc. The richer probability baseline uses p, p^2, c, pc and missingness. The identity

$$p = \tau + m - 2mc$$

shows that m, c, mc exactly recover p and its linear effect in the risk model; margin alone folds opposite sides of the threshold together.

The intermediate baseline reaches AUROC 0.81973, versus 0.60482 for margin and missingness. The richer baseline changes AUROC by -0.00001; raw-width addition to the intermediate baseline gives $+0.00014$, both with intervals spanning zero. Quadratic features are unnecessary for this aggregate finding; L2 regularization remains parameterization-dependent.

Table 6: Retrospective NIH-DEP error prediction, mean across five heads. All rows share participant folds, fold-local scaling and logistic risk models. Feature sets are defined above; AUPRC is average precision.

Risk features	Error AUROC	Error AUPRC	Error NLL
Margin	0.61379	0.34266	0.52713
Margin + missingness	0.60482	0.33003	0.52670
Margin + missingness + raw	0.70146	0.37195	0.49157
Margin + missingness + FOU	0.70218	0.37201	0.49084
Intermediate	0.81973	0.59962	0.39715
Intermediate + raw	0.81987	0.59914	0.39705
Intermediate + FOU	0.81987	0.59903	0.39705
Rich probability	0.81972	0.59914	0.39714
Rich probability + raw	0.81985	0.59875	0.39704
Rich probability + FOU	0.81985	0.59866	0.39704

Against the original margin and missingness baseline, raw and FOU add about 0.097 AUROC. Against the richer baseline (0.8197), their increments are only $+0.00013$ and $+0.00014$, with both 95% intervals $[-0.00021, 0.00047]$. Intervals jointly resample participants and head seeds while fixing the out-of-fold risks. The result establishes sensitivity to baseline specification; it does not isolate a causal effect of probability direction alone.

6 Discussion and Conclusion

At $S = 10$ and fixed patient count, independent randomness alone removes near-perfect localization: MC-only discrepancies resemble the intervention comparisons. This locates a numerical noise-cancellation boundary before changing patient mix; paired localization alone cannot establish independent-window monitoring ability. Larger windows improve the targeted mask diagnostic but produce no sustained raw-width alarm response over $n = 20$–100.

The imputer performs worse than interpolation and widens only slightly after context removal, restricting generalization. Ranking requires no minimum discrepancy magnitude, so noise cancellation can make small responses conspicuous. Statistical fixtures and input manipulation checks cannot replace this imputer-quality assessment.

Table 1 makes the protocol reusable: assess imputer quality and conditional response; vary patient and random-stream coupling separately; retain matched unperturbed windows and continuous scales; and preserve baseline probability information. Another imputer may produce different outcomes under these same comparisons.

Timing and difficulty findings remain conditional on one coronary-trained model and retrospective PN2012 cohorts. Three MC banks do not capture training uncertainty; the difficulty $S = 50$ check does not establish timing robustness at larger S. Source validation was previously exposed, and NIH-DEP is not cross-architecture replication. The 223-patient pool permits at most 111 disjoint encounters per arm. Distributional equivalence, larger-window power and prospective clinical action value remain unresolved. Each use therefore requires separate validation, alongside absolute imputation quality.

AI USE STATEMENT

Generative AI was used to assist with grammar and language checking during manuscript preparation, as well as to review code for potential bugs and implementation issues. All suggested changes were independently reviewed and verified by the authors, who remain fully responsible for the final scientific content, code, results, and disclosure.

REPRODUCIBILITY STATEMENT

Executable statistical replay. `pn2012_reproduction.zip` supplies frozen PN2012 representations for 30 timing conditions, unperturbed banks, window assignments, scoring functions and CPU replay. Direct record identifiers and clinical labels are omitted; row indices preserve disjointness. After installing the pinned dependencies, `python replay.py -check-saved` recomputes ranks and rates from every saved matrix. `python replay.py -phase fresh -workers 12` refits the domain classifiers and recomputes all 199 permutations per window. The `sanity`, `timing` and `anchors` phases retain pooled-null/positive checks and earlier window controls; `-smoke` checks representative draws. This reproduces statistics conditional on frozen imputation draws, not model training or raw-data export.

Frozen-checkpoint export and current core results. `upstream_export.zip` includes the PN2012 imputer weights, normalizer, membership quantiles, loading code, public-record split lists, availability lookups and timing specifications. Given PhysioNet Challenge 2012 set-a records, `python export.py -records /path/to/set-a -device cuda:0 -timing -out export` rebuilds the original banks. Parsing matches all 577 coronary and 1481 medical records and valid lengths. Hourly bins retain the last nonnegative value; padding is excluded. CPU export uses a different random stream. This is frozen-checkpoint export, not retraining.

The new `boundary_reproduction.zip` supplies anonymous public-data exports, frozen scoring/plotting code and permutation matrices for the current localization, window-detection and imputer-state analyses. Its README and Appendix A.9 give one replay entry point and its scope. `python replay.py -out replay` verifies hashes, recomputes localization/quality from arrays and continuous changes/window curves from complete saved scores, and regenerates matching figures/tables. `-rescore-smoke` or `-rescore-all` also recomputes representative or all raw-KL/gap permutation tests. This package does not regenerate the upstream imputation draws.

Additional evidence and limits. `bridge_evidence.zip` provides protocols, analysis/plotting code, continuous responses, localization, paired public-data summaries and $S = 50$ difficulty results with source hashes; source snapshots document that workflow. Historical panels remain in a separate PDF. A broader software release is planned after publication. NIH-DEP records, identifiers, out-of-fold predictions and clinical checkpoints are excluded, so its risk probe cannot be independently refitted from the anonymous supplement. These packages neither reproduce the entire private-data study nor establish independent clinical replication.

ETHICS STATEMENT

This study retrospectively analyzes public PhysioNet research data and a private dataset from an NIH-funded project, referred to as NIH-DEP during review. The authors confirm that the private-data analysis is lawful and authorized for research. To preserve double-blind review, identifying project and access details are withheld; the data-access procedure will be described after publication. Participant-level NIH-DEP records and identifiers are not included in the supplement. The analyses evaluate monitoring behavior and do not support clinical diagnosis or intervention without further validation.

REFERENCES

Wei Cao, Dong Wang, Jian Li, Hao Zhou, Lei Li, and Yitan Li. BRITS: Bidirectional recurrent imputation for time series. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 31, 2018.

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

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.

Vincent Fortuin, Dmitry Baranchuk, Gunnar Raetsch, and Stephan Mandt. GP-VAE: Deep probabilistic time series imputation. In *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics*, volume 108 of *Proceedings of Machine Learning Research*, pp. 1651–1661, 2020. URL https://proceedings.mlr.press/v108/fortuin20a.html.

João Gama, Indrė Žliobaitė, Albert Bifet, Mykola Pechenizkiy, and Abdelhamid Bouchachia. A survey on concept drift adaptation. *ACM Computing Surveys*, 46(4):44:1–44:37, 2014. doi: 10.1145/2523813.

Saurabh Garg, Sivaraman Balakrishnan, Zachary C. Lipton, Behnam Neyshabur, and Hanie Sedghi. Leveraging unlabeled data to predict out-of-distribution performance. In *International Conference on Learning Representations (ICLR)*, 2022.

Antonio A. Ginart, Martin J. Zhang, and James Zou. MLDemon: Deployment monitoring for machine learning systems. In *Proceedings of the 25th International Conference on Artificial Intelligence and Statistics (AISTATS)*, volume 151 of *PMLR*, pp. 3962–3997, 2022.

Arthur Gretton, Karsten M. Borgwardt, Malte J. Rasch, Bernhard Schölkopf, and Alexander Smola. A kernel two-sample test. *Journal of Machine Learning Research*, 13:723–773, 2012.

Oleg Ivanov, Michael Figurnov, and Dmitry Vetrov. Variational autoencoder with arbitrary conditioning. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=SyxtJh0qYm.

Pang Wei Koh, Shiori Sagawa, Henrik Marklund, Sang Michael Xie, Marvin Zhang, Akshay Balsubramani, Weihua Hu, Michihiro Yasunaga, Richard Lanas Phillips, Irena Gao, Tony Lee, Etienne David, Ian Stavness, Wei Guo, Berton A. Earnshaw, Imran S. Haque, Sara Beery, Jure Leskovec, Anshul Kundaje, Emma Pierson, Sergey Levine, Chelsea Finn, and Percy Liang. WILDS: A benchmark of in-the-wild distribution shifts. In *Proceedings of the 38th International Conference on Machine Learning (ICML)*, volume 139 of *PMLR*, pp. 5637–5664, 2021.

Sean Kulinski, Saurabh Bagchi, and David I. Inouye. Feature shift detection: Localizing which features have shifted via conditional distribution tests. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 33, 2020.

Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. Simple and scalable predictive uncertainty estimation using deep ensembles. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 30, 2017.

Zachary C. Lipton, Yu-Xiang Wang, and Alexander Smola. Detecting and correcting for label shift with black box predictors. In *Proceedings of the 35th International Conference on Machine Learning (ICML)*, volume 80 of *PMLR*, pp. 3122–3130, 2018.

Roderick J. A. Little and Donald B. Rubin. *Statistical Analysis with Missing Data*. Wiley, 3rd edition, 2019. doi: 10.1002/9781119482260.

David Lopez-Paz and Maxime Oquab. Revisiting classifier two-sample tests. In *International Conference on Learning Representations (ICLR)*, 2017.

Scott M. Lundberg and Su-In Lee. A unified approach to interpreting model predictions. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 30, 2017.

Pierre-Alexandre Mattei and Jes Frellsen. MIWAE: Deep generative modelling and imputation of incomplete data sets. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 4413–4423, 2019. URL https://proceedings.mlr.press/v97/mattei19a.html.

Jose G. Moreno-Torres, Troy Raeder, Rocío Alaiz-Rodríguez, Nitesh V. Chawla, and Francisco Herrera. A unifying view on dataset shift in classification. *Pattern Recognition*, 45(1):521–530, 2012. doi: 10.1016/j.patcog.2011.06.019.

Carlos Mougan, Klaus Broelemann, Gjergji Kasneci, Thanassis Tiropanis, and Steffen Staab. Explanation shift: How did the distribution shift impact the model? *Transactions on Machine Learning Research*, 2025. URL https://openreview.net/forum?id=MO1slfU9xy.

Yaniv Ovadia, Emily Fertig, Jie Ren, Zachary Nado, D. Sculley, Sebastian Nowozin, Joshua V. Dillon, Balaji Lakshminarayanan, and Jasper Snoek. Can you trust your model's uncertainty? Evaluating predictive uncertainty under dataset shift. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 32, 2019.

Joaquin Quiñonero-Candela, Masashi Sugiyama, Anton Schwaighofer, and Neil D. Lawrence (eds.). *Dataset Shift in Machine Learning*. MIT Press, Cambridge, MA, 2009. ISBN 9780262170055.

Stephan Rabanser, Stephan Günnemann, and Zachary C. Lipton. Failing loudly: An empirical study of methods for detecting dataset shift. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 32, 2019.

Donald B. Rubin. Inference and missing data. *Biometrika*, 63(3):581–592, 1976. doi: 10.1093/biomet/63.3.581.

Hidetoshi Shimodaira. Improving predictive inference under covariate shift by weighting the log-likelihood function. *Journal of Statistical Planning and Inference*, 90(2):227–244, 2000. doi: 10.1016/S0378-3758(00)00115-4.

Bernard W. Silverman. *Density Estimation for Statistics and Data Analysis*, volume 26 of *Monographs on Statistics and Applied Probability*. Chapman and Hall, London, 1986. ISBN 0412246201.

Adarsh Subbaswamy, Peter Schulam, and Suchi Saria. Preventing failures due to dataset shift: Learning predictive models that transport. In *Proceedings of the 22nd International Conference on Artificial Intelligence and Statistics (AISTATS)*, volume 89 of *PMLR*, pp. 3118–3127, 2019.

Masashi Sugiyama, Matthias Krauledat, and Klaus-Robert Müller. Covariate shift adaptation by importance weighted cross validation. *Journal of Machine Learning Research*, 8:985–1005, 2007.

Yusuke Tashiro, Jiaming Song, Yang Song, and Stefano Ermon. CSDI: Conditional score-based diffusion models for probabilistic time series imputation. In *Advances in Neural Information Processing Systems*, volume 34, pp. 24804–24816, 2021. URL https://papers.nips.cc/paper_files/paper/2021/hash/cfe8504bda37b575c70ee1a8276f3486-Abstract.html.

Ryan J. Tibshirani, Rina Foygel Barber, Emmanuel Candès, and Aaditya Ramdas. Conformal prediction under covariate shift. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 32, 2019.

Alexandre B. Tsybakov. *Introduction to Nonparametric Estimation*. Springer Series in Statistics. Springer, New York, 2009. ISBN 9780387790510. doi: 10.1007/b13794.

Stef van Buuren and Karin Groothuis-Oudshoorn. mice: Multivariate imputation by chained equations in R. *Journal of Statistical Software*, 45(3):1–67, 2011.

Jinsung Yoon, James Jordon, and Mihaela van der Schaar. GAIN: Missing data imputation using generative adversarial nets. In *Proceedings of the 35th International Conference on Machine Learning (ICML)*, volume 80 of *PMLR*, pp. 5689–5698, 2018.

Margaux Zaffran, Aymeric Dieuleveut, Julie Josse, and Yaniv Romano. Conformal prediction with missing values. In *Proceedings of the 40th International Conference on Machine Learning (ICML)*, volume 202 of *PMLR*, pp. 40578–40604, 2023.

A Additional Evaluation Protocols and Results

A.1 Frozen Resources and Controlled Localization

The main controlled PN2012 experiments monitor eight numerical coordinates: HR, DiasABP, MAP, RespRate, Temp, BUN, Creatinine and WBC. The coronary-trained imputer, normalizer and membership functions are fixed across the coronary and medical cohorts. The coronary training/validation/test pools contain 403/86/88 encounters; the medical validation/test pools contain 222/223. The five-component suite uses a source-trained mortality logistic head. Later conditional-deletion and recovery analyses also reuse three existing clinical Transformer heads (training seeds 42–44), with actual predictions recomputed after each intervention. These are different heads from the NIH-DEP classification-error probe.

Conditional imputation in the PN2012 case study. The PN2012 imputer is trained on source-normalized coronary training sequences by conditional flow matching. Observed cells are randomly partitioned into conditioning, pseudo-missing targets, and remaining observed cells, with nominal probabilities 0.70/0.15/0.15. Binary and non-scope features are moved into conditioning. For target mask M, Gaussian noise z, and time t sampled uniformly from $[10^{-4}, 1-10^{-4}]$, the linear bridge sets $x_t = (1-t)z + tx$ on M and retains the data elsewhere. The masked training loss is

$$\mathcal{L}(\theta) = \mathbb{E}\left[\frac{\sum_{u,j} M_{uj}\{v_\theta(x_t, t, \mathcal{C})_{uj} - (x_{uj} - z_{uj})\}^2}{\max(1, \sum_{u,j} M_{uj})}\right].$$

Here $\mathcal{C}$ comprises conditioning/target/remaining-observed/availability masks, positional and horizon embeddings, and valid sequence lengths. The velocity Transformer receives the complete 41-feature sequence and four masks; the loss and generated outputs are restricted to the eight numerical scope features. Natural missing cells supply no reconstruction target. At inference, Gaussian initializations are integrated on missing, available, in-scope positions while observed cells are projected to their input values. Source normalization and model parameters remain fixed. These generated quantiles carry no coverage guarantee and the model is not trained to predict downstream clinical errors.

The alternative FOU representation uses $R = 5$ Gaussian membership functions,

$$\mu_{jr}(z) = \exp\{-\tfrac{1}{2}((z - c_{jr})/s_j)^2\}.$$

$$g_j^{\mathrm{FOU}}(x) = \frac{1}{L_x R}\sum_{t=1}^{L_x}\sum_{r=1}^{R}\left[\max_{z\in[\ell_{tj},u_{tj}]}\mu_{jr}(z) - \min_{z\in[\ell_{tj},u_{tj}]}\mu_{jr}(z)\right].$$

FOU is a model-derived transformed width, not a predictive interval with a numerical coverage guarantee. Its source-fitted membership parameters remain frozen; numerical settings and interval extrema are given in Appendix A.

Unless varied explicitly, $S = 10$ without validation-derived interval expansion. Historical cohort panels retain their FOU representation; controlled experiments compare both representations.

The flow imputer uses hidden size 128, four attention heads, three layers, feed-forward size 512, dropout 0.2, a linear flow schedule and 50 Euler steps. The width estimator uses Scott bandwidth separately for each sample, floored at the larger of 10^{-10} and 10^{-3} times the pooled range. The common 200-point grid extends past the pooled range by the larger of 5% of the range and five times the larger bandwidth. KDE values receive additive 10^{-8} regularization before grid normalization.

The implementation evaluates regularized Gaussian KDEs on a shared 200-point grid; bandwidth and boundary rules are in Appendix A. With additive regularization $\varepsilon = 10^{-8}$,

$$\pi_{jv}^S = \frac{\hat{p}_j^S(z_{jv}) + \varepsilon}{\sum_w(\hat{p}_j^S(z_{jw}) + \varepsilon)}, \quad \pi_{jv}^T = \frac{\hat{p}_j^T(z_{jv}) + \varepsilon}{\sum_w(\hat{p}_j^T(z_{jw}) + \varepsilon)}, \quad D_2 = \sum_{j\in\mathcal{A}}\underbrace{\sum_v \pi_{jv}^S \log\frac{\pi_{jv}^S}{\pi_{jv}^T}}_{d_j}.$$

The centers c_{jr} are source-training observed-value quantiles at levels $0.05, 0.275, 0.5, 0.725, 0.95$. The shared feature scale s_j is their mean adjacent spacing, floored at 10^{-4}. For a Gaussian, the maximum is one if its center lies inside the interval, and otherwise is the larger endpoint membership;

the minimum is the smaller endpoint membership. These source-initialized functions are frozen in the reported width exports. FOU is a model-derived transformed width, not a predictive interval with a numerical coverage guarantee.

The archived paired-localization benchmark uses all 223 medical test encounters. There are ten perturbation seeds, three severity levels (0.2, 0.4, 0.6), and one, two or four affected variables per intervention family. Coverage dropout deletes observations; matched-coverage timing moves the same observed values in time without changing their count or multiset. Feature AUROC and tie-aware Recall@k use the known affected variables as labels. Uncertainty across perturbation seeds is conditional on the same encounters, imputer and choices of interventions. The matched-coverage negative controls are uninformative by construction: neither observation counts nor marginal observed-value summaries change.

The historical width-surrogate SHAP adaptation has feature AUROC 0.836 and Recall@k 0.703 on the matched-coverage timing benchmark. Its surrogate distinguishes cohorts using widths (validation AUROC 0.620); it is not an explanation of the frozen mortality predictor and cannot establish superiority over clinical-predictor Explanation Shift. The expanded triage controls instead recompute the actual clinical head's logit changes after source-defined column occlusion. These diagnostics are still not a complete implementation of Explanation Shift.

A.2 Alarm Protocols and Complete Frozen-Threshold Results

The corrected source-frozen multi-target natural-cohort audit contains 26 cross-cohort pairs, with 8/8/6 width coordinates for NIH-DEP/PN2012/PN2019. Whole-cluster and cluster-mean permutation p-values range from 0.0005 to 0.006; 24 pairs reach the $1/2000$ resolution floor. Pair-specific row and cluster counts are included in the aggregate evidence. These are detected cohort discrepancies, not incremental utility over the strong raw/imputed baselines in the historical descriptive comparison.

Its Monte Carlo p-value is $(1+\sum_{b=1}^{B}\mathbf{1}\{D_2^{(b)} \geq D_2^{\mathrm{obs}}\})/(B+1)$ with $B = 1999$, exchanging whole participants/encounters under the tested null. Cluster bootstrap intervals resample these units within each cohort. Cluster-mean sensitivity first averages repeated rows within each person and therefore changes their weighting; repeated rows are not independent replications. These corrected tests are not attached to historical single-target scores with different scopes and references.

The historical same-cohort FOU checks average widths within each participant/encounter and use 500 randomized splits of each held-out cluster pool and 1999 shared null draws. Their numerical results are retained in Table 7. The NIH-DEP/PN2012/PN2019 scopes have 8/8/6 coordinates, respectively. Reusing clusters makes these conditional randomization summaries, not independent deployment replications. NIH-DEP exceeds the nominal error level; increasing the number of permutations alone does not repair this discrepancy.

Table 7: Historical same-cohort FOU null checks, separate from independent-window and suite protocols.

Dataset / source	Clusters	FPR at 5%	FPR at 1%
NIH-DEP / WHI-OS	25,128	0.072	0.020
PN2012 / coronary	174	0.054	0.004
PN2019 / hospital A	6,099	0.054	0.010

In the five-component frozen-threshold suite, source-only monitoring divides source validation into 43 reference and 43 calibration encounters. Target warm-up uses 111 medical-validation encounters for each role. Test encounters (88 source or 223 target) are disjoint from validation. Component scales are fixed before complete-suite threshold calibration; the baseline and each augmented suite have separate thresholds at the same nominal 5% total budget. Calibration and evaluation draw many windows from their respective finite pools. In particular, two calibration draws can reuse encounters; they are not independent samples merely because their random seeds differ. Table 8 retains the high realized rejection rates. Cross-cohort unperturbed target windows already differ from the source, so their rejection rate is not a false-positive rate.

Table 8: Frozen-threshold suite results. Power averages the archived observation-dropout, matched-coverage-burst and value-offset intervention windows. Unpert.: unperturbed rejection rate; Interven.: intervention rejection rate. Fixed-pool comparisons are not assumed to be exchangeable null draws.

Protocol	n	Suite		Suite + raw		Suite + FOU	
		Unpert.	Interven.	Unpert.	Interven.	Unpert.	Interven.
Source	20	0.064	0.253	0.064	0.270	0.064	0.270
Source	40	0.236	0.381	0.236	0.406	0.252	0.407
Cross-cohort	20	1.000	1.000	1.000	1.000	1.000	1.000
Cross-cohort	40	1.000	1.000	1.000	1.000	1.000	1.000
Target warm-up	20	0.104	0.296	0.116	0.336	0.100	0.318
Target warm-up	40	0.236	0.550	0.100	0.446	0.140	0.458
Target warm-up	80	0.896	0.997	0.884	0.996	0.876	0.996

Joint permutation comparison. The additional protocol uses independent imputation streams across validation and test, with shared random numbers only within each baseline/intervention pair. Each window has 40 reference and 40 target encounters. Pool the 80 encounter representations and apply the same 199 whole-encounter label permutations to all seven statistics: the five original suite components, raw-width D_2, and FOU D_2. All C2ST models are refitted for each assignment. Including the observed assignment gives $M = 200$ scores T_{bc} per component. For every assignment, define

$$p_{bc} = \frac{1}{M}\sum_{a=1}^{M} \mathbf{1}\{T_{ac} \geq T_{bc}\}, \qquad m_b = \min_{c \in \mathcal{C}} p_{bc}, \qquad p_{\mathcal{C}} = \frac{1}{M}\sum_{a=1}^{M} \mathbf{1}\{m_a \leq m_{\text{obs}}\}.$$

The complete suite is thus recalibrated, including dependence between its components and conservative treatment of ties. All suites use the same encounter draws and permutations. This is a separate protocol rather than a reinterpretation of the failed frozen-threshold calibration. Exchangeability of encounters and their processed representations is still needed for a null interpretation.

We fix 500 unperturbed draws and three draws for each of the 60 medical deletion-burst conditions with two affected variables: ten seeds, three severities, and two source-sensitivity groups. The resulting 180 alternatives change observation coverage, unlike the earlier timing benchmark. Reference data are unlabeled medical validation encounters; this is not source-only deployment. Conditional power intervals resample the ten perturbation seeds and hold clinical pools, fitted models and window draws fixed. A four-window engineering pilot is excluded from the final rates.

Table 9: Joint-permutation comparison on common windows. Unperturbed rejection is conditional on the fixed 222/223 reference/evaluation pools. Intervention rejection concerns coverage-changing deletion, not clinical failure.

Monitor	Unperturbed rejection (%)	Intervention rejection (%)
Input/mask/output suite	7.8	6.7
Suite + raw	5.0	9.4
Suite + FOU	4.8	9.4
Raw width only	3.6	10.0
FOU only	4.2	12.2
Observed-value KS	4.4	12.2
Temporal-mask C2ST	3.0	2.8
Imputed-summary MMD	20.6	0.6
Imputed-summary C2ST	6.2	0.0
Prediction-probability KS	5.0	0.0

A.3 Pooled Randomization, Positive Controls and Matched Timing

The frozen protocol distinguishes exchangeable random assignment from a comparison of two fixed clinical pools. Pooled nulls draw 80 distinct encounters from the 223-person medical-test pool, randomly assigning 40 per side. We perform 500 independently randomized draws, each

with the same 199 whole-encounter permutations for all components. Their rejection-rate intervals describe Monte Carlo sampling conditional on this empirical pool. Reusing the pool does not make those random draws additional independent patients, but neither does it inherently invalidate label randomization. Fixed-pool unperturbed comparisons instead draw 40 encounters from each of the medical validation/test pools (222/223); these are unperturbed comparisons whose conditional exchangeability is not assumed.

The positive control is explicitly a representation-level fixture. Fifty seeds generate separated groups: raw-width coordinates are log-normal with log-means zero/two and SD 0.1; FOU and probability blocks are uniform on [0.1,0.2]/[0.8,0.9]; other statistic-input blocks have Gaussian means zero/three and unit SD. All blocks use the clinical representations' dimensions. The synthetic RFF vectors test the MMD mean-distance calculation rather than a clinical kernel export. Every classifier and the complete joint-rank calculation are rerun. This is an implementation check, not a clinically plausible shift or imputation-model validation.

For timing power, we select all 30 $k = 2$ matched-coverage conditions from the paired-localization result by their saved feature selection, severity and intervention seed. The same observed-value multisets and observation counts are checked after reconstruction of each intervention. We export all comparator representations together with $S = 10$. Validation and test use different imputation random streams (991003/991004); baseline and intervention within test share the latter stream. A full unperturbed export is replayed against the previously frozen bank before reuse. The common-bank localization in Figure 2 now uses these exact representations. They reproduce the archived input interventions, but not the archived Monte Carlo width arrays.

Matched unperturbed controls and window sampling. An initial diagnostic fixed three window pairs across all 30 conditions. In the within-test comparison, temporal-mask C2ST, imputed C2ST and the baseline suite each rejected 30/90 altered windows, all from the same one of the three pairs. Recomputing the six unique unperturbed reference/target comparisons with the same permutations showed that this pair already rejected for all three methods. Thus their 33.3% altered-window rejection did not establish a response to timing. Suite + raw was unperturbed-negative on all three pairs and rejected 11/90 altered windows; its lower overall rejection cannot by itself be interpreted as poorer intervention detection. Complete initial results and every score matrix remain in the supplement.

We then froze a follow-up without changing any method, threshold, representation or model. For each condition and each of three draws, a new random set of 80 medical-test encounters is divided 40/40, and a new medical-validation reference of 40 is drawn. This yields 90 distinct window pairs per reference setting. Every pair is evaluated both unperturbed and timing-altered, with exactly the same encounters and 199 permutations. The total is 360 score matrices, using RNG seed 2026092602 and permutation seed $985000+$ pair index. All 30 conditions and all methods are retained. This is a disclosed follow-up on previously analyzed data, not independent clinical confirmation.

The reported follow-up intervals use 2000 resamples of ten perturbation-seed blocks, then the three window draws within each severity of each selected block. Clinical pools, imputation draws, models and timing conditions remain fixed. These conditional intervals reflect the observed between-window variation; they do not quantify new-population uncertainty. The within-test comparison provides a matched unperturbed control, while the fixed validation/test comparison still need not compare equal empirical pools. No threshold is adjusted to the realized null rates. When every observed paired alarm difference is zero, resampling produces a degenerate $[0, 0]$ interval; it is retained in the numerical archive but is not an equivalence test or evidence of zero population effect.

A separate post-hoc resolution diagnostic fixes all 500 completed pooled-null score banks and counts which of their 200 rows could reject under conservative ties. Mean eligible fractions are 4.454% for the suite and 3.593% for suite + raw. These finite-bank fractions explain one source of conservatism; they are not population-FPR estimates and do not change any decisions.

Table 10: Complete protocol checks and fresh-window timing comparison (percent). Implementation checks use 500 pooled nulls, 50 positive fixtures and 500 fixed-pool unperturbed draws. Fresh pooled/fixed columns each pair 90 unperturbed (Base) and altered (Timing) windows. Fixed-pool unperturbed rejection need not estimate a same-distribution FPR.

	Implementation checks			Fresh pooled		Fresh fixed	
Method	Pooled	Positive	Fixed	Base	Timing	Base	Timing
Value-summary KS	5.8	100.0	2.6	4.4	4.4	1.1	1.1
Temporal-mask C2ST	4.8	100.0	6.4	8.9	7.8	7.8	7.8
Imputed MMD	4.8	100.0	18.0	3.3	3.3	27.8	25.6
Imputed C2ST	6.2	100.0	5.4	4.4	4.4	12.2	12.2
Prediction KS	5.2	100.0	5.8	7.8	7.8	3.3	6.7
Raw-width KL	5.6	100.0	3.2	6.7	6.7	5.6	5.6
FOU KL	5.0	100.0	5.0	7.8	7.8	5.6	5.6
Raw-width MMD	5.6	100.0	4.6	4.4	4.4	7.8	7.8
Raw-width C2ST	5.2	100.0	4.2	2.2	2.2	5.6	5.6
Input/mask nonlinear C2ST	3.6	100.0	5.8	10.0	10.0	10.0	12.2
Suite	3.0	100.0	9.4	4.4	4.4	13.3	13.3
Suite + raw KL	2.4	100.0	6.4	3.3	4.4	7.8	8.9
Suite + FOU KL	2.6	100.0	6.4	3.3	4.4	8.9	8.9
Expanded suite	2.2	100.0	8.6	5.6	4.4	10.0	10.0
Expanded suite + raw KL	1.8	100.0	5.8	1.1	4.4	8.9	10.0

A.4 Implementation Specifications

Table 11: Monitoring implementation. Transform parameters learned on source training are frozen; group-dependent operations are recomputed for every permutation.

Component	Executed definition
Value summaries	Per variable: mean, population SD, and four means over relative-time bins [0,.25), [.25,.5), [.5,.75), [.75,1). Raw summaries use observed cells; imputed summaries use interval centers. Empty-bin means are zero; padding is excluded. Source normalization is frozen.
Temporal mask	The six analogous observation-mask summaries plus the flattened full valid-time mask. Nonlinear input/mask C2ST concatenates these masks with the six raw-value summaries, not the complete raw-value sequence.
KS	Maximum two-sample KS statistic across value-summary columns; output KS operates on the frozen clinical prediction probabilities. The joint permutation, not a marginal KS p-value, defines the suite alarm.
Linear C2ST	Three-fold stratified, shuffled encounter split; fold-local StandardScaler and L2 logistic regression, $C = 1$, liblinear, 1000 iterations, seed 20260924. Pooled out-of-fold domain AUROC is the statistic.
Nonlinear C2ST	Same three-fold structure; fold-local StandardScaler and RBF SVC, $C = 1$, gamma=scale, seed 730013. Out-of-fold decision scores define domain AUROC. Every assignment refits all three models.
MMD approximation	Squared Euclidean difference of mean random Fourier features. Imputed summaries: 512 RFFs, source-trained StandardScaler, gamma=$1/(6D)$, seed 20260924. Raw widths: 512 RFFs, source-trained scaler, gamma=$1/8$, seed 20260926. No group-specific fit enters this frozen feature map.
Width KL	Eight width coordinates; separate Scott bandwidths, pooled-range bandwidth floor, shared 200-point grid and 10^{-8} density floor as specified above. Bandwidths, grid and degenerate-coordinate checks are recomputed for each assignment.
Joint suite	For every observed/permuted row, take conservative upper-tail ranks of all component scores, then the minimum component p within a fixed suite, then its conservative lower-tail rank across the same 200 rows. Original suite: value KS, temporal-mask C2ST, imputed MMD/C2ST and output KS. Expanded suite adds nonlinear input/mask C2ST.

Table 12: Conditional difficulty and secondary triage implementation; no target clinical labels enter feature construction or ranking.

Component	Executed definition
Difficulty training	Three independent 20% pseudo-masks of coronary training observations. Per-variable ExtraTreesRegressor: 64 trees, maximum depth 6, minimum leaf 16, seed $820019 + j$, target $\log(1 + \lvert\hat{x} - x\rvert)$, inverse encounter-frequency sample weights. Two separate source-validation pseudo-masks diagnose ranking without selecting a target model.
Difficulty inputs	Current-time observed values and masks excluding the queried cell; normalized time; existence, normalized distances and values of nearest observed same-variable neighbors on both sides; neighbor value difference; remaining same-variable observation fraction; local mask density. There are $2D + 10$ inputs. Within-history future neighbors are permitted by the retrospective imputation task; future clinical outcomes are excluded.
Matched deletion	For each encounter and selected variable, delete $\lfloor sn_{\mathrm{obs}} + U \rfloor$ observed cells, using the identical count across easier/harder/random arms. Order cells by predicted difficulty with fixed random ties, or random order. All retained values are unchanged.
Source sensitivity	For each scope variable, permute its twelve clinical-input summary columns together across source-validation encounters, ten times. Average the absolute probability change and divide by the maximum across variables. These weights multiply featurewise raw KS scores in the source-weighted diagnostic.
Head-specific explanation	Logistic: sum standardized linear-logit contributions for each variable. Transformer: original logit minus the logit after replacing that variable's values, centers, interval endpoints and observation masks by their source-mean/zero references, leaving availability fixed. Source-validation mean absolute changes define normalized head-specific weights. These are occlusion diagnostics, not Shapley values.
Frozen-threshold suite	Separate calibration-window draws define component empirical upper-tail probabilities and the 95th percentile (higher-quantile rule) of the complete suite's maximum negative log tail. Alarm requires strict exceedance. Finite-pool transfer is assessed empirically rather than assumed.
Fixed-budget suite	Component null medians and IQR/1.349 define standardized scores, with SD then unit-scale fallbacks. Rank the maximum within each fixed suite. Select $\lceil qN \rceil$ windows and average uniformly across boundary ties. Expected random event recall is $\lceil qN \rceil / N$ when any events exist; it is exactly 20% in the reported primary settings.

A.5 Paired Encounter–Feature Difficulty Summaries

We first average error, interval width and interval coverage over the artificially hidden cells of each encounter–feature in each condition. Harder and easier arms have identical deletion counts, so their differences are paired at this level. Repeated conditions are then averaged within encounter–feature, before macro averaging across those pairs. This gives 584 pairs from 87 source encounters and 1380 pairs from all 223 medical encounters; the excluded source encounter has no eligible paired hidden cells. The 2000 bootstrap draws resample whole encounters with all their feature pairs, holding intervention seeds, model and difficulty scorer fixed. Medians and featurewise effects guard against interpreting a pooled mean as a uniform response. No interval is a test of equality of width distributions.

Table 4 reports the primary paired macro effects; the table below resolves their heterogeneity by feature.

Table 13: Featurewise paired macro means. Positive and negative width changes coexist; these heterogeneous effects must not be replaced by a claim that every interval is insensitive to difficulty.

Cohort	Feature	Pairs	ΔMAE	Δwidth	Δcoverage (pp)
Source	BUN	87	+0.0842	-0.0438	-4.6939
Source	Creatinine	87	+0.0937	+0.0435	-4.2558
Source	DiasABP	62	+0.2681	-0.0506	-12.5235
Source	HR	86	+0.3517	+0.0392	-16.4097
Source	MAP	62	+0.3203	+0.0456	-11.7341
Source	RespRate	27	+0.1518	-0.0066	-4.6730
Source	Temp	86	+0.2213	+0.0555	-8.1140
Source	WBC	87	+0.0997	-0.0336	-3.8156
Medical	BUN	218	+0.0640	+0.0157	-0.2008
Medical	Creatinine	219	+0.0120	-0.0263	-2.5006
Medical	DiasABP	99	+0.4492	+0.0098	-22.3600
Medical	HR	217	+0.4251	-0.0071	-16.6317
Medical	MAP	97	+0.6314	-0.0444	-30.6625
Medical	RespRate	94	+0.1847	+0.0075	-7.3672
Medical	Temp	217	+0.1767	+0.0070	-7.3294
Medical	WBC	219	-0.0228	-0.0047	+2.7181

A.6 Common-Bank Representation Bridge

The reviewer-motivated bridge is a separate exploratory analysis of the frozen 30-condition timing bank and 360 conditional-deletion exports. Its specification fixes endpoints before follow-up computation; it does not turn previously exposed PN2012 encounters into confirmation data. Every condition is retained. The shared-stream and disjoint-patient anchors in Figure 2 use exactly the baseline/altered width arrays used for independent-window detection. Appendix A.8 adds independent-stream baselines to separate patient pairing from Monte Carlo noise cancellation.

The three reference choices are: all 223 identical encounters before/after; the same 40 target encounters before/after in each fresh window; and the disjoint 40-encounter reference versus that altered target. The latter also uses unperturbed target counterparts. Across the 90 windows, unperturbed raw feature AUROC/Recall@2 are 0.448/0.217 when scored against the variables about to be intervened on, versus 0.447/0.217 after intervention. FOU gives 0.420/0.200 in both cases. This is feature ranking conditional on known intervention labels, not a diagnostic prediction on unknown clinical changes. The full-cohort raw AUROC has perturbation-seed interval $[0.992, 1.000]$; the disjoint-window interval is $[0.338, 0.556]$. These intervals condition on the clinical pool and common draw bank.

Continuous scores and transitions. For each matched window and every component, the archive retains both observed statistics, all 199 permutation scores, their means/SDs, the raw paired difference and its standardized version. We also calculate

$$z^{(a)}_{\Delta,\text{centered}} = \frac{(T^{(a)}_{\text{timing}} - \bar{T}^{(a)}_{\pi,\text{timing}}) - (T^{(a)}_{\text{base}} - \bar{T}^{(a)}_{\pi,\text{base}})}{s^{(a)}_{\pi,\text{base}}}.$$

The denominator remains the unperturbed sample SD ($\text{ddof} = 1$), so a changed null scale is not silently folded into the comparison. All 180 paired windows across the two reference settings have nonzero denominators for all ten components. Table 14 summarizes the within-test result; per-severity and fixed validation/test results are included in machine-readable form. Mean intervals resample ten seed blocks 2000 times after averaging severities and windows within seed, holding window assignments fixed. This differs from the earlier nested window resampling for alarm-rate intervals. Empirical quantiles describe the observed change distribution and are not uncertainty intervals for a population effect. Both alarm transition directions are retained for every component and complete suite.

Direct timing manipulation check. For each scope variable, scan valid time positions, resetting the missing-run length at an observed position and taking its maximum, including leading and

Table 14: Paired changes standardized by each component's unperturbed permutation SD. The 5–95% column is an empirical window range. Components have different raw units, so only their within-method standardized changes are juxtaposed.

Component	Mean z_Δ	Median	5–95%	Null-centered mean
fou_d2	+0.0031	+0.0003	[-0.008, +0.023]	+0.0020
imputed_joint_c2st	+0.0060	+0.0079	[-0.173, +0.160]	+0.0056
imputed_joint_mmd	+0.0045	+0.0040	[-0.074, +0.086]	+0.0019
input_mask_rbf_c2st	+0.0400	+0.0228	[-0.084, +0.236]	+0.0402
prediction_ks	+0.0280	+0.0000	[-0.677, +0.445]	+0.0324
raw_ks	+0.0207	+0.0000	[+0.000, +0.000]	+0.0228
raw_width_c2st	+0.0050	+0.0070	[-0.083, +0.107]	+0.0057
raw_width_d2	+0.0010	-0.0007	[-0.024, +0.050]	-0.0008
raw_width_mmd	+0.0008	+0.0011	[-0.082, +0.074]	-0.0002
temporal_mask_c2st	+0.1014	+0.0914	[-0.145, +0.366]	+0.1021

trailing missing runs. Divide by valid length. The maximum KS statistic across the eight resulting coordinates is calibrated using the same 199 encounter-label permutations, with a conservative upper-tail rank including the observed assignment. This summary, test and eight-variable scope were fixed before scoring this follow-up; no clinical outcome or favorable-feature selection enters it. The mean increase in affected-feature normalized gap is 0.051 over the 30 conditions. Because availability and numerical interval construction are not part of this direct mask summary, it is a manipulation check of observation timing, not an alternative imputation uncertainty estimator. It was not added retrospectively to the historical jointly calibrated suite.

Relation to the archived benchmark. The original paired benchmark additionally varies affected-feature count over one, two and four. It gave raw AUROC/Recall@k 1.000/1.000, FOU 0.985/0.942 and temporal-mask C2ST 0.982/0.950. The FOU-minus-mask AUROC difference was 0.003 with perturbation-seed interval $[-0.013, 0.016]$. These use archived Monte Carlo realizations, so the common-bank two-feature values replace them in the primary cross-task comparison. Observation-count and marginal observed-value controls are chance-level by construction under timing rearrangement; they recover variables under dropout. The earlier width-only target-validation-threshold timing test had raw rejection rates 1.3%, 2.3%, 1.8% at $n = 20, 40, 80$ and unperturbed rates 1.8%, 1.4%, 0.8%. It is retained as a distinct protocol, not combined with fresh-window estimates.

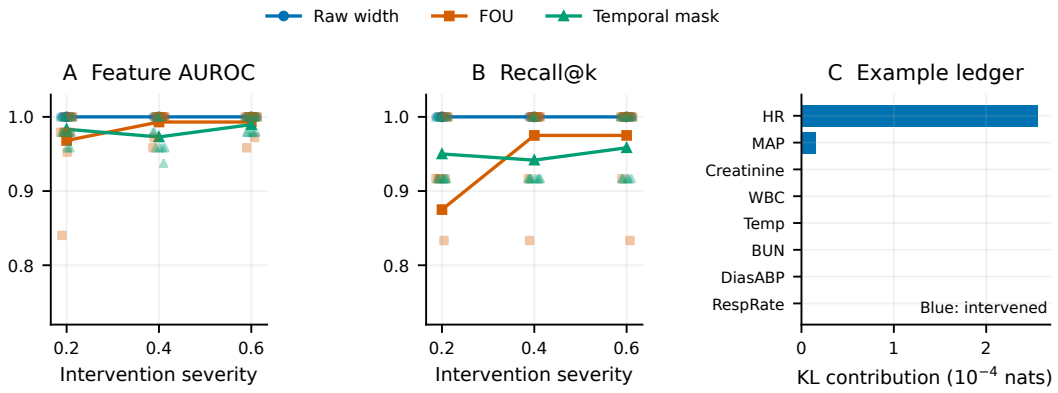


Figure 5: Archived localization on a different Monte Carlo bank, including one/two/four affected variables. This broader benchmark is supplementary to the exact common-bank comparison in Figure 2. Faint points are perturbation seeds; the example is fixed by first seed, two features and severity 0.4.

A.7 Monitor-Input Connection and Monte Carlo Sensitivity

For every easier/harder encounter–feature with at least one artificially hidden cell in both arms, we compute hidden-cell MAE, width and coverage, all-valid-time g_j^{raw}, and mean width across all

missing valid positions. The final two summaries include naturally missing cells, whose reconstruction errors are unknown. Easier/harder arms match encounter, feature, deletion count and fraction, but select different hidden positions. MAE uses the arithmetic mean of imputation draws on each arm's own query cells. Their contrast is not a fixed-query context intervention. In the $S = 10$ versus $S = 50$ check, each arm's removed-cell mask is replayed exactly across draw counts. Every cached g_j^{raw} agrees with the interval-derived value within numerical tolerance; $g_j^{\mathrm{raw}} = (1 - c_j)\bar{u}_j^{\mathrm{miss}}$ also holds. Averaging repeated conditions within encounter–feature prevents repeated perturbations being treated as independent patients. Bootstrap draws resample all feature pairs from a selected encounter together.

Full-grid $\Delta g_j^{\mathrm{raw}}$ medians are 0.00065 and 0, with 51.7% and 49.6% positive for source/medical. Their empirical 5th–95th percentile ranges are $[-0.073, 0.075]$ and $[-0.073, 0.066]$. These descriptive ranges show that a small mean coexists with substantial individual variation; they are not uncertainty intervals for a population effect.

Missing-position width changes have medians 0.0015 and 0, with 52.1% and 49.6% positive. Thus appreciable individual variation coexists with a small macro mean. Spearman correlations between paired ΔMAE and $\Delta g_j^{\mathrm{raw}}$ are -0.015 and -0.014. These descriptive pair-level associations neither estimate a causal response nor test equality of distributions; pairs sharing an encounter are dependent.

Table 5 reports the matching subset means.

The 40 $S = 50$ exports reuse the exact observed masks from the easier/harder $S = 10$ outputs. A complete first-condition $S = 10$ replay in each cohort agrees with saved centers, interval endpoints and masks to 10^{-6} before proceeding. Export batch size stays 32 and the within-cohort Monte Carlo seed stays fixed; because the draw count changes random-number consumption across batches, the full $S = 10$ bank is not a nested prefix of the $S = 50$ bank. Easier/harder arms still share random numbers within each S. This sensitivity therefore varies both Monte Carlo size and its resulting realization, while keeping patient selection, masks and model fixed.

At $S = 50$, source/medical MAE-effect intervals are $[0.166, 0.250]$ and $[0.176, 0.229]$; hidden-width intervals are $[-0.020, 0.037]$ and $[-0.022, 0.013]$. Monitor-input intervals are $[-0.0027, 0.0023]$ and $[-0.0036, -0.0003]$. Coverage effects are -4.21 points ($[-5.81, -2.70]$) and -7.85 points ($[-9.21, -6.51]$). The source coverage loss changes materially from $S = 10$, even though the error/mean-width pattern persists. We therefore do not describe all conclusions as Monte Carlo invariant or use this subset to infer full-grid distributional robustness. Historical CES-D $S = 50, 100$ results concern a separate dataset and do not substitute for this PN2012 check.

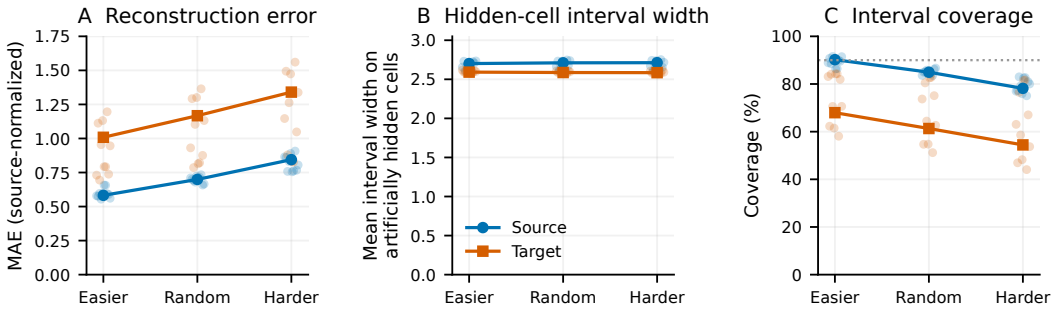


Figure 6: Original cell-weighted easier/random/harder summaries, retained alongside the paired macro analyses. Mean width here is restricted to artificially hidden cells, not the monitoring input. The dashed 90% level is a nominal quantile level, not a guarantee.

A.8 Imputer State, Random-Stream Coupling and Window-Size Boundaries

These are exploratory follow-ups on previously exposed PN2012 cohorts. Endpoints, masking rules, Monte Carlo seeds and window sizes were fixed before the new scores were computed. The frozen imputer, normalizer, numerical feature scope and existing timing interventions were retained. No checkpoint or condition was selected for a favorable response. The checks distinguish finite Monte

Carlo variability, patient sampling and basic reconstruction quality; they do not estimate variability across trained imputers.

Source-validation quality and fixed-query context removal. We reuse the source-validation split of 86 encounters and the original 20% pseudo-mask seeds 820201 and 820202. For each encounter and scope feature, the hidden count is stochastically rounded from 20% of available observed cells, which are then sampled without replacement. These two masks supply 4358 query-cell evaluations, including repeated cells across masks. They are not 4358 independent patients. Complete $S = 10$ replays, batch size 32 and MC seed equal to mask seed plus 401, reproduce all previously saved hidden-cell absolute errors to 10^{-6}. The source-validation data had already been used for difficulty-model diagnostics; this is a descriptive working-state audit, not an untouched confirmation set.

Flow MAE is absolute error of the arithmetic sample mean, $\widehat{x}_{tj} = S^{-1} \sum_{s=1}^{S} \tilde{x}_{tj}^{(s)}$, not the sample median or interval midpoint. The mapper computes `samples.mean(dim=0)` and exports it as `mu_center`; the quality scorer reads that tensor at the fixed query coordinates. Source-mean imputation is zero in source-normalized coordinates. Linear interpolation uses only retained observations of the same feature within the same encounter, with nearest-endpoint extrapolation and source-mean fallback if none remain. Its use of later observations within the fixed history matches the retrospective imputer's permitted context; it is not an online causal imputation comparator. No clinical label or hidden query value enters either reference. Macro RMSE is the square root of the macro mean squared error: random-mask values are 0.966 for the flow imputer, 0.986 for source mean and 0.816 for linear interpolation. After thinning, they are 0.983, 0.986 and 0.845. Thus both MAE and RMSE show the flow imputer's limited performance against this simple reference.

The input control applies the same masking algorithm to 50% of the remaining observed scope cells, using seed equal to the query-mask seed plus 100000. The originally hidden queries stay fixed and are disjoint from the additional deletions; retained values are unchanged. Shared random numbers are used between the two context arms. Width on fixed query cells is the primary conditional-response endpoint. We additionally evaluate width on the *common* missing set from the random-mask arm, each arm's full missing set, and the deployed g_j^{raw}. The common-set comparison excludes changes in the composition of missing positions. In both arms, $g_j^{\mathrm{raw}} = (1 - c_j)\bar{u}_j^{\mathrm{miss}}$ is checked against the interval arrays.

Average cells within encounter–feature–mask, then average the masks available for each encounter–feature. The resulting 532 pairs span all 86 encounters. Bootstrap samples select encounters, retaining all their feature pairs, for 2000 replicates. Table 15 reports the levels and paired effects. Random-query width has median 2.654, interquartile range $[2.225, 3.111]$ and empirical 5th–95th percentile range $[1.674, 3.813]$; the thinned-context values are 2.674, $[2.236, 3.144]$ and $[1.693, 3.843]$. These pooled-cell quantiles describe dispersion, not confidence intervals. Width increases on the fixed query and common-missing sets are small but distinguishable from zero under this conditional bootstrap. They do not establish useful error ranking, correct conditional coverage, or equality of width distributions.

Table 15: Imputer-state macro means and 95% encounter-cluster intervals. Both arms evaluate the same original hidden queries; common-missing width fixes the random-mask arm's missing positions. Paired effects are formed within pairs before aggregation.

Metric	Random	Less context	Paired difference
Flow MAE	0.6804 [0.6268, 0.7384]	0.6964 [0.6418, 0.7554]	0.0160 [0.0112, 0.0209]
Source-mean MAE	0.6995 [0.6435, 0.7596]	0.6995 [0.6435, 0.7596]	0.0000 [0.0000, 0.0000]
Linear MAE	0.3284 [0.2869, 0.3878]	0.3781 [0.3349, 0.4334]	0.0497 [0.0333, 0.0665]
Coverage (%)	85.6152 [83.2372, 87.7689]	84.8338 [82.3752, 87.1465]	-0.7814 [-1.4608, -0.1904]
Query width	2.6212 [2.5885, 2.6539]	2.6329 [2.5996, 2.6656]	0.0117 [0.0095, 0.0139]
All-missing width	2.6443 [2.6342, 2.6536]	2.6539 [2.6462, 2.6614]	0.0096 [0.0042, 0.0150]
Common-missing width	2.6443 [2.6342, 2.6536]	2.6550 [2.6448, 2.6646]	0.0107 [0.0093, 0.0121]
Monitor g^{raw}	1.7409 [1.6970, 1.7857]	2.1999 [2.1777, 2.2218]	0.4590 [0.4357, 0.4820]
Observed fraction	0.3359 [0.3185, 0.3530]	0.1678 [0.1593, 0.1763]	-0.1681 [-0.1767, -0.1593]

Independent random-stream paired localization. `infer` and `infer_temporal` reset Python, NumPy and PyTorch CPU/CUDA seeds once at the start of every export. They retain encounter order,

padded shape and batch size 32. For each batch, `sample_imputations` draws a full $B \times 48 \times 41$ Gaussian tensor with `torch.randn_like(x_obs)` for each of the S draws, then applies the fill/keep masks. The network is in evaluation mode and the Euler solver is deterministic. Thus the shared stream aligns by batch, draw, encounter, time and feature, including positions whose mask changes; it does not allocate noise only to missing cells. The 223 encounters form six batches of 32 and one of 31. Changing order, batch size, S or backend can change this alignment; using the same seed alone is insufficient. These are code-path facts, supported by replay parity of the saved representations, not a claim that noise tensors themselves were archived.

The old baseline MC seed is 991004; its representation replay agrees with the saved bank to 10^{-6}. New unperturbed exports use seeds 992101, 992102 and 992103, the same encounter order and batch size 32, and $S = 10$. All input summaries and masks are unchanged. Each new baseline is compared with all 30 original altered banks, both on all 223 encounters and on the same target subsets of 40 used in the 90 fresh windows. No new intervention export or patient cohort is introduced. Three MC banks assess sensitivity to three numerical realizations; they are not three imputation models.

The MC-only control compares each new unperturbed export with the old unperturbed export. It uses the planned intervention-variable labels solely as negative-control ranking labels; no variable was altered in this control. The raw $n = 40$ AUROCs for the three independent-baseline comparisons are 0.387, 0.381 and 0.390. Their mean 0.386 and the MC-only mean 0.384 have intervals spanning 0.5; we do not interpret them as inverse localization. Table 16 retains every seed, both representations and both sample sizes. Mean intervals first average MC banks, severities and window draws within each perturbation seed, then resample the ten perturbation seeds 2000 times. They condition on patients, model and the chosen MC banks.

Table 16: All random-stream comparisons. KL is the median total discrepancy in nats; independent-mean and MC-only rows pool the three fixed new baseline banks. Disjoint rows use 40 different reference patients and 40 target patients.

n	Comparison	Raw AUROC	Recall@2	KL	FOU AUROC	Recall@2	KL
40	Shared RNG	0.987	0.972	0.00115	0.960	0.861	0.00021
40	Independent 1	0.387	0.178	0.25125	0.387	0.178	0.09149
40	Independent 2	0.381	0.172	0.21103	0.405	0.206	0.08910
40	Independent 3	0.390	0.144	0.21847	0.382	0.172	0.08515
40	Independent mean	0.386	0.165	0.22212	0.391	0.185	0.08921
40	MC-only mean	0.384	0.170	0.22105	0.386	0.181	0.08945
40	Disjoint patients	0.447	0.217	1.13766	0.420	0.200	1.33869
223	Shared RNG	0.997	0.983	0.00031	0.953	0.817	0.00008
223	Independent 1	0.403	0.133	0.08252	0.425	0.200	0.03802
223	Independent 2	0.392	0.200	0.10492	0.428	0.183	0.05773
223	Independent 3	0.353	0.083	0.07317	0.417	0.200	0.03801
223	Independent mean	0.382	0.139	0.08252	0.423	0.194	0.03829
223	MC-only mean	0.383	0.133	0.08262	0.428	0.200	0.03801

Nested window-size checks. For each of the 90 existing within-test pairs, retain the original $n = 40$ assignments. With sampling seed $260926100 + \text{pair_id}$, independently select 20 patients from each original arm; shuffle the remaining 143 patients and allocate the first 60 to reference extensions and the next 60 to target extensions. The $n = 80$ extensions use the first 40 of these respective pools; $n = 100$ uses all 60. This preserves nesting and disjointness within every comparison. The 223-patient pool and original 30 timing interventions are reused across sizes; resampling supplies no new independent patients.

Only canonical raw-width KL and the already specified longest-gap KS check are evaluated. Each matched unperturbed/intervened pair uses its original permutation seed, 199 whole-encounter permutations and the observed assignment. KDE bandwidths and grids are recomputed after each assignment. All $n = 40$ raw-KL permutation matrices reproduce the previous scores, and all gap-check p-values reproduce the previous result. Table 17 retains both transition directions; there is one new and two lost raw alarms at $n = 80$, and one new raw alarm at $n = 100$. The near-flat raw response therefore is not described as literally invariant at every tested size. Seed-block intervals hold nested assignments, draw bank and the patient pool fixed.

The gap diagnostic becomes more responsive over this range while raw-width alarms remain sparse. This is a boundary over $n = 20$–100 and one $S = 10$ representation bank, not a power limit for all

Table 17: Window-size boundary, 90 matched pairs at each n, 199 permutations and nominal 5% per-method level. Unperturbed comparisons are from the same fixed pool. z_Δ uses each method's unperturbed permutation SD.

n	Method	Unperturbed	Timing	$0 \to 1$	$1 \to 0$	Mean z_Δ
20	Raw KL	1	1	0	0	+0.0063
20	Gap KS	2	20	18	0	+0.9067
40	Raw KL	6	6	0	0	+0.0010
40	Gap KS	5	31	27	1	+1.6259
80	Raw KL	7	6	1	2	-0.0001
80	Gap KS	6	41	35	0	+3.0402
100	Raw KL	3	4	1	0	-0.0019
100	Gap KS	8	45	37	0	+3.3871

sample sizes or imputers. The independent-stream localization check and the difficulty $S = 50$ check answer different questions; neither substitutes for timing analysis at larger S. The accompanying `boundary_reproduction.zip` provides executable replay of these core results from saved public-data exports. Its README and Appendix A.9 give the replay command, verification and scope. A broader software release remains planned after publication.

A.9 Reproduction Entry Point

The anonymous `boundary_reproduction.zip` provides one CPU entry point for the current localization, window-detection and imputer-state analyses. After extracting the archive and installing `requirements.txt`, run `python replay.py -out replay`. The destination must be new. No parent checkout, GPU or private data is required. The included README explains the data layout; `manifest.json` verifies supplied-file hashes and `provenance.json` distinguishes frozen source records from the anonymous review copies.

The command recomputes localization and quality from saved arrays, continuous changes and window summaries from complete permutation matrices, and checks regenerated outputs against frozen expected results. It writes `replay/validation.json`. Adding `-rescore-smoke` recalculates eight representative matched-window tests; `-rescore-all -workers 12` recalculates all 360 raw-KL/gap comparisons and their 199 permutations. The default mode does not refit C2ST classifiers; their earlier full replay remains in `pn2012_reproduction.zip`.

Row indices preserve pairing while direct identifiers and clinical labels are excluded. Model parameters and numerical inputs are unchanged by neutral implementation names and metadata cleanup. These materials reproduce statistics conditional on saved exports, not all upstream imputation draws, model training or private NIH-DEP analyses. The separate public-record export package retains its stated scope.

A.10 Fixed-Budget Triage and Strong Controls

Fixed-budget analysis asks which windows to inspect, independent of a 5% alarm claim. The offline event is a Brier-score increase of at least 0.01 on the same encounters after intervention. The ranking uses no evaluation outcomes. At the primary 20% audit budget, the selected count is exactly one fifth of the candidate windows; uniform random selection therefore has expected event recall 20% when any events exist. Boundary ties receive fractional selection weights. Signed losses, zero-event settings (NA recall), alternative budgets, all single components and original sampling-stream results are retained in the aggregate artifacts. Conditional seed intervals exclude patient sampling, model fitting and development uncertainty.

The matched-acquisition comparison pairs absolute deletion counts across higher/lower source-sensitivity feature groups. Because the variables have different observation densities, their relative deletion fractions are not matched. The second Monte Carlo stream uses independent draws across train/validation/test exports, while preserving paired baseline/intervention draws within each split. This changes sampling noise, not the clinical patient population. Figure 7 shows both streams, in-

cluding reversal of the source $n = 40$ increment. The same-variable conditional-deletion comparison further fixes patient, variable, deletion count and fraction; Table 18 retains all 24 predictor/settings.

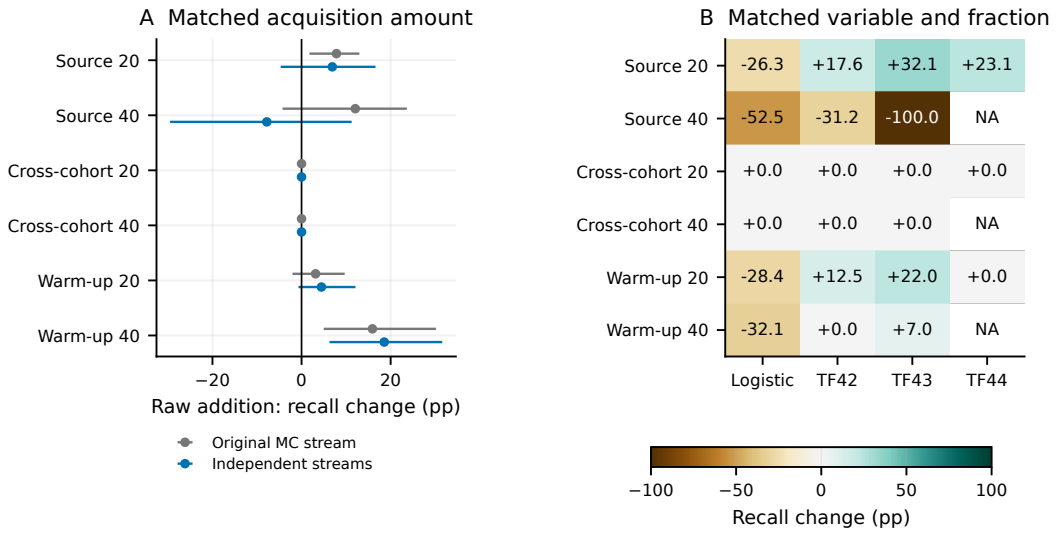


Figure 7: Conditional audit gains across protocols. A: paired raw-addition recall differences for two Monte Carlo streams on the same patients; bars are perturbation-seed intervals. B: all predictor/settings after matching the variable and deletion fraction. NA denotes no loss events, not zero recall. Cross-cohort rows use only source reference data; warm-up rows use target reference data.

The expanded suite adds a nonlinear input/mask C2ST using six value summaries per feature and complete temporal masks, scope-matched raw and mask controls, source-weighted statistics, and actual-predictor column-occlusion diagnostics. It does not compare against an arbitrary optimal classifier of the full raw sequence. Best-single values below are descriptive maxima over the reported controls, not a strategy selected without evaluation information.

Probability-scale sensitivity. The three clinical Transformers were trained with positive-class-weighted BCE, weight $344/59$. A separate sensitivity subtracts $\log(344/59) = 1.763104$ from logits. This source-weight-only offset neither fits a calibration model nor guarantees calibrated probabilities; detector rankings remain fixed. Source $n = 20$ event counts 17/28/13 become zero, and all source $n = 40$ settings also have zero events. Some target $n = 20$ raw additions and temporal-mask controls both reach complete recall. We retain original and shifted scales rather than choosing whichever favors widths. A fixed absolute loss threshold is inseparable from the probability scale being evaluated.

Table 18: All conditional-deletion loss-event recall results at 20% budget (percent). S: source; C: cross-cohort with source reference; W: target warm-up. Events count overlapping intervention windows. *Best single is the non-width descriptive envelope.

Predictor	Protocol / n	Events	Suite	+raw	+FOU	Raw only	Best single*
logistic	S / 20	78	40.4	14.1	3.8	12.8	62.8
TF42	S / 20	17	29.4	47.1	47.1	47.1	50.5
TF43	S / 20	28	25.0	57.1	46.4	57.1	35.7
TF44	S / 20	13	15.4	38.5	38.5	38.5	38.5
logistic	S / 40	101	52.5	0.0	0.0	0.0	65.3
TF42	S / 40	7	45.5	14.3	14.3	14.3	100.0
TF43	S / 40	3	100.0	0.0	0.0	0.0	100.0
TF44	S / 40	0	NA	NA	NA	NA	NA
logistic	C / 20	116	12.9	12.9	12.9	3.4	51.3
TF42	C / 20	8	12.5	12.5	12.5	62.5	62.5
TF43	C / 20	50	34.0	34.0	34.0	40.0	64.0
TF44	C / 20	5	0.0	0.0	0.0	100.0	100.0
logistic	C / 40	193	21.2	21.2	16.6	0.0	48.2
TF42	C / 40	3	100.0	100.0	100.0	0.0	100.0
TF43	C / 40	43	44.2	44.2	41.9	18.6	48.8
TF44	C / 40	0	NA	NA	NA	NA	NA
logistic	W / 20	116	29.3	0.9	1.7	0.9	41.4
TF42	W / 20	8	50.0	62.5	62.5	62.5	62.5
TF43	W / 20	50	18.0	40.0	40.0	40.0	38.0
TF44	W / 20	5	100.0	100.0	100.0	100.0	100.0
logistic	W / 40	193	32.1	0.0	0.0	0.0	42.5
TF42	W / 40	3	0.0	0.0	0.0	0.0	100.0
TF43	W / 40	43	7.0	14.0	14.0	14.0	34.9
TF44	W / 40	0	NA	NA	NA	NA	NA

A.11 Actual Variable Recovery

We evaluate an action after diagnostic ranking by restoring originally observed values of the top-ranked variables, rerunning the frozen imputer, and recomputing the actual clinical prediction. Each query selects from the same eight numerical variables; budgets are one, two or four variables. Ties are averaged over the possible query sets. The benefit is perturbed-minus-restored Brier loss, so positive means improvement. Variables not deleted by the intervention may consume a query without restoring any cells. This is a retrospective oracle restoration experiment: the deleted true measurements are available for the simulation, not an assumption that they can always be recollected clinically.

The main $n = 40$, one-variable comparison has negative mean raw/FOU recovery benefit in all 12 predictor/reference settings, also for their source-weighted versions (Table 19). Some source-weighted mask controls are positive and are retained. Selected $n = 20$ width results are positive; benefits therefore vary with setting. Even complete restoration of the original record has negative average benefit at $n = 40$, demonstrating that deleting observations did not reliably worsen these frozen predictors. This is why finding a changed variable does not itself establish a useful action. Conditioning on a loss event can describe recovery in that subset, but cannot supply an unlabeled deployment trigger.

The source-weight logit-offset sensitivity preserves the negative $n = 40$, one-query raw/FOU benefits and their differences from random selection. It changes the numerical losses, not the original rankings or recovered variables. Recovery is not used to select a favorable alarm threshold or clinical loss definition.

Table 19: Brier improvement $\times 1000$ after one-variable restoration, $n = 40$. No action gives zero in every row; negative is worse. S/C/W are defined in Table 18; w-mask is the source-weighted temporal-mask control; w-raw/w-FOU are source-weighted width ledgers. Full query-budget and $n = 20$ results are retained in the aggregate CSV.

Head	Protocol	Random	Input/mask	w-mask	Raw	FOU	w-raw	w-FOU
logistic	S	-1.64	-16.90	-19.11	-18.88	-18.95	-18.99	-18.99
logistic	C	-3.53	-37.44	-41.95	-39.54	-37.47	-40.02	-37.65
logistic	W	-3.53	-36.42	-40.16	-37.77	-37.50	-38.54	-38.06
TF42	S	-0.60	-2.65	+1.20	-2.73	-2.81	-2.05	-2.10
TF42	C	-0.74	-5.19	+0.84	-5.24	-5.10	-4.80	-4.80
TF42	W	-0.74	-5.20	-0.34	-5.13	-5.11	-4.81	-4.89
TF43	S	-0.34	-1.51	+0.61	-1.17	-1.20	-0.56	-0.42
TF43	C	-0.27	-4.87	+3.21	-4.76	-4.78	-2.35	-2.15
TF43	W	-0.27	-4.67	+3.21	-4.29	-4.27	-1.87	-1.91
TF44	S	-0.71	-3.41	+0.29	-3.23	-3.25	-2.76	-2.75
TF44	C	-0.60	-3.93	+0.46	-3.92	-3.79	-3.58	-3.59
TF44	W	-0.60	-3.89	+0.48	-3.95	-3.85	-3.34	-3.34

A.12 Classification-Error Probe and Monte Carlo Interval Resolution

The NIH-DEP analysis reuses five existing PR2-mild clinical heads (42–46), implemented as FuzzyTokenTransformer classifiers with fuzzy attention and an MLP head, each evaluated on the same 52,185 windows from 23,054 people. The recorded binary endpoint is `hm_dep` at the target wave; input histories precede that wave. This analysis does not create a new diagnostic threshold from CES-D scores. Checkpoint metadata and source configuration fix the single CES-D imputation target. Each row is keyed by participant and target wave; label, wave, sequence length and tensor order are checked. Replaying all five sets of logits differs from their saved values by at most 3.58×10^{-6} (predeclared absolute/relative tolerances $10^{-5}/2 \times 10^{-5}$). Original, uncalibrated sigmoid probabilities are used, with each head's existing validation-selected threshold. Direct checks of the original Stage-B participant IDs give zero train/validation/test overlap; their window counts are 243,463/51,904/52,185 and participant counts are 107,583/23,053/23,054. Clinical-head fitting uses train; checkpoint and threshold selection use validation. This audit does not turn the repeatedly analyzed test data into independent confirmation.

Width summaries use available valid cells in that source-defined scope, with zero if no cells are available; they are not relabeled as the multi-target valid-time averages. GroupKFold assigns every participant to one of five folds. Within each training fold, a StandardScaler and $C = 1$ L2 logistic risk model are fitted; no auxiliary hyperparameter or feature search uses the held-out fold. The intermediate ablation uses m, c, mc and missingness, with $m = -|p - \tau|$ and $c = \mathbf{1}[p \geq \tau]$. It restores the threshold side and an interaction without p^2; indeed $p = \tau + m - 2mc$ is in the intermediate feature span. The richer baseline includes p, p^2, predicted class, their interaction $p\mathbf{1}[p \geq \tau]$, and missingness. Width features are appended to the identical baseline and folds. This is supervised retrospective error-risk analysis within these data, not unlabeled monitoring.

Intervals use 2000 paired bootstrap draws over participants and the five head seeds, holding out-of-fold predictions fixed. They exclude auxiliary-model refitting and prior development uncertainty. Metrics score windows; AUPRC is computed as average precision and NLL as binary log loss. Beyond AUROC, raw addition changes error AUPRC by -0.00039 (95% interval $[-0.00171, 0.00102]$) and NLL by -0.000099 ($[-0.000233, 0.000033]$). FOU changes AUPRC by -0.00049 ($[-0.00165, 0.00074]$) and NLL by -0.000099 ($[-0.000230, 0.000027]$). Thus the richer-baseline comparison has no demonstrated incremental benefit on any of these three metrics. The original margin results remain in Table 6 for direct comparison.

Numerical interval coverage. On the historical controlled CES-D holdout, empirical coverage of the Monte Carlo 5%–95% quantile interval is 0.7632 for $S = 10$, 0.8938 for $S = 50$, and 0.9083 for $S = 100$. Validation-fitted interval expansion gives 0.9004 for $S = 10$; the main historical monitors use unexpanded widths. Under an ideal continuous, correctly specified distribution with independent draws, the empirical quantile interval lies inside the sample extrema. An independent

truth is between the extrema with probability $(S-1)/(S+1)$, since its rank among $S+1$ draws is uniform. At $S=10$ even this upper bound is $9/11 \approx 81.8\%$. Finite Monte Carlo resolution can therefore contribute to low coverage; the ideal calculation is not a coverage prediction for the real CES-D distribution. FOU inherits no numerical interval-coverage guarantee. In the original five pseudo-mask seeds, Spearman associations with absolute reconstruction error are 0.0139 ± 0.0190 for FOU and 0.2044 ± 0.0319 for raw width.

Within-feature dependence versus cross-feature dependence. An interval-level experiment fixes midpoint and width marginal multisets but changes their coupling. FOU, joint midpoint/width MMD and a Spearman-dependence diagnostic all localize the changed coordinates perfectly; separate midpoint and raw-width marginal monitors are at chance. The projection exposes information already in the joint interval representation. Dependence changes *among* width coordinates are a different question and may escape any marginal ledger (see the separate archived-results supplement).

Dependent-dyad correction. This supplementary PN2019 cohort-performance analysis uses source-matched single-HR FOU/prediction replays and six-node QAP permutations. It is separate from the NIH-DEP individual-error analysis in Section 5.3. Width discrepancies compare source-training and target-test strata. Both source and target subgroup AUROCs use the same source-hospital predictor and test encounters. For hospital A, low/medium/high missingness strata contain 4592/3325/6313 training encounters and 963/710/1377 test encounters; for hospital B the counts are 4820/6088/3089 and 1034/1307/659. These form six reused nodes, not 30 independent samples.

The dependence-aware $\rho = 0.1153$, $p = 0.5819$ concerns the absolute source/target AUROC gap; it must not be relabeled as signed degradation. Signed source-minus-target AUROC instead gives $\rho = 0.0065$, $p = 0.9736$. Both use the internally consistent replay, rather than mixing older widths with newly replayed predictions. These corrected associations supersede the earlier row-wise analysis and describe a secondary cohort-performance check; they do not support the main PN2012 timing or difficulty conclusions.

Separate sample-deviation diagnostic. The separate score

$$a_j(x) = \left| \frac{g_j(x) - \mu_j^S}{\sigma_j^S + \varepsilon} \right|$$

measures sample-level deviation from source width moments. It is not a sample decomposition of KL and should not be conflated with intervention localization accuracy. The main error-ranking analysis instead fits a supervised risk model to width and probability features; it does not validate $a_j(x)$.

Estimator scope. Observed-cell zero widths and degenerate coordinates can create point masses. The continuous-density consistency proposition below does not cover such distributions. Bandwidth floors and skip rules describe the implemented estimator, rather than furnishing finite-sample detection guarantees. Fixed-grid results must be interpreted with their actual null assessments.

Provenance. The revision preserves frozen experiment protocols, source snapshots, model/input hashes, component statistics, aggregate results and deterministic figure scripts. Public-release artifacts exclude NIH-DEP participant identifiers, row-level records and prediction arrays. The public PN2012 experiments supply the principal localization, alarm, deletion-difficulty and recovery comparisons. The NIH-DEP aggregates can be inspected, but reproducing their row-level analysis currently requires access to the laboratory's private data; the authors are authorized to use these data, and access procedures and links will be disclosed after publication to preserve review anonymity.

B Formal Properties

Proposition 1 (Marginal width divergence). *Let $P_j = P \circ g_j^{-1}$ and $Q_j = Q \circ g_j^{-1}$ be the one-dimensional pushforward distributions for width coordinate j, and let*

$$D_g(P, Q) = \sum_{j \in \mathcal{A}} \mathrm{KL}(P_j \| Q_j),$$

where $\mathcal{A}$ is the active-coordinate set and $\mathrm{KL}(P_j\|Q_j) = +\infty$ when P_j is not absolutely continuous with respect to Q_j. Then $D_g(P,Q) \geq 0$. If $D_g(P,Q) < \infty$, then $D_g(P,Q) = 0$ iff $P_j = Q_j$ for all $j \in \mathcal{A}$. With natural logarithms,

$$\sum_{j\in\mathcal{A}} \|P_j - Q_j\|_{\mathrm{TV}}^2 \leq \frac{1}{2} D_g(P,Q).$$

The statistic is marginal: it does not identify changes that preserve every P_j and Q_j while changing only the dependence among width coordinates.

Proof. Non-negativity and the equality condition follow from Gibbs' inequality applied to each KL summand. Pinsker's inequality gives $\|P_j - Q_j\|_{\mathrm{TV}}^2 \leq \mathrm{KL}(P_j\|Q_j)/2$ for each active coordinate; summing over j gives the displayed bound. For the dependence limitation, compare two bivariate width distributions with identical uniform marginals but different copulas, such as an independent copula and a Gaussian copula with nonzero correlation. All one-dimensional width marginals match, so $D_g = 0$, although the joint width distributions differ.

Proposition 2 (Regularized KDE plug-in consistency)**.** *Fix an active coordinate j. Let p_j and q_j be the source and target width densities supported on a compact monitoring interval I_j and normalized to integrate to one there, with $q_j(z) \geq c > 0$ on I_j, and assume the KDEs $\hat{p}_{j,h}^S$ and $\hat{p}_{j,h}^T$ are uniformly consistent on I_j under bandwidths $h_S, h_T \to 0$ with $n_S h_S / \log n_S \to \infty$ and $n_T h_T / \log n_T \to \infty$. Let the grid spacing $\Delta_n \to 0$ and the density floor $\varepsilon_n \to 0$. Define normalized grid probabilities by adding ε_n to the KDE values and renormalizing over the grid, as in Section 3.2. Then the grid KL for coordinate j converges in probability to $\mathrm{KL}(P_j\|Q_j)$. For fixed finite $|\mathcal{A}|$, the summed statistic converges in probability to $D_g(P,Q)$.*

Proof sketch. Uniform KDE consistency gives $\sup_{z\in I_j} |\hat{p}_{j,h}^S(z) - p_j(z)| \to 0$ and $\sup_{z\in I_j} |\hat{p}_{j,h}^T(z) - q_j(z)| \to 0$ in probability under standard smoothness and boundary-correction conditions (Silverman, 1986; Tsybakov, 2009). Because q_j is bounded away from zero and $\varepsilon_n \to 0$, the regularized and normalized target estimate is eventually bounded away from zero with high probability. The map $(u,v) \mapsto u\log(u/v)$ is continuous on this region, so the integrand converges uniformly. The Riemann-grid approximation error vanishes as $\Delta_n \to 0$, yielding convergence of the coordinatewise grid KL. A finite sum over active coordinates preserves convergence. This is an asymptotic statement for a vanishing grid step and vanishing regularizer; the finite-sample implementation uses a fixed 200-point grid and 10^{-8} floor and has no guarantee from this asymptotic statement. If the monitoring interval truncates probability mass, normalization instead targets the corresponding conditional distributions; the implemented finite-grid statistic requires its own empirical null assessment.