MEASURING THE TEMPORAL AND IDENTITY READABILITY OF OPERATIONAL LOGS ACROSS SIX DOMAINS

Anonymous authors
Paper under double-blind review

ABSTRACT

Retrospective event-prediction evaluations can reward recognition of temporal position or unit identity, making apparent predictive performance difficult to interpret. We introduce a reusable audit that measures these signals in operational logs before an event-prediction model is evaluated. Using windows from units without recorded adverse events, the audit asks two questions: which of two windows from the same unit came later, and whether two windows matched on time and offset came from the same unit. These tasks measure *temporal readability* and *identity exposure*, respectively; under their stated exchangeability nulls, expected AUC is exactly 0.5. We release both audits as one harness, driven by a small adapter and a declaration file, and apply it across six public domains. Removing amortization clocks reduces temporal AUC in mortgage records from 0.834 to 0.549, while identity AUC stays at 0.994. In ICU records, measurement counts alone reveal temporal position (AUC 0.634) and distinguish same-patient from different-patient pairs (AUC 0.879). A controlled case-control experiment connects identity exposure to evaluation inflation: carry-forward imputation increases the AUC gap between overlapping-unit and unit-disjoint splits by 0.118. This experiment measures susceptibility to unit overlap rather than prospective prediction skill. Together, the audits characterize temporal and identity signals that can confound evaluation and motivate controls appropriate to the intended prediction task.

1 INTRODUCTION

A model watches a hard drive, a vessel, or a patient and raises an alarm when recent data looks like trouble. Behind every such claim sit two rarely tested null hypotheses: in the absence of trouble, one stretch of a unit's data looks like any other stretch of that unit's data, and one unit's ordinary data looks like another's. If the first fails, an evaluation can score temporal position, phase, season, drift, as if it were anticipation. If the second fails, an evaluation whose sampling correlates with units can score recognition, learning *who* the risky units are, as if it were foresight about *state*. Both failures are documented. Leakage taxonomies catalog the failure class (Kaufman et al., 2012), including its identity route (Kapoor & Narayanan, 2023), patient-overlap studies measure its inflation (Tampu et al., 2022), permutation tests detect it in fitted pipelines (Chaibub Neto et al., 2019), and the case-crossover literature has long repaired time trends that survive self-matching (Suissa, 1995). Their *preconditions*, however, live in the dataset before an event-prediction model or evaluation protocol is chosen, and this paper is about measuring those preconditions instead of assuming them away. What we contribute is an auditing protocol and a benchmark, a reusable measurement of two latent vulnerabilities of a dataset taken before any event-prediction model is evaluated on it, with the six-domain study as its validation.

Both measurements instantiate one construction. We build pairs of ordinary windows such that, under a declared null, the two classes are identically distributed, so symmetry forces every classifier's expected AUC to 0.5 (Clémençon et al., 2009); split the units before drawing pairs; train a classifier on the training units and score the held-out ones; read the measured excess as readability along the axis the pairing targets. Along time this is the *temporal A/A floor*, two ordinary windows of one unit labeled earlier and later and nothing else, whose excess prices the clocks, seasonality, phase, and drift that any pipeline can read with no event in sight. Along identity it is the *identity A/A ex-*

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.

posure, an anchor paired with a same-unit or different-unit window at the same calendar time and offset, whose excess is the dataset's fingerprintability, the identity signal available to leak into any evaluation whose sampling correlates with units.

For example, take hourly rows from the PhysioNet/Computing in Cardiology 2019 sepsis challenge data (Reyna et al., 2020), restrict to stays without sepsis, choose two rows 12 hours apart, and ask a classifier which came later. Under exchangeability across that interval expected AUC is 0.5, so a larger value shows that ordinary features reveal position in the stay although the classifier never sees a sepsis label. Holding the two times fixed and asking whether the rows came from the same patient is the identity task.

Each axis has a mechanical ancestor with decades of standing, the classifier two-sample test (Lopez-Paz & Oquab, 2017; Clémençon et al., 2009) and biometric verification (Jain et al., 2004), and we build on both. We combine them into one harness with one eligibility test and one forced-null framing, apply it to both axes of the same panels, and measure each axis under controls that target the other, in service of evaluation auditing rather than detection, pretext learning, or authentication. The combination is the instrument; what it reveals across six domains is the contribution.

Contributions. (1) Temporal floors and identity exposures across six public domains spanning hourly, daily, and monthly sampling, with offset curves and a cross-unit placebo that distinguishes readability persisting across units from unit-anchored structure. Readability is not reliably inferred from domain labels alone; attribution differs in kind (clock floors versus structural floors, physiological versus manufactured exposure); the instrument calibrates, reading 0.500 where symmetry forces it; and identity is easier to read than time on every panel where both axes are measurable, an ordering claimed only in its weak form (Sections 4–7, 9, Appendix B). (2) A controlled bridge experiment in which changing carry-forward imputation alone moves the AUC gap between overlapping-unit and unit-disjoint splits from +0.033 to +0.151, and measured exposure separates high-inflation from low-inflation designs (Section 8). (3) A released harness, driven by a small adapter and a declaration file (Section 3). (4) The construction behind both measurements, a proposition stating when its null forces expected AUC to 0.5 in finite samples, cross-axis controls, pooled scoring for aging domains, and tier ladders attributing each number to clocks, hidden structure, or single channels (Section 2).

2 One Construction, Two Axes

Identified units i produce feature rows $x_{i,t}$. Fix an offset range $[\delta_{\min}, \delta_{\max}]$ in the domain's natural rhythm; both instantiations use the same range per dataset.

Temporal instantiation. A pair is (x_{i,t_1}, x_{i,t_2}), $t_2 - t_1 \in [\delta_{\min}, \delta_{\max}]$, drawn from a unit with no recorded adverse event; rows are labeled by temporal position alone. A classifier trains on 80% of units and is evaluated on held-out units; the floor is the pooled AUC with a unit-clustered bootstrap interval. The null is exchangeability of a unit's ordinary windows across the offset range, $(X_{i,t}, X_{i,t+\delta}) \overset{d}{=} (X_{i,t+\delta}, X_{i,t})$ for $\delta \in [\delta_{\min}, \delta_{\max}]$ under the declared pairing and weighting, under which the labels are a symmetric convention and expected AUC is 0.5 for any scorer.

Identity instantiation. Sample an anchor (i, t_1) and offset δ; emit $(x_{i,t_1}, x_{i,t_1+\delta})$ with label 1 and, for the same anchor and the same $t_1+\delta$, $(x_{i,t_1}, x_{j,t_1+\delta})$ for a random $j \neq i$ observed at that time, with label 0. Both classes share the anchor distribution, the calendar, and the offset; only identity differs. Features are $(x_a, x_b, |x_a - x_b|)$; units are split before any pair is drawn and anchor and impostor come from the same side, so every matched pair reaches the scorer intact; intervals come from a dyadic cluster bootstrap that resamples units over both pair roles (validated against a full-regeneration bootstrap, Appendix D). The null is that no persistent unit-specific dependence remains once calendar and offset are matched, $(X_{i,t_1}, X_{i,t_1+\delta}) \overset{d}{=} (X_{i,t_1}, X_{j,t_1+\delta})$ under the matching controls, under which the two classes are identically distributed.

The null is forced, exactly and in finite samples. Both instantiations are cases of one statement, which we name so that its scope is unambiguous.

Proposition 1 (Forced null)**.** *Let the test set be m matched pairs from the declared protocol, pair k contributing a scored item U_k labeled 1 and an item V_k labeled 0, and let s be any scorer that is a fixed measurable function of the features and independent of the held-out labels. If the joint law of*

$((U_1, V_1), \ldots, (U_m, V_m))$ is invariant to exchanging U_k with V_k for every k at once, then the pooled empirical AUC with mid-rank ties satisfies $\mathbb{E}[\widehat{\mathrm{AUC}}] = 1/2$, for every such s and every m.

The exchange is a measure-preserving involution sending $\widehat{\mathrm{AUC}}$ to $1-\widehat{\mathrm{AUC}}$ pointwise, so the proof is four lines. Appendix A gives it, fixes what the expectation is over and what *every classifier* quantifies over, and notes where the released construction departs from the idealization. Two consequences matter here. The invariance is a property of the ordered pair rather than of the two marginals, which is why the temporal null is displayed above as exchangeability of the ordered window pair and not as stationarity of $X_{i,t}$. And the independence requirement is what the unit-disjoint split buys, so the split is load-bearing rather than hygiene.

The controls target the other axis. In the temporal pairing both windows come from one unit, so unit identity does not directly separate the labels. Population-level structure may still distinguish earlier from later rows; the cross-unit placebo in Appendix B measures how much temporal readability persists when within-unit pairing is removed. In the identity pairing, offset matching prevents same-unit pairs from being systematically closer in time, and calendar matching cancels shared environment, so what remains is unit signature rather than shared context. The match is on the panel's time coordinate, which for the ICU is hours since admission rather than a calendar date, so cohort and site differences can contribute to that panel's exposure (Appendix D reports the cohort-matched control where it was run). The identity exposure reads feature values only; identity carried by *when* a unit is observable is availability exposure, outside this number by construction.

Outcomes are an eligibility filter, never a target. Every declaration names the panel's outcome column and every measurement drops units with a recorded event, so all windows are ordinary (the wind panel contains no recorded failures at all). That is their only use. No scorer is fit to an outcome, no outcome is a feature, no reported quantity is a detection score. These measurements are label-free in the sense an audit needs, no event detector and no prediction target, not in the stronger sense of never reading the outcome column. Outcome labels appear in exactly one experiment, the bridge (Section 8), as unit-level case-control labels in a deliberately vulnerable protocol, which is not one of the audit measurements.

Pooled, never paired (temporal axis). Scoring is row-level across the pooled test set. Comparing the two windows of one unit directly is degenerate whenever any feature is a monotone counter, a cumulative counter orders any within-unit pair by construction; under pooled scoring between-unit variance dilutes clocks and the question becomes empirical. Paired variants must not be reported as floors.

What the numbers are. Each population quantity is the Bayes-optimal AUC of its pairing task, a functional of the dataset's distribution jointly with the declared protocol (pairing rule, offset range, feature tier). Any measured value is a reader-specific lower bound, here one fixed gradient-boosting configuration (Appendix C varies the reader). *Floor* is meant only in this diagnostic sense and is never a bound on any detector's AUC. Detection is a different task on different labels, so the audit AUC cannot be subtracted from, added to, or treated as the starting performance of an event detector. Its role beside a warning claim is comparative: it shows how readily the same features reveal temporal position without an event label. We call the identity quantity an *exposure*. It saturates, and the correct reading is not a level to clear but a measure of how much identity signal is available to leak. The null is a ruler, not a belief. A reading of exactly 0.5 is evidence of calibration, not vindication.

Tier ladder and reporting. Each quantity is measured at the *full* tier (all features included in our domain panel, the quoted value), a *reduced* tier with declared explicit clocks removed, and a single *benign channel*. Collapse under clock removal means clock arithmetic; persistence means structure that deleting a column cannot fix. Published values are means over five pair-draw seeds, because a single draw carries sampling variance comparable to the bootstrap width, so the quoted interval carries both variance sources at once. The unit-cluster bootstrap runs inside *every* seed and the replicates are pooled, making the interval the percentile interval of a distribution that varies both the pairs drawn and the units resampled. A single-seed interval conditions on that seed's pairs, which is how a five-seed mean can fall outside it; the pooled interval describes the spread of one measurement while the quoted point averages five, so it describes a different target from a confidence interval for the five-seed mean. No general coverage guarantee is claimed. Per-seed values and single-seed intervals ship with the results, and Appendix D compares the pooled interval against a full-regeneration bootstrap.

Domain	Offsets	Full	Reduced	Benign channel
Mortgage loans	6–18 mo	0.834	0.549	0.500 (rate)
ICU vitals	6–24 h	0.718	0.677	0.634 (counts)
Drive telemetry	30–90 d	0.598	0.527	0.507 (temp.)
Wind SCADA	14–60 d	0.539	0.540[b]	0.517 (ambient)
Student clicks	14–60 d	0.534	0.520	n/a
Vessel tracks	14–60 d	0.527	0.517	0.505 (kinematics)

Table 1: Temporal A/A floors of six operational-log domains (five-seed means). [b]at wind's near-null floor with twenty units the tiers are statistically indistinguishable.

3 ELIGIBILITY

Both quantities require five things: (1) identified units observed repeatedly; (2) enough units for a unit-disjoint split, tens of thousands is comfortable, twenty works with wide intervals on the temporal axis and fails on the identity axis; (3) per-unit spans longer than the offset range; (4) features at each timestamp; (5) a clock declaration from the analyst. They do not require a prediction task, labeled events beyond the column that excludes event-bearing units (the wind farms ran without a single failure), completeness, or any particular sampling rate. The test collapses to one sentence: do many identified units each have timestamped feature rows spanning more than the offset you care about? Two public datasets that fail it (MetroPT, one unit; Scania APS, one snapshot per truck) are in Appendix F.

4 THE TEMPORAL FLOOR ACROSS SIX DOMAINS

Table 1 is the temporal atlas. Each domain follows the same protocol: an adapter maps raw data to the panel contract, a declaration fixes offsets, tiers, and clocks, and the measurement runs. Full domain narratives are in Appendix E; four results carry the section.

The mortgage panel spans the whole range. With amortization clocks (balance, age, remaining term) the floor is 0.834, a vintage's loans amortize in lockstep; without them it drops to 0.549, and the loan's fixed rate alone reads exactly 0.500 to four decimals, the value symmetry forces for a within-unit-constant feature. The harness returns the forced value where the null holds by construction (a sanity check on a constructed null, not evidence about hard cases).

The ICU floor is high and is not simple clock arithmetic. It is 0.718 [0.706, 0.730] on never-septic stays; values alone read 0.677, and measurement *counts* alone, with no physiology whatsoever, read 0.634. Time is written into the rhythm of care. A cross-unit placebo (Appendix B) supports a shared stay-phase interpretation: the readability is almost entirely structure, available across patients at matched stay hours, with a within-unit AUC contrast of only 0.021 at these offsets.

Independent reimplementation shifts rungs by more than seed noise. An earlier independent implementation of the same protocol reads 0.580/0.520/0.514 across the drive ladder, about 0.02 below this harness while seed ranges span only $\sim$0.01; the ordering is identical and both values are disclosed (Appendix F).

Three floors are near-null. Wind SCADA reads 0.539 (weather decorrelates in days), student clickstreams read 0.534 (on active days the *mix* of activity is nearly stationary; the calendar governs which days are active), and casualty-free vessel tracks read 0.527 [0.496, 0.558], 0.517 without position (Appendix E).

5 THE IDENTITY EXPOSURE ACROSS SIX DOMAINS

Table 2 is the identity survey, on the same panels and offset ranges as the temporal atlas. Four results carry it.

Saturation is individual, not cohort. Drive exposure is 1.000 with clocks and 0.998 without, so the wear-state pattern itself is unit-unique. Restricting every impostor to the anchor's own drive

Domain	Offsets	Exposure	Reduced tier
Hard drives	30–90 d	1.000	0.998 (no clocks)
Mortgage loans	6–18 mo	1.000	0.994 (no clocks)
ICU patients (LOCF)	6–24 h	0.993	0.879 (rhythm only)
Vessel tracks	14–60 d	0.955	0.886 (no location)
ICU patients (raw)	6–24 h	0.850	0.848 (vitals only)
Student clicks	14–60 d	0.660	0.540 (totals)
Wind turbines	14–60 d	unmeasurable at 20 units[b]	

Table 2: Identity A/A exposures (five-seed means; the ICU raw row is the imputation dissection of the LOCF row). [b]five-seed means range 0.747–0.948 and the pooled interval [0.617, 0.999] spans more than a third of the axis; the row decides nothing and is reported as such.

model, so that model and capacity are uninformative by construction, leaves 0.996 [0.992, 0.998]. Individual wear trajectories identify a drive among its manufacturing siblings, a control that closes the "you are recognizing drive models" objection by measurement.

Measurement rhythm alone identifies patients. Under the field's standard carry-forward imputation, ICU exposure is 0.993. Raw values with missingness intact read 0.850 [0.828, 0.871], consistent with physiological biometrics (Odinaka et al., 2012). Baselines are personal. Measurement rhythm alone, how often each channel is observed, with no physiological values, reads 0.879 [0.860, 0.897], comparable to the raw tier with overlapping intervals. That care patterns identify patients is known: laboratory orders without result values re-identify a patient among 100 candidates from a single visit at least a quarter of the time (Johnson et al., 2019). The measurement here differs in unit and purpose, per-channel observation counts in hourly ICU windows matched on time and offset, read as an evaluation precondition rather than a privacy attack. In this panel the same channels also tell time (0.634), so texture leaks twice, extending the documented finding that observation patterns encode the care process (Agniel et al., 2018).

Preprocessing manufactures exposure. Isolating the imputation's contribution requires a control that changes *only* the imputation. A naive comparison of the imputed tier against raw values alone also deletes the entire measurement-count family, the stronger identity channel of the two, and would attribute 0.14 to carry-forward imputation. The matched tier, raw values plus rhythm, differing from the imputed tier in imputation and nothing else, measures 0.916 [0.900, 0.930], so the isolated effect is about 0.08: roughly half the naive estimate is the deleted texture family, a confound invisible in a results table, which is why the matched comparison is the reported one. The effect is mechanism-specific: stale values shared as literal copies across a patient's windows add pipeline-induced record linkability, letting a classifier solve record linkage rather than recognize a body, while the same choice leaves the temporal floor essentially untouched. Preprocessing can move one axis and not the other, the argument for putting preprocessing inside the audit boundary.

A vessel is recognizable without its position. Vessel exposure is 0.955 [0.932, 0.973] on windows whose temporal floor is 0.527. Removing latitude and longitude leaves 0.886: speed profile and reporting rhythm identify the hull, and the rhythm alone reads 0.732, the counterpart of the ICU measurement-count result.

Small fleets fail loudly, which is the multi-seed rule earning its keep. Wind identity is unmeasurable at twenty turbines: the seed range and the pooled interval together span more than a third of the AUC axis (Table 2), and a single draw (our pilot read 0.917) is unrepresentative of its own seed family. The contrast with the same fleet's temporal row is informative. Pooled temporal scoring draws power from thousands of rows, while identity's effective sample size is the number of units.

6 A Floor Is a Property of Its Offset

Offset curves in two domains measure the full tier at a short, published, and long offset range. Drives read 0.525, 0.598, 0.734 at 7–29, 30–90, and 90–180 days; the ICU reads 0.592, 0.718, 0.881 at 3–6, 6–24, and 24–48 hours. Both curves are monotone and steep. The consequence: within a single domain the floor spans more than the entire cross-domain spread of Table 1, so a floor is a property of a dataset *jointly with a declared offset range*, and a warning claim must be read against the floor

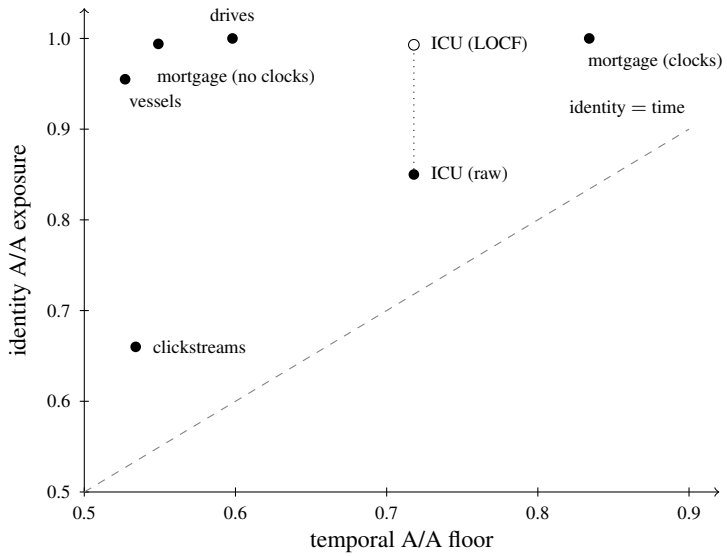

Figure 1: The readability plane. Each measured panel is a point (temporal floor, identity exposure); the ICU LOCF point (hollow) sits above the raw measurement, the mortgage pair spans the temporal axis at saturated exposure, and vessels sit beside mortgages without clocks. Every point lies above the dashed equal-reading line, but the ordering is claimed only in its weak form. The two axes are different tasks on different feature spaces (the identity reader receives both windows and a difference kernel), so part of every vertical gap is structural. Wind is omitted (identity is unmeasurable at twenty units). The plane is descriptive; taxonomy claims await a larger atlas.

at its own lead time. A model claiming 24-hour warning in the ICU should therefore be audited against temporal readability near 0.881, not only the mixture 0.718. The cross-unit placebo tracks both curves in their entirety (Appendix B), so the rises are population-level structure, stay phase in the ICU and fleet drift on the drives.

7 The Readability Plane

Measuring both axes on the same panels places each dataset at a point in the (floor, exposure) plane (Figure 1). On every panel where both are measurable, identity AUC exceeded temporal AUC. The two quantities are not directly comparable because their tasks and feature spaces differ, so the ordering is claimed only in its weak form, but their joint measurement shows that both nuisance signals can be substantial in retrospective evaluation. Position guides the audit. High-exposure, low-floor datasets (mortgages without clocks) motivate unit-disjoint evaluation when the target is unseen units; high-floor high-exposure datasets (ICU) motivate placement controls and patient-level splitting when evaluating generalization to new patients. No panel with two measurable axes is benign on both, although the small sample supports a descriptive finding rather than a general taxonomy.

8 From Readability to Protocol Inflation

Readability is availability, not use: the quantities above measure what a pipeline *could* exploit, not what any particular published detector did. A factorial experiment demonstrates the mechanism, by running a protocol we know to be vulnerable as a deliberate control arm against the same task under progressively stricter controls. The claim is narrow. When labels are unit-correlated, measured exposure predicts susceptibility to unit overlap. The task is unit-labeled case-control classification: every row of an event unit is labeled positive, every row of a sampled non-event unit negative, with no pre-event censoring or prediction horizon, so the experiment probes recognition leakage in the sampling design, not prospective warning skill. Wind and vessels lack such labels and are absent. We cross the feature tier (the domain's full versus reduced tier; for the ICU, the imputation pair

Domain	Tier	Exposure	Row-random	Unit-disjoint	Δ	Δ 95%
ICU	LOCF full	0.993	0.935	0.784	+0.151	[+0.131, +0.167]
ICU	raw, matched	0.916	0.776	0.743	+0.033	[+0.008, +0.045]
Mortgage	full	1.000	0.961	0.877	+0.084	[+0.057, +0.092]
Mortgage	no clocks	0.994	0.836	0.759	+0.077	[+0.043, +0.085]
Drives	full	1.000	0.925	0.861	+0.064	[+0.046, +0.075]
Drives	no clocks	0.998	0.886	0.822	+0.064	[+0.044, +0.080]
Clickstreams	full	0.660	0.587	0.581	+0.005	[-0.009, +0.016]
Clickstreams	totals	0.540	0.542	0.541	+0.001	[-0.009, +0.012]

Table 3: The bridge experiment: unit-labeled case-control AUC under a vulnerable versus a controlled split, beside the panel's measured identity exposure (Δ is a diagnostic contrast, not a decomposition). Δ averages three paired row/split seeds over one pool; the final column is the 2.5–97.5 percentile interval over 200 replicates that redraw the unit pool at the same size (Appendix D), so it carries pool, row-sampling, split, and fitting variability together. Δ is computed on one fixed pool and the interval on redrawn ones, so Δ need not sit centrally within its interval; the mortgage full cell lies near its upper bound, meaning that pool was a favorable draw for inflation. The saturated cells clear zero and the moderate-exposure clickstream cells do not.

differing in carry-forward imputation alone) with a grid of protocols formed by two independent controls. The *identity* control is either absent or unit-disjoint, holding every unit to one side. The *time* control is absent, a calendar split holding out the tail of the record, or a calendar split with an embargo: training ends at the 70th percentile of the span, the next fifth is discarded, and the test period is the final tenth. The embargo matters because under a plain split the first test rows sit immediately after the last training rows and are nearly unchanged from them, so a unit crosses the boundary by adjacency alone. The harness reader is trained in every cell, and the inflation statistic is the diagnostic contrast $\Delta = \mathrm{AUC}_{\text{leaky}} - \mathrm{AUC}_{\text{unit-disjoint}}$ between two protocols that share a time control and differ only in whether units persist: Δ against row-random, Δ_t under a calendar split, Δ_g under the embargo.

Three readings (Table 3). First, a coarse separation, not a monotone ordering. Calling a cell *saturated* when its exposure is at least 0.99, all 5 such cells inflate by +0.064 or more; the one intermediate cell (the ICU imputation-matched control, exposure 0.916) inflates +0.033; and the moderate-exposure clickstream cells sit at +0.005 and +0.001 with pool intervals covering zero. Within the saturated group, Δ does not track exposure rank (the ICU inflates most at the lowest saturated exposure), so exposure separates vulnerable from resistant designs without calibrating the inflation. Second, the ICU imputation pair: +0.151 under carry-forward imputation against +0.033 for the control differing in imputation alone, a 0.118 difference in inflation from a preprocessing choice: manufactured exposure is exploited, not merely available. Third, every unit-disjoint AUC stays above 0.5; because labels are unit-level, this shows persistent between-class separability under the controlled protocol, not that prospective signal survives; a pre-event-censored, fixed-horizon variant is natural follow-up work.

8.1 Holding Out the Future Does Not Hold Out the Unit

A temporal holdout is the field's usual answer to leakage, and it is often assumed to subsume the identity control: if the test period is later than the training period, the reasoning goes, nothing can be memorized. That reasoning is sound only when units do not survive the boundary. They usually do. At the 80th-percentile cut, 95% of drives, 75% of loans, 64% of learners and 52% of stays have rows on both sides. A drive present in training is the same physical drive in test, with the same wear signature, and its failure is scored as a prediction.

Table 4 separates the two controls on the two panels whose clock is a wall clock. Elsewhere the panel's time column is anchored per unit rather than to a calendar (ICU rows are indexed by hours since admission, clickstream rows by days since the module opened), so its tail is late-in-the-stay rather than later-in-history; those panels admit no out-of-time split at all, and the domain whose literature leans hardest on temporal validation is among them.

		Calendar split			With embargo		
Domain	Tier	Time	Unit+time	Δ_t	Time	Unit+time	Δ_g
Mortgage	full	0.915	0.890	+0.025	0.775	0.747	+0.028
Mortgage	no clocks	0.801	0.740	+0.061	0.688	0.650	+0.038
Drives	full	0.869	0.829	+0.039	0.864	0.854	+0.010
Drives	no clocks	0.855	0.826	+0.029	0.843	0.827	+0.015

Table 4: Identity inflation under two time controls, wall-clock panels. Each Δ compares two protocols sharing a time control and differing only in whether units persist across the boundary, yielding a diagnostic contrast between evaluation designs. This is not a causal decomposition of identity effects. *Calendar split* holds out the last fifth of the record; *with embargo* trains on the first 70%, discards the next 20%, and tests on the final 10%.

Two readings. First, a calendar split is a partial control, never a complete one. Δ_t is positive in all four cells with seed ranges excluding zero, so identity inflation survives holding out the future everywhere we can measure it; how much of the row-random inflation the split removes is not predictable from the domain, ranging from most of it (mortgages with clocks, +0.084 to +0.025) to little of it (mortgages without, +0.077 to +0.061).

Second, distance helps unevenly and costs accuracy. The embargo lowers absolute performance in these case-control evaluations (mortgages 0.915 to 0.775, drives 0.869 to 0.864). Its effect on identity inflation is domain-specific rather than uniform: on mortgages Δ_g stays clearly positive (+0.028, +0.038), while on drives it shrinks toward zero (+0.010) and in the no-clocks tier its seed range spans zero [-0.013, +0.043], which we report as deciding nothing for that cell. Pushing the test period further out can reduce identity inflation, then, but does not reliably remove it, and a protocol cannot assume it has. The two controls are separate instruments, and applying one does not earn the other's guarantee.

9 FINDINGS

Readability varies widely and resists domain intuition. Floors range from exactly 0.500 to 0.834 and exposures from unmeasurable to saturation, and cannot be inferred reliably from domain labels alone: the highest structural floor arises where stays are individually indexed, and the lowest floors arise where course calendars and seasonality seem strongest. Every within-unit warning evaluation that does not measure or control temporal readability effectively assumes it is negligible, and a unit-correlated evaluation that does not control identity assumes the same of exposure.

Attribution differs in kind, not degree. The mortgage floor is clock arithmetic, delete three columns and it vanishes, and the placebo shows its clock is *loan-anchored*: cross-unit pairs cannot read amortization levels, making it the one measured cell where the floor clearly exceeds its placebo (Appendix B). The ICU floor survives every deletion, and the placebo locates both it and the drive floor at the population level, shared clocks rather than individual trajectories. Drive exposure survives clock removal and cohort matching (wear states are individually unique); mortgage exposure survives for the inverse reason (levels constant within a loan differ across loans); ICU exposure combines a genuine physiological signature with an imputation-manufactured contribution; vessel exposure survives removing position (0.886). A clock floor yields to feature hygiene; a structural floor can only be read comparatively; saturated exposure motivates unit-disjoint evaluation for unseen units, and a temporal holdout does not supply it (Section 8.1).

The verdicts survive a change of reader. Under a linear reader every floor shifts down or within noise and no verdict changes; calibration cells are reader-invariant to the fourth decimal; structural floors drop by about 0.08, telling us their structure is substantially nonlinear (Appendix C, which also reports a foundation-model pilot that reads *less* than the from-scratch tree on identical input).

Implication for the warning literatures. In the features used by within-stay sepsis evaluations, temporal position is readable at 0.718 and patient identity at 0.850 without a deterioration label; in drive data, 0.598 and 1.000. Neither literature reports such measurements, and the harness makes reporting both cheap.

10 Related Work

Both instantiations operationalize negative-control reasoning (Lipsitch et al., 2010), and each axis has a long-standing mechanical ancestor. Along time, the case-crossover design and its time-trend repairs (Maclure, 1991; Suissa, 1995), placebo tests (Abadie et al., 2015), and classifier two-sample tests (Lopez-Paz & Oquab, 2017; Clémençon et al., 2009). Along identity, biometric verification (Jain et al., 2004; Odinaka et al., 2012) and the privacy literature's fingerprintability results (de Montjoye et al., 2013; Rocher et al., 2019), including re-identification of patients from laboratory orders alone (Johnson et al., 2019). This paper combines these established tools rather than replacing them, placing both axes of one panel under one forced-null harness with controls targeting the other axis.

Leakage taxonomies (Kaufman et al., 2012; Kapoor & Narayanan, 2023), reporting standards (Kapoor et al., 2024), and meta-reviews of evaluation failure (Liao et al., 2021), and a catalog of forecasting pitfalls (Hewamalage et al., 2023) prescribe against this failure class, which asks a reader to trust a protocol or audit it by hand. The quantities here are its dataset-side preconditions, measured before any event-prediction model is fit. Where unit identity crossing a split has inflated published results (Wen et al., 2020) or, in controlled experiments, drastically inflated performance (Rosenblatt et al., 2024), the exposure was a property of the corpus before any of those models were fit. Appendix H gives the full taxonomy. (Concurrent work develops protocol-level audit instruments for the same failure class; neither paper depends on the other.)

11 Limitations

Quantities measured at different offset ranges are different measurements, and Section 6 shows the dependence dominates cross-domain differences, so the tables are a survey of six declarations, not a ranking of six datasets; offset normalization remains future work and would have to be curve-level. Pairing also selects a population: each number describes units observable at both endpoints (retention ships with every result and exceeds 95% of ordinary units in every large panel except vessels, 4,628 of 6,248). Both quantities are scorer-dependent lower bounds. The identity design cancels shared environment by construction, so exposures are lower bounds on what an unmatched adversary could read; the match is on the panel's time coordinate, and relative stay time is not a match on admission date or hospital site. The plane comparison is structurally uneven (different tasks, different feature spaces) and is claimed only in weak form. The temporal floor is blind to placement and identity artifacts, the exposure is blind to phase, and the two together do not audit a protocol; they characterize the data it runs on. Mechanisms acting through neither axis, such as resampling before the split (Vandewiele et al., 2021) or model selection on the evaluation data (Neunhoeffer & Sternberg, 2019), are outside both. The bridge uses unit-level case-control labels with no prediction horizon, so it demonstrates the mechanism and scores no published predictor; and unit overlap is not inherently invalid when deployment targets future observations of known units and training uses only information available at the prediction time. Six domains are six points; the adapter contract exists so the atlas can grow.

12 Conclusion

One forced-null construction turns two unstated assumptions of retrospective evaluation, ordinary data carries no time, ordinary data carries no identity, into measured numbers. Across six domains both assumptions fail by amounts that vary substantially across domains and on axes that fail differently in kind, and the bridge shows the mechanism by which the second failure becomes inflation under unit-correlated case-control sampling. These measurements establish susceptibility under the tested sampling designs; they do not determine whether a particular predictor exploits time or identity. We recommend reporting both diagnostics alongside warning-model results to motivate controls appropriate to the intended deployment and prediction task.

AI Use Statement

Large language models were used to retrieve and discover related literature, support research ideation and hypothesis refinement, assist with methodology and experiment design, and aid software implementation (including the code that synthesizes the released demonstration's panel from

a random seed; no model produced any data), mathematical formulation and proof assistance for Proposition 1 and its proof in Appendix A, interpretation of results, and manuscript drafting and editing, including drafting sections and polishing their writing. Empirical values are outputs of executed experiments, with the synthetic demonstration distinguished from the public-data measurements. Harness values pass through generated macros and transcription audits; the original drive comparison is stored as an attributed constant. These audits check consistency with stored results, not the validity or provenance of the experiments themselves. The research questions and study design were chosen, the experiments executed and audited, the proof checked, and the reported results and citations verified by the human authorship of this paper, which takes responsibility for its contents.

Ethics Statement

The identity A/A exposure quantifies how re-identifiable units are from ordinary operational logs, which is dual-use knowledge. Three design choices bound the risk. All measurements use public, de-identified datasets under their published terms (PhysioNet/CinC 2019, Backblaze drive stats, Freddie Mac loan-level samples, OULAD, Kelmarsh and Penmanshiel SCADA, and the casualty-free normal pool of a public linked AIS corpus), and no measurement links any window to an external identity. The construction only asks whether two windows *within* a dataset share an anonymous unit label, a quantity the privacy literature already studies per-domain with stronger attack models than ours. Second, the exposure is reported as an aggregate AUC, not as per-unit re-identifications. Third, the findings favor defenders. They argue for unit-disjoint evaluation, identify carry-forward imputation as a pipeline choice that manufactures linkability clinicians can avoid, and give data stewards a cheap instrument for measuring exposure before release. We believe measuring and publishing these quantities reduces net risk relative to the status quo, in which the exposure exists, adversaries can measure it, and evaluators assume it away.

Reproducibility Statement

The supplementary material contains the complete harness (an installable package with invariant tests for the forced-null, split, placebo, and bridge constructions), panel adapters, declarations and stored results for the six-domain harness measurements, and the scripts that build the ICU and drive panels from the public raw downloads. The vessel adapter reads per-window summaries published with the linked maritime corpus, whose computation from raw tracks is not included. The original drive reimplementation is an imported value, not an experiment reproducible from this zip. Appendix G and the supplementary README state these boundaries. Sections 2 and 3 specify the construction and its preconditions; Appendix D validates the interval machinery. Re-running the whole pipeline reproduces the published values: on our machine 28 of the 30 measurement cells returned bit-identical per-seed AUCs and intervals, and two moved in the fourth decimal, which is gradient-boosted histogram reduction order under multithreading. Values are quoted to three decimals, so none is affected.

Every panel measured here needs a credentialed or multi-gigabyte download, so a reader who has just received the harness can read it but cannot run it. The package therefore also ships a self-contained demonstration, `aa-floor-demo`, that needs no download and no credentials. It synthesizes a panel with readability planted on known columns, runs both instruments on it through the same entry points used for the six-domain audit measurements, and checks each tier against what was planted: each instrument must read its own axis above the null, read the other instrument's axis at the null, and read pure noise at the null. It prints a per-tier verdict and exits non-zero if any tier misses. Appendix G gives the construction.

References

Alberto Abadie, Alexis Diamond, and Jens Hainmueller. Comparative politics and the synthetic control method. *American Journal of Political Science*, 59(2):495–510, 2015.

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

Backblaze, Inc. Hard drive reliability test data. https://www.backblaze.com/cloud-storage/resources/hard-drive-test-data, 2026. Daily drive snapshots released quarterly, accessed August 2026.

Elias Chaibub Neto, Abhishek Pratap, Thanneer M. Perumal, Meghasyam Tummalacherla, Phil Snyder, Brian M. Bot, Andrew D. Trister, Stephen H. Friend, Lara Mangravite, and Larsson Omberg. Detecting the impact of subject characteristics on machine learning-based diagnostic applications. *npj Digital Medicine*, 2:99, 2019.

Stéphan Clémençon, Marine Depecker, and Nicolas Vayatis. AUC optimization and the two-sample problem. In *Advances in Neural Information Processing Systems 22*, pp. 360–368, 2009.

Yves-Alexandre de Montjoye, César A. Hidalgo, Michel Verleysen, and Vincent D. Blondel. Unique in the crowd: the privacy bounds of human mobility. *Scientific Reports*, 3:1376, 2013.

Peter Eckersley. How unique is your web browser? In *Privacy Enhancing Technologies (PETS 2010)*, volume 6205 of *Lecture Notes in Computer Science*, pp. 1–18. Springer, 2010.

Freddie Mac. Single family loan-level dataset. https://www.freddiemac.com/research/datasets/sf-loanlevel-dataset, 2026. Sample acquisition and monthly performance files, accessed August 2026.

Igor Goldenberg and Geoffrey I. Webb. Survey of distance measures for quantifying concept drift and shift in numeric data. *Knowledge and Information Systems*, 60(2):591–615, 2019.

Sander Greenland. Confounding and exposure trends in case-crossover and case-time-control designs. *Epidemiology*, 7(3):231–239, 1996.

Andrew Hendel. A deterministically linked corpus of U.S. Coast Guard marine casualty investigations, public AIS vessel tracks, and vessel regulatory history, 2024–2025. Zenodo, 2026. https://doi.org/10.5281/zenodo.21743546.

Hansika Hewamalage, Klaus Ackermann, and Christoph Bergmeir. Forecast evaluation for data scientists: common pitfalls and best practices. *Data Mining and Knowledge Discovery*, 37(2): 788–832, 2023.

Anil K. Jain, Arun Ross, and Salil Prabhakar. An introduction to biometric recognition. *IEEE Transactions on Circuits and Systems for Video Technology*, 14(1):4–20, 2004.

Kipp W. Johnson, Jessica K. De Freitas, Benjamin S. Glicksberg, Jason R. Bobe, and Joel T. Dudley. Evaluation of patient re-identification using laboratory test orders and mitigation via latent space variables. In *Pacific Symposium on Biocomputing*, volume 24, pp. 415–426, 2019.

Sayash Kapoor and Arvind Narayanan. Leakage and the reproducibility crisis in machine-learning-based science. *Patterns*, 4(9):100804, 2023.

Sayash Kapoor, Emily M. Cantrell, Kenny Peng, Thanh Hien Pham, Christopher A. Bail, Odd Erik Gundersen, Jake M. Hofman, Jessica Hullman, Michael A. Lones, Momin M. Malik, Priyanka Nanayakkara, Russell A. Poldrack, Inioluwa Deborah Raji, Michael Roberts, Matthew J. Salganik, Marta Serra-Garcia, Brandon M. Stewart, Gilles Vandewiele, and Arvind Narayanan. REFORMS: consensus-based recommendations for machine-learning-based science. *Science Advances*, 10(18):eadk3452, 2024.

Shachar Kaufman, Saharon Rosset, Claudia Perlich, and Ori Stitelman. Leakage in data mining: formulation, detection, and avoidance. *ACM Transactions on Knowledge Discovery from Data*, 6 (4):15:1–15:21, 2012.

Jakub Kuzilek, Martin Hlosta, and Zdenek Zdrahal. Open university learning analytics dataset. *Scientific Data*, 4:170171, 2017. doi: 10.1038/sdata.2017.171.

Thomas Liao, Rohan Taori, Inioluwa Deborah Raji, and Ludwig Schmidt. Are we learning yet? A meta-review of evaluation failures across machine learning. In *Advances in Neural Information Processing Systems, Datasets and Benchmarks Track*, 2021.

Marc Lipsitch, Eric Tchetgen Tchetgen, and Ted Cohen. Negative controls: a tool for detecting confounding and bias in observational studies. *Epidemiology*, 21(3):383–388, 2010.

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

Malcolm Maclure. The case-crossover design: a method for studying transient effects on the risk of acute events. *American Journal of Epidemiology*, 133(2):144–153, 1991.

Ishan Misra, C. Lawrence Zitnick, and Martial Hebert. Shuffle and learn: unsupervised learning using temporal order verification. In *European Conference on Computer Vision*, pp. 527–544, 2016.

Fabian Monrose and Aviel D. Rubin. Keystroke dynamics as a biometric for authentication. *Future Generation Computer Systems*, 16(4):351–359, 2000.

Marcel Neunhoeffer and Sebastian Sternberg. How cross-validation can go wrong and what to do about it. *Political Analysis*, 27(1):101–106, 2019.

Ikenna Odinaka, Po-Hsiang Lai, Alan D. Kaplan, Joseph A. O'Sullivan, Erik J. Sirevaag, and John W. Rohrbaugh. ECG biometric recognition: a comparative analysis. *IEEE Transactions on Information Forensics and Security*, 7(6):1812–1824, 2012.

Charlie Plumley. Kelmarsh wind farm data. Zenodo, 2023a. https://doi.org/10.5281/zenodo.8252025.

Charlie Plumley. Penmanshiel wind farm data. Zenodo, 2023b. https://doi.org/10.5281/zenodo.8253010.

Matthew A. Reyna, Christopher S. Josef, Russell Jeter, Supreeth P. Shashikumar, M. Brandon Westover, Shamim Nemati, Gari D. Clifford, and Ashish Sharma. Early prediction of sepsis from clinical data: The PhysioNet/computing in cardiology challenge 2019. *Critical Care Medicine*, 48(2):210–217, 2020. doi: 10.1097/CCM.0000000000004145.

Michael Roberts, Derek Driggs, Matthew Thorpe, Julian Gilbey, Michael Yeung, Stephan Ursprung, Angelica I. Aviles-Rivero, Christian Etmann, Cathal McCague, Lucian Beer, Jonathan R. Weir-McCall, Zhongzhao Teng, Effrossyni Gkrania-Klotsas, AIX-COVNET, James H. F. Rudd, Evis Sala, and Carola-Bibiane Schönlieb. Common pitfalls and recommendations for using machine learning to detect and prognosticate for COVID-19 using chest radiographs and CT scans. *Nature Machine Intelligence*, 3(3):199–217, 2021.

Luc Rocher, Julien M. Hendrickx, and Yves-Alexandre de Montjoye. Estimating the success of re-identifications in incomplete datasets using generative models. *Nature Communications*, 10:3069, 2019.

Matthew Rosenblatt, Link Tejavibulya, Rongtao Jiang, Stephanie Noble, and Dustin Scheinost. Data leakage inflates prediction performance in connectome-based machine learning models. *Nature Communications*, 15:1829, 2024.

Samy Suissa. The case-time-control design. *Epidemiology*, 6(3):248–253, 1995.

Iulian Emil Tampu, Anders Eklund, and Neda Haj-Hosseini. Inflation of test accuracy due to data leakage in deep learning-based classification of OCT images. *Scientific Data*, 9:580, 2022.

Gilles Vandewiele, Isabelle Dehaene, György Kovács, Lucas Sterckx, Olivier Janssens, Femke Ongenae, Femke De Backere, Filip De Turck, Kristien Roelens, Johan Decruyenaere, Sofie Van Hoecke, and Thomas Demeester. Overly optimistic prediction results on imbalanced data: a case study of flaws and benefits when applying over-sampling. *Artificial Intelligence in Medicine*, 111:101987, 2021.

Daniel Vela, Andrew Sharp, Richard Zhang, Trang Nguyen, An Hoang, and Oleg S. Pianykh. Temporal quality degradation in AI models. *Scientific Reports*, 12:11654, 2022.

Vladimir Vovk, Ilia Nouretdinov, and Alex Gammerman. Testing exchangeability on-line. In *International Conference on Machine Learning*, pp. 768–775, 2003.

Donglai Wei, Joseph J. Lim, Andrew Zisserman, and William T. Freeman. Learning and using the arrow of time. In *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 8052–8060, 2018.

Junhao Wen, Elina Thibeau-Sutre, Mauricio Diaz-Melo, Jorge Samper-González, Alexandre Routier, Simona Bottani, Didier Dormont, Stanley Durrleman, Ninon Burgos, and Olivier Colliot. Convolutional neural networks for classification of Alzheimer's disease: overview and reproducible evaluation. *Medical Image Analysis*, 63:101694, 2020.

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

Zhirong Wu, Yuanjun Xiong, Stella X. Yu, and Dahua Lin. Unsupervised feature learning via non-parametric instance discrimination. In *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3733–3742, 2018.

A Proof of the Forced Null

Proof of Proposition 1. Write the pooled empirical AUC with the mid-rank tie correction as

$$\widehat{\mathrm{AUC}} = \frac{1}{m^2} \sum_{k=1}^{m} \sum_{l=1}^{m} \Big[\mathbf{1}\{s(U_k) > s(V_l)\} + \tfrac{1}{2}\mathbf{1}\{s(U_k) = s(V_l)\} \Big].$$

Let σ exchange U_k with V_k for every k at once. By hypothesis σ preserves the joint law of the test sample, and s is held fixed, so $\widehat{\mathrm{AUC}} \circ \sigma$ has the same distribution as $\widehat{\mathrm{AUC}}$. Relabelling the summation indices, the (k, l) term of $\widehat{\mathrm{AUC}} \circ \sigma$ pairs with the (k, l) term of $\widehat{\mathrm{AUC}}$ to give $\mathbf{1}\{a > b\} + \mathbf{1}\{b > a\} + \mathbf{1}\{a = b\} = 1$ with $a = s(U_k)$ and $b = s(V_l)$, so $\widehat{\mathrm{AUC}} \circ \sigma = 1 - \widehat{\mathrm{AUC}}$ pointwise. Taking expectations, $\mathbb{E}[\widehat{\mathrm{AUC}}] = \mathbb{E}[1 - \widehat{\mathrm{AUC}}]$, and the claim follows. □

Why the ordered pair and not the marginals. Equal marginals do not force the null, because the two items of a pair are dependent and the within-pair comparisons enter the statistic. The temporal null is therefore exchangeability of the ordered window pair, which is strictly stronger than stationarity of $X_{i,t}$ and is the sense in which the paper uses the word throughout. The identity null is already in this form, the same-unit and different-unit continuations of a shared anchor being exchangeable once calendar and offset are matched.

Why the split is load-bearing. A scorer fit on the rows it later scores is not independent of their labels, the involution argument fails, and the measured excess would confound memorization of those rows with readability of the panel. Unit-disjoint evaluation is what makes the proposition apply, not a hygiene convention layered on top of it.

What the statement fixes. *Randomness.* The expectation is over the draw of the test pairs, with the training units, the fitted scorer, and the scorer's own seed held fixed by conditioning. The published values additionally average over pair-draw seeds and the intervals resample units; both sit outside the proposition and are reported separately. *Independence.* The scorer must not depend on the held-out labels. The unit-disjoint split supplies this, and in the identity instantiation the units are split before any pair is drawn, so no unit contributes to both fitting and scoring. *Ties.* Gradient-boosted ensembles produce ties in quantity, so the mid-rank convention is not cosmetic. The involution needs the $\frac{1}{2}$ weight for the three indicators to sum to one, and it is the convention the reported AUCs already use. *Scope of "every classifier."* Every scorer measurable in the features and independent of the held-out labels, which covers any model fit on the training units by any procedure. It excludes a procedure that reads the held-out labels, which is the failure mode the design exists to prevent. *Finite sample.* The conclusion concerns the expected empirical AUC at the realized m and not a population

limit, so the bootstrap width and the seed spread quoted throughout are dispersion around a mean the null pins at exactly one half.

Where the released construction departs from the idealization. The proposition assumes every anchor yields a matched pair. In the identity instantiation the units are split before any pair is drawn and the impostor is drawn from the anchor's own side, so the split discards nothing; the one residue is an anchor whose matched time offers no impostor on its side, which emits its same-unit row unmatched. The proposition applies to the matched subsample under its stated joint-invariance assumption; the unmatched residue requires a separate qualification. Class imbalance alone does not establish AUC bias. The harness records the per-side class counts in every results file, and Table 5 replays the published seeds and reports them: the test sides are balanced to within the residue, which is zero on every large panel, one or two rows under cohort-matched impostors, and at most fifteen on the twenty-turbine wind panel, where a matched time sometimes offers no impostor among four test units. The temporal instantiation has no analogous residue, since both windows of a pair come from the same unit by construction. Neither instantiation assumes the null is true. The proposition says what the reading would be if it were, which is what licenses reading the excess as readability.

Panel	Train same	Test same	Test different	Test unmatched
Drives	8,000	2,000	2,000	0
Drives (cohort matched)	8,000–8,002	2,001	1,999–2,000	1–2
Mortgage	8,000	2,000	2,000	0
Vessels	3,200	800	800	0
Clickstreams	8,000	2,000	2,000	0
ICU (LOCF)	8,000	2,000	2,000	0
ICU (raw)	8,000	2,000	2,000	0
Wind	251–335	58–137	58–122	0–15

Table 5: Identity matching diagnostics, ranges across the five published seeds, replayed from the stored seeds without refitting. Train same is the number of same-unit training pairs; test same and test different are the two test classes; test unmatched counts same-unit test rows whose matched time offered no impostor on the test side. Per-seed counts ship in `results/identity_balance.json` and in each results file.

B The Cross-Unit Placebo. Population Structure versus Within-Unit Readability

Under pooled scoring, label-1 rows sit on average one offset later than label-0 rows on whatever clock the panel keeps, so structure shared across the population at matched clock positions is readable by a classifier that has learned nothing about any individual unit. The control is a cross-unit placebo. Draw the anchor (x_i, t_1) and the offset exactly as the real construction does, but take the second window from a *different* unit observed at $t_2 = t_1 + \delta$, label earlier/later, and score pooled as usual. The pairing carries no within-unit information, so whatever it reads is population structure. We compare its AUC with the within-unit pairing AUC. Their difference is a diagnostic contrast, not an additive statistical decomposition; AUC is nonlinear, and the contrast may be negative.

ICU. The placebo tracks the floor at all three offsets (Table 6), contrasts of 0.005–0.021. Although stays are individually indexed from admission, hours-since-admission is itself a shared clock, and features that evolve with time-in-stay population-wide (the admission burst of measurement counts, the fill-in density of carried-forward labs) order rows by stay position across patients: the ICU cells and their steep offset curve measure shared stay-phase structure. The floor's audit role, the tier ladder, and every identity result are unaffected (the identity design is time-and-offset matched by construction).

Drives. The placebo rises with offset and tracks the floor throughout, exceeding it at 7–29 days (0.544 against 0.525) and sitting within 0.019 at 30–180 days. The drive floor and its offset curve read as population-level calendar drift (fleet turnover, model mix, firmware waves, seasonality) rather than within-unit aging.

Mortgages, clickstreams, wind, vessels. Clickstream, wind, and vessel placebos track or slightly exceed their floors (course calendar and season are shared; the vessel placebo reads 0.546 against a

Domain	Offsets	Floor	Placebo	Difference
Drives	7–29 d	0.525	0.544 [0.521, 0.565]	-0.020
Drives	30–90 d	0.598	0.592 [0.579, 0.605]	0.006
Drives	90–180 d	0.734	0.715 [0.702, 0.727]	0.019
ICU	3–6 h	0.592	0.586 [0.570, 0.602]	0.005
ICU	6–24 h	0.718	0.697 [0.684, 0.708]	0.021
ICU	24–48 h	0.881	0.872 [0.857, 0.891]	0.009
Mortgage (clocks)	6–18 mo	0.834	0.736 [0.718, 0.753]	0.098
Mortgage (no clocks)	6–18 mo	0.549	0.529 [0.516, 0.542]	0.020
Mortgage (rate only)	6–18 mo	0.500	0.507 [0.490, 0.523]	-0.007
Clickstreams	14–60 d	0.534	0.555 [0.542, 0.567]	-0.021
Wind	14–60 d	0.539	0.563 [0.529, 0.599]	-0.024
Vessels	14–60 d	0.527	0.546 [0.510, 0.580]	-0.019

Table 6: Cross-unit placebo beside the matched within-unit floor (five-seed means, seed-0 unit-clustered bootstrap intervals on the placebo; full tier except the labeled mortgage variants). "Difference" is floor AUC minus placebo AUC, reported as a diagnostic contrast rather than an additive within-unit component. The placebo tracks or exceeds the floor everywhere except the mortgage clock tier, the one measured cell whose temporal structure is unit-anchored.

floor of 0.527), the mortgage no-clock placebo tracks its floor (the surviving signal is a population LTV index), and the rate-only placebo calibrates at 0.507. The mortgage *clock* tier is the exception: its placebo reads 0.736 against a floor of 0.834, a 0.098 shortfall, because amortization is a *loan-anchored* clock: balance and age levels order time within a loan but are incomparable across loans, so cross-unit pairs cannot read them. The placebo does not return "population" indiscriminately; it distinguishes unit-anchored clock structure from shared structure.

What the placebo changes, and what it does not. Outside the mortgage clock tier, the absolute AUC contrast between within-unit and placebo tasks is no larger than about 0.02 in any measured cell, so measured temporal readability in five of six domains is a property of the population's shared clocks rather than of individual trajectories. The floors remain diagnostic references for temporal readability without an event prediction target. The null is still forced at 0.5 under exchangeability, the excess is still real readability, and the placebo tells you which structure carries it.

C Scorer Sensitivity and a Foundation-Model Pilot

Full-tier temporal floors under the harness tree and under a linear reader (logistic regression, median-imputed, standardized; five seeds each). Mortgage 0.834 vs 0.821, ICU 0.718 vs 0.636, drives 0.598 vs 0.522, wind 0.539 vs 0.536, clickstreams 0.534 vs 0.542, vessels 0.527 vs 0.506. Calibration rows are reader-invariant to the fourth decimal; all shifts are downward or noise-level, consistent with floors as lower bounds; within-domain tier conclusions survive everywhere, and two reader-dependence findings are reported. The sepsis values and texture tiers exchange places under the linear reader, and mid-table domain order is reader-dependent.

One pilot measured whether a pretrained reader raises the floor (same populations, three cells). On sepsis pairs with a full 24-hour history and vitals only, a snapshot read by the harness tree (0.511 [0.495, 0.528]), the trailing 24-hour sequence read by the same tree (0.576 [0.559, 0.592]), and the same sequences through chronos-t5-small embeddings with a tree head (0.551 [0.535, 0.568]). Sequence history adds readable time (+0.064); the generic embedding reads less than the from-scratch tree on identical input, so the published floors' conservatism is not an artifact of declining a fashionable reader; and public datasets may appear in foundation pretraining corpora, so any FM cell is a knowledgeable-reader variant by construction.

D Interval Validation and Additional Controls

Checks on the measurement machinery. **Coverage.** the dyadic identity bootstrap was validated against a 200-repetition full-regeneration bootstrap on the two headline unsaturated cells, which resamples units, redraws pairs, refits, and rescores, and so prices every source at once. Regener-

ation intervals [0.639, 0.674] (clickstreams) and [0.836, 0.859] (ICU raw) sit inside the reported pooled intervals [0.624, 0.694] and [0.828, 0.871], with regeneration means within 0.004 of the five-seed means. Pooling widens the cheap interval against its primary-seed version ([0.624, 0.684] and [0.823, 0.865]), so the pooled intervals are wider than the regeneration intervals in these two checked cells. This comparison does not establish nominal coverage or conservativeness on other panels. **Bridge pool resampling.** Δ in Table 3 averages three row/split seeds over one fixed unit pool, so it carries no uncertainty from which units that pool contained. The interval column redraws the pool 200 times at the same size, redraws rows, reapplies both splits, refits, and takes the 2.5–97.5 percentile of Δ. This is not a restatement of seed variability: the pool interval is wider than the three-seed range in 8 of 8 cells, and for mortgage the whole seed range sits above the pool mean, so seeds agreeing tightly with each other on one pool is no evidence that Δ is pinned down. Same-size redraws are required rather than convenient: Δ is not a fixed parameter estimated at rate $1/\sqrt{n}$, because a smaller or unevenly weighted pool starves the unit-disjoint arm of units to generalize across while the row-random arm can still recognize units, so bootstrap and subsample replicates inflate Δ instead of describing its spread. Pool headroom differs by class, and a class no larger than the pool contributes no pool variability by construction; this is the case for the 2,932 ICU positives, which are the entire available population. Both tiers of a domain share a replicate's pool, so the ICU imputation contrast is paired: the gap is +0.122 [+0.107, +0.136], positive in 100% of replicates. We report that contrast as a difference rather than a ratio: the paired ratio has median 5.8 but interval [3.5, 18.0], because its denominator is a small Δ with spread of its own, so the ratio is well defined but poorly determined even where the difference is tight. **Cohort matching.** drive impostors restricted to the anchor's model leave exposure at 0.996; the equivalent test has not been run for clickstreams or ICU, and course cohort and hospital site remain live alternative explanations for part of those numbers. **Matched imputation control.** our first LOCF comparison changed two things at once (imputation and the deletion of the measurement-count family) and overstated the imputation effect as 0.14; the matched tier isolates it at 0.08. The error was invisible in a results table, which is why the matched tier is the reported one. **Retention.** the pairing rule narrows no domain by more than about a quarter, and eligible units exceed 95% of ordinary units in every large panel except vessels, whose weekly windows leave 4,628 of 6,248 with two windows 14–60 days apart.

E Domain Narratives

Each domain followed the same protocol. An adapter maps raw data to the panel contract, a declaration fixes offsets, tiers, and clocks, and the measurement runs.

E.1 Temporal floors

ICU vitals (PhysioNet 2019, 40,336 stays, hourly). The highest structural floor measured (the clock-driven mortgage full tier is larger), 0.718 [0.706, 0.730] on never-septic stays at 6–24 hour offsets. The ladder shows the structure is deep. values alone read 0.677, and measurement counts alone, with no physiology whatsoever, read 0.634. The cross-unit placebo (Appendix B) supports a shared stay-phase interpretation.

Hard-drive telemetry (Backblaze, 30–90 day offsets). The public drive-day logs come from Backblaze (Backblaze, Inc., 2026). 0.598 [seed range 0.594–0.603] with SMART's explicit clocks (power-on hours, total writes), 0.527 without them, temperature alone 0.507.

Wind SCADA (Kelmarsh and Penmanshiel, 20 turbines, daily). The public farm records are documented by Plumley (2023a;b). 0.539 [0.509, 0.580]. Weather decorrelates in days, so 14–60 day ordinary windows are nearly exchangeable, and even a literal energy-export counter dilutes under pooled scoring across twenty baselines.

Student clickstreams (OULAD, 29,228 registrations, active days). We use the Open University Learning Analytics Dataset (Kuzilek et al., 2017). 0.534. On days a student is active, the mix of activity is nearly stationary; the calendar governs which days are active more than what activity looks like. The floor is measured on active days and that scope is disclosed.

Mortgage performance (Freddie Mac samples, 162,490 loans, monthly). We use the public Single-Family Loan-Level Dataset samples (Freddie Mac, 2026). The widest spread in the atlas. 0.834 [0.826, 0.842] with amortization clocks (a vintage's loans amortize in near lockstep, making

age a shared calendar), 0.549 [0.536, 0.561] without them (the surviving signal is carried by the quarterly-refreshed LTV index), and the loan's own rate alone exactly 0.500.

Vessel tracks (linked maritime corpus, 6,248 casualty-free vessels, weekly windows). The panel is the normal pool of a public corpus that links casualty investigations to AIS tracks (Hendel, 2026): every vessel with no casualty on record, one ordinary window about every seven days across two years, each window summarized into anchor kinematics, recent and voyage-scale history statistics, AIS reporting texture, and anchor position. The corpus's own casualty-vessel experiment uses event-relative windows and is not quoted here. 0.527 [0.496, 0.558] at 14–60 day offsets, 0.517 without position, kinematics alone 0.505, reporting rhythm alone 0.506, which is below the range recorded for it before the run (we expected fleet-wide AIS coverage drift to act as a shared clock, and it does not). The lowest full-tier floor in the atlas; the placebo exceeds it (Appendix B).

E.2 Identity exposures

Hard drives. 1.000 [0.998, 1.000] full, 0.998 without declared clocks (remaining cumulative error counters are disclosed as quasi-clocks), 0.996 under cohort-matched impostors. Drives are re-identifiable from windows a month to a season apart, individually, among their own model siblings.

Mortgage loans. 1.000 [0.998, 1.000] full and 0.994 without amortization clocks. Rate and LTV levels differ across loans even though constant within, so the feature set that is temporally near-exchangeable (floor 0.549) is identity-saturated. One panel spans the plane.

ICU patients. 0.993 [0.989, 0.995] under carry-forward imputation; raw values with missingness intact 0.850 [0.828, 0.871] (vitals alone 0.848); measurement rhythm alone 0.879; the matched raw-plus-rhythm tier 0.916. The three channels are partly complementary, and no single one explains the saturated imputed cell.

Student clickstreams. 0.660 [0.624, 0.694]. Activity-type mixes overlap more than anticipated, and totals alone carry little (0.540).

Vessel tracks. 0.955 [0.932, 0.973] full, 0.886 [0.851, 0.919] without position, reporting rhythm alone 0.732. Test sides balanced at 800 same-unit against 800 different-unit pairs in every seed with no unmatched row, since about eighty vessels are observed on any calendar day.

Wind turbines. Unmeasurable at 20 units (Section 5).

F Boundary Conditions of the Measurements

The eligibility requirements and reporting rules of Sections 2 and 3 each have an empirical basis, summarized here.

One unit is disqualifying (eligibility item 2). MetroPT, a predictive-maintenance dataset comprising a single compressor unit on one metro train, admits no unit-disjoint split and no cluster to bootstrap; neither quantity is measurable on it.

Repeated observations are the one hard requirement (eligibility item 1). The Scania APS dataset records one aggregated snapshot per truck; identified units without a time axis yield no pairs.

Small fleets are measurable on the temporal axis and not on the identity axis (eligibility item 2). Twenty turbines produce a temporal floor whose unit-clustered interval spans nearly 0.08, wide but usable, and an identity row whose five-seed range spans chance to near-perfect, which decides nothing. The asymmetry has a reason: pooled temporal scoring draws power from thousands of rows, while identity's effective sample size is the number of units.

Published values are multi-seed means (reporting rule). An independent reimplementation of the drive-telemetry measurement, identical protocol, fresh random pair draw, reproduces the tier ladder but shifts each rung by up to about 0.02, more than the $\sim$0.01 spanned by within-implementation seed ranges: a single pair draw under-reports uncertainty, and the wind identity row shows the failure at its largest.

Pooled, never paired (scoring rule). Under paired scoring a monotone counter such as power-on hours orders any within-unit pair by construction and the test is degenerate; under pooled scoring,

between-unit variance dilutes clocks and the measurement is informative (the drive floor reads 0.580, not the saturation paired scoring would force).

Cohort holdouts (protocol grid). A third identity control, holding out whole cohorts so that neither a unit nor its siblings appear in training, was measured on the two panels carrying a cohort label (drive model; loan vintage) and is reported as deciding nothing. Of 12 cells, only 3 have a seed range excluding zero, and the widest spans 0.436. The instability is structural rather than a sample-size accident: cohorts differ by orders of magnitude in membership, so which cohorts a seed draws changes the test population itself, and on the mortgage panel vintage is close to collinear with the calendar, so the cohort and time controls are not separable there. A cohort holdout is also the one arm whose contrast is not a leakage measure, since cohorts genuinely differ and part of any drop is real heterogeneity. We report the grid without it and note that a panel with many comparable cohorts would support the arm.

G HARNESS

The released harness measures both quantities, and the placebo, for any eligible panel from one parquet and one JSON declaration; full results (per-seed values, both interval sources) ship in the supplementary material, and harness values are generated from those results files. The original drive comparison is stored as an attributed constant in the number generator.

Running the instruments without any of these panels. `aa-floor-demo` synthesizes a panel in which the two axes are planted on disjoint columns and checks that each instrument recovers its own and only its own. A counter that advances with the calendar and a population-level trend shared by every unit carry time and no identity; per-unit constant offsets plus noise carry identity and are flat in time, so they cannot order a unit's own earlier and later windows; a third group is iid noise and carries neither. The demonstration then runs the published `measure` and `measure_identity` entry points, unmodified, and reports a verdict per tier.

The cross terms are the substance of the check. A temporal floor that rose on the per-unit offsets would be reporting identity as time, and an identity exposure that rose on the shared clocks would be reporting the calendar as fingerprinting; both are the confusions the two constructions are designed to prevent, and both would be invisible in a demonstration that only planted signal where it expected to find it. Verdicts are read off the seed mean rather than off any single seed, using the half-width of the harness's own bootstrap interval as the scale of the noise, because a single seed's interval conditions on the pairs that seed drew and clips the null often enough to report a spurious defect on a routine run. The synthetic panel omits one feature of real data, staggered unit entry, which would make age itself partly identity-readable at matched calendar time; the demonstration is a check on the instruments, not a miniature of any domain. Its outputs are an ordinary parquet and two ordinary declarations, so the published command-line entry points run on them unchanged.

H RELATED WORK IN FULL

Section 10 states the positioning; this appendix gives the taxonomy behind it and the boundary of what the two quantities cover.

Temporal ancestors. The case-crossover design and its time-trend repairs (Maclure, 1991; Suissa, 1995; Greenland, 1996) self-match a unit to cancel fixed confounders, and the repairs exist precisely because time trends survive that cancellation. Placebo tests in synthetic-control studies (Abadie et al., 2015) apply a treatment where none occurred and read the estimate as a null distribution. Classifier two-sample tests (Lopez-Paz & Oquab, 2017; Clémençon et al., 2009), known to practice as adversarial validation, supply the estimator and the symmetry argument. Concept-drift magnitudes (Goldenberg & Webb, 2019) measure population-level change with no null guarantee; the floor is the unit-paired, offset-resolved counterpart. Temporal model degradation (Vela et al., 2022) measures the consequence downstream of a fitted model. Sequential exchangeability testing (Vovk et al., 2003) answers whether exchangeability holds; the floor answers how much it fails by and from which features.

Identity ancestors. Biometric verification (Jain et al., 2004; Odinaka et al., 2012; Monrose & Rubin, 2000) supplies the mechanical ancestor, transplanted here to logs never designed as biometrics. The privacy literature documents per-domain fingerprintability (de Montjoye et al., 2013; Rocher et al., 2019; Eckersley, 2010), privacy-motivated and metric-fragmented. The closest clinical precedent re-identifies patients from which laboratory tests were ordered, without result values (Johnson et al., 2019). It ranks a single day's laboratory orders against candidate patients to price a privacy risk, where the ICU texture tier reads per-channel observation counts at matched time and offset to price an evaluation confound. The evaluation literature polices the consequence (Tampu et al., 2022; Chaibub Neto et al., 2019), where the exposure measures the dataset-side precondition instead.

The same signals, valued oppositely. Temporal-order pretext tasks (Misra et al., 2016; Wei et al., 2018) and instance discrimination (Wu et al., 2018) treat as free supervision exactly what adverse-event evaluation must price as confound. That a literature has built representation learning on the readability of time and of unit identity is independent evidence that both signals are substantial in ordinary data.

Standards and their dataset-side preconditions. Leakage taxonomies (Kaufman et al., 2012; Kapoor & Narayanan, 2023) document the failure class across fields. Reporting standards (Kapoor et al., 2024) and a meta-review of evaluation failures (Liao et al., 2021) prescribe against it, and forecasting-specific catalogs (Hewamalage et al., 2023) do the same on the temporal axis. Each asks a reader to trust a protocol description or audit it by hand. Measuring the preconditions gives the same reader a number obtained before any event-prediction model is fit, on the data the protocol will run on.

Where the precondition became the failure. A review of deep models for Alzheimer's classification found that about half of the surveyed studies may have suffered from data leakage, with splits not made at the subject level a recurring cause (Wen et al., 2020). A factorial study of connectome models isolated repeated subjects as one of the two leakage forms that drastically inflate performance (Rosenblatt et al., 2024). A screen of COVID-19 imaging models found none of the assessed models clinically usable, with images duplicated across partitions and data-source confounds among the recurring flaws (Roberts et al., 2021). In each case an identity exposure was a property of the corpus before any of those models were fit, and none of those literatures had a routine way to measure it. The Epic sepsis model, deployed at hundreds of hospitals, validated externally far below its reported performance (Wong et al., 2021), which is the operational stake behind the ICU numbers in Section 4.

What these instruments do not cover. The two quantities read ordinary windows along two axes, so leakage mechanisms acting through neither are outside their scope. Resampling applied before the split overturned a body of preterm-birth results (Vandewiele et al., 2021), and model selection carried out on the evaluation data overturned a civil-war prediction result (Neunhoeffer & Sternberg, 2019). Neither is visible to a construction that never consults an outcome and never inspects a pipeline. Feature-selection leakage, treatment leakage, and label circularity are likewise outside. The floor and the exposure characterize a dataset; they do not audit a protocol.