Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

A Multidimensional Data-Quality Profile for Clinical Prediction: An Exploratory Benchmark in MIMIC-IV and eICU

Yunguo Yu

yuyunguo@gmail.com

Mayo Clinic https://orcid.org/0009-0006-2003-3813

Research Article

Keywords: data quality; clinical prediction; electronic health records; distributional shift; label noise; missingness; representativeness; temporal integrity; MIMIC-IV; eICU

Posted Date: October 23rd, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11460866/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: The authors declare no competing interests.

A Multidimensional Data-Quality Profile for Clinical Prediction: An Exploratory Benchmark in MIMIC-IV and eICU

Yunguo Yu

October 21, 2026

Abstract

Background: Clinical prediction models are trained on routine care data, and their quality in deployment depends on that data. However, data-quality assessment in healthcare AI remains largely heuristic: practitioners report missingness tables, class-balance ratios, or demographic summaries, but these numbers are not connected to what actually happens to a model trained on the data. A key open question is whether information available *before* training can flag which dataset conditions will harm a deployed model.

Methods: We report an exploratory benchmark of a five-part, task- and target-specific data-quality profile covering: completeness (how much of the data are missing), label noise (how often recorded labels disagree with the true condition), distributional similarity (how far the training population deviates from the population the model will serve), representativeness (how well the training data cover that population), and temporal integrity (whether future information leaks into training records). Each part is an operational proxy, not a validated instrument, and carries an explicit availability state: a part that cannot be measured is reported as "not measured" rather than imputed as perfect. For every dataset condition we train a real model on the possibly corrupted data and measure how much its log-loss and calibration change on held-out patients, rather than asserting a closed-form relationship. We evaluate three cohorts from the Medical Information Mart for Intensive Care IV (MIMIC-IV; Sepsis-3, N=2,000; intensive care unit (ICU) mortality, N=3,000; KDIGO-staged acute kidney injury (AKI), N=2,293) under controlled corruptions, and one external multi-institution database (eICU Collaborative Research Database, eICU-CRD; 202 hospitals).

Results: Five findings emerge. First, in the controlled cohorts the composite index moved in the expected direction (a direction-only finding at four conditions per cohort, not distinguishable from chance), while each of three common heuristics pointed the wrong way in at least one cohort. Second, the temporal component flagged a training-only leakage corruption whose apparent severity exceeded its measured harm; in a separate train-and-test leakage experiment, apparent discrimination (the area under the receiver operating characteristic curve, AUROC) was inflated by 0.025 (95% interval 0.011–0.039), while the penalty once the leaked features were removed was concentrated in log-loss (log-loss worsening 0.009, interval 0.002–0.017); the calibration penalty was smaller and not distinguishable from zero (ECE worsening 0.014, interval −0.000 to +0.024). Third, in the corruption pairs evaluated, combining two corruptions did not multiply the harm: all interaction contrasts were sub-additive. Fourth, in a pre-specified, patient-grouped, leave-one-family-out evaluation, the full component vector did not outperform either the best scalar baseline or an always-zero prediction of degradation (mean advantage +0.009, 95% interval $[-0.035, +0.051]$ versus the best scalar; +0.005, interval $[-0.020, +0.042]$ versus zero, across five corruption families); the evaluation is inconclusive. Fifth, across 21 analyzable eICU hospitals the composite score tracked calibration—a property that shapes how confidently clinicians can act on predictions—but not detectably discrimination (AUROC); this calibration association

was largely explained by hospital size and mortality prevalence (correlation 0.60 unadjusted, 0.29 after joint adjustment, 0.17 when hospitals were equalized in size; the observed 0.60 lies at the 94th percentile of estimation-artifact-only null experiments, $P = 0.057$). Three results undercut the scalar summary: geometric and arithmetic aggregation were indistinguishable on these cohorts (a rank-agreement finding at four conditions per cohort, not an equivalence), signal was concentrated in a subset of components, and the scalar is a partial index wherever temporal integrity is unmeasured.

Conclusions: We do not claim that any scalar score certifies patient safety. The contribution is a transparent benchmark of which quality signals do and do not track downstream degradation—reported with the components that worked and the ones that failed.

Keywords: data quality; clinical prediction; electronic health records; distributional shift; label noise; missingness; representativeness; temporal integrity; MIMIC-IV; eICU

1 Introduction

Machine-learning models are increasingly used in clinical decision support, from sepsis prediction and mortality risk stratification to early warning systems [1, 2, 3]. A growing body of evidence documents failures of these systems upon deployment that are traceable to the underlying training data rather than to model architecture [4, 5, 6]. Clinical datasets are subject to a set of well-documented pathologies: structured missingness driven by clinical protocols (e.g., lactate measured selectively in severely ill patients) [7, 8]—and informative missingness can itself act as a clinical signal in ICU data [9]—plus label noise from inter-rater disagreement and coding variability [10, 11, 12], demographic bias from single-site or historically skewed cohorts [13, 14, 15], sparse coverage of rare presentations [16], and temporal-integrity violations including leakage of future information across care episodes [17, 18, 19].

The prevailing approach to data-quality assessment in healthcare AI is heuristic and task-agnostic: practitioners report feature-level missingness rates, class-imbalance ratios, or demographic tables as proxies for dataset fitness, without connecting them to model behavior. This gap is consequential because healthcare AI systems increasingly face regulatory scrutiny under good-machine-learning-practice (GMLP) guidance [20], and because data quality is a primary consideration for responsible deployment and clinical adoption [21, 22].

A substantive body of prior work already *organizes* data-quality assessment, from the fitness-for-use multidimensional tradition in data management [23]. Kahn and colleagues harmonized electronic health record (EHR) data-quality terminology around conformance, completeness, and plausibility, with verification and validation contexts [24], and complementary reviews catalogued the methods and dimensions of EHR data-quality assessment [25] and proposed pragmatic data-quality frameworks for the EHR [26]. METRIC, a systematic framework for medical training data, specifies 15 awareness dimensions across five clusters and explicitly makes assessment dependent on the intended application, identifying measurement, downstream-effect quantification, and overall aggregation as open problems [27]. The Observational Health Data Sciences and Informatics (OHDSI) Data Quality Dashboard demonstrates a deployed, configurable framework that expands 20 check types into more than 3,300 executable checks [28]. More recently, the Metric Hub preprint operationalizes medical data-quality dimensions through metric cards and selection workflows, and MetricLib provides an extensible toolkit for medical-dataset quality evaluation [29, 30]. Complementary assurance work frames clinical AI quality as a lifecycle problem, spanning governance and assurance structures [31], post-deployment surveillance [32], and the operating trajectory of deployed models [33]. These developments advance the operationalization of data-quality assessment but do not address whether

pre-model quality measurements predict downstream model-performance degradation—the question this benchmark investigates.

This paper does not claim to be the first to propose multidimensional data-quality assessment. It asks a narrower, testable question:

> *Can information available before fitting the final clinical model identify harmful dataset conditions, and does that information predict measured changes in model performance under controlled and naturally occurring shifts?*

Our contribution is (i) a task- and target-specific five-dimensional quality profile with explicit availability states and an explicit account of which uncertainty estimates were and were not produced (Section 2); (ii) a modest, explicitly stated bound relating an idealized (joint-divergence) formulation of one component to a bounded loss difference, with the measured marginal divergence as a proxy whose limitations are stated in place (Section 3, Appendix B); and (iii) an exploratory benchmark that measures—for each controlled condition—the profile components *and* the resulting change in trained-model performance, reporting both the components that work and the components that fail (Section 4). We deliberately do not claim that a single scalar score certifies reliability or satisfies regulatory requirements.

2 A Multidimensional Data-Quality Profile

2.1 Setup and the reference-distribution problem

We model a clinical dataset as n triplets $D = \{(x_i, y_i, t_i)\}_{i=1}^{n}$ drawn i.i.d. from an observed distribution Q over $\mathbb{R}^p \times \mathcal{Y} \times \mathbb{R}$, where $x_i \in \mathbb{R}^p$ is a feature vector, $y_i \in \mathcal{Y}$ a discrete outcome, and t_i an event timestamp. Data quality is the fidelity of Q to a target distribution P over the same space, connecting dataset quality to the dataset-shift literature [34, 35]. The central difficulty is that P—the population who will encounter the deployed model—is not directly observable. We therefore evaluate quality relative to an identifiable proxy $\hat{P}$ (e.g., a held-out partition, a registry, or a federated multi-site cohort). The following is immediate:

Lemma 1 (Proxy triangle inequality)**.** *For any metric divergence d satisfying the triangle inequality, and any dataset distribution Q,*

$$d(P, Q) \;\le\; d(\hat{P}, Q) + \epsilon_{\text{proxy}}, \qquad \epsilon_{\text{proxy}} = d(P, \hat{P}). \tag{1}$$

Every quantity below is stated relative to $\hat{P}$, with the understanding that true-population statements carry an additional, generally unknown, proxy-error slack ϵ_{proxy}.

2.2 The profile is primary; the scalar is a hypothesis

We define a quality *profile*

$$\mathbf{q}(D; \mathcal{T}, \hat{P}, \mathcal{A}) = (q_C,\, q_N,\, q_B,\, q_R,\, q_T), \tag{2}$$

where $\mathcal{T}$ is the clinical task, $\hat{P}$ the target reference, and $\mathcal{A}$ the audit protocol (the estimator choices, reference construction, and thresholds). Each component lies in $[0, 1]$, with 1 denoting the best-measured state. Each component carries an *availability status*—*measured*, *not measured*, or *not applicable*—and an explicit statement of how its sampling uncertainty was or was not estimated

(Section 2.5). An incomplete audit must not be silently completed: we report a component as *not measured* rather than imputing a perfect score.

A *scalar* quality index can be formed by aggregating the profile (the data-quality index, DQI), for example as the weighted geometric mean

$$\mathrm{DQI}(D) = \prod_k q_k^{w_k}, \qquad w_k \geq 0. \tag{3}$$

We treat this scalar as a *hypothesis to be tested*, not as a conclusion, consistent with broader cautions that a single aggregate statistic can obscure important differences between models [36]. The geometric form is one of several candidate aggregators—arithmetic mean, minimum, and a regularized learned combination—each of which must earn its role through held-out performance, interpretation, and operational usefulness. Two properties of the geometric mean are immediate and valid without further assumptions: it is non-decreasing in each component (for $w_k \geq 0$), and it is bounded in $[0, 1]$. We make no uniqueness claim: the geometric mean is a modeling choice, not a theorem.

Remark 1 (Two restrictions on any continuity claim)**.** *Any claim that* DQI *is Lipschitz in a dataset distance must be restricted. With equal weights $w_k = 1/5$ and a single varying component $q \in [0, 1]$ (all others fixed at 1), the score is $q^{1/5}$, whose derivative $q^{-4/5}/5$ is unbounded as $q \to 0$. A valid Lipschitz statement therefore requires each component to be bounded away from zero—$q_k \in [a, 1]$ for $a > 0$—with a constant depending on a. This tension (behavior near zero vs. continuity) is inherent to a product aggregator and is stated here explicitly rather than claimed away.*

2.3 Component definitions

Completeness q_C (missingness). Let m_j be the fraction of missing values in feature j. We define

$$q_C = \sum_{j=1}^{p} \alpha_j (1 - m_j), \qquad \sum_j \alpha_j = 1,\ \alpha_j \geq 0, \tag{4}$$

where the weights α_j reflect each feature's predictive relevance to the task. The implementation estimates α_j by feature–outcome mutual information (MI) on the available cases for each feature in the clean reference. This available-case MI estimator has two distinct sources of bias: (i) *selection distortion*: under MNAR (missing not at random) the missingness mechanism itself biases the available cases [37, 38]; (ii) *finite-sample estimator bias*: even under MCAR (missing completely at random; where selection distortion is absent), the k-nearest-neighbor (k-NN) mutual-information estimator with fixed k has nonzero finite-sample bias and sampling variability. A validation experiment (Experiments F6–F10) showed: mean absolute bias 0.005 under MCAR and 0.010 under MAR (missing at random); bootstrap coefficient of variation 0.15 (max 0.32); k-sensitivity (mean absolute difference < 0.008 for $k \in \{3, 7, 10, 15\}$ vs $k = 5$). Zero-MI features correctly receive zero weight. The statement that available-case MI is "unbiased only under MCAR" is therefore false for finite samples; MCAR removes selection distortion but not finite-sample estimator bias. We treat the completeness weights as an open identifiability question rather than a solved one [39, 40].

Label noise q_N. The idealized definition models label noise through a class-conditional transition matrix T with $T_{ij} = \Pr(\tilde{Y} = j \mid Y = i)$ and severity weights Λ:

$$q_N^{\mathrm{ideal}} = \exp\Big(-\big\|\Lambda^{1/2}(T - I)\Lambda^{1/2}\big\|_F\Big). \tag{5}$$

This is a heuristic severity-weighted Frobenius penalty; it does not preserve the full eigenstructure of T and assumes mislabeling depends only on the true class (not features), a known simplification of clinical coding error [41, 42].

Operationally, the transition matrix T is unobserved. The implemented proxy uses cross-validated logistic-regression disagreement:

$$\eta = \frac{1}{n}\sum_{i=1}^{n} \mathbb{I}[\hat{y}_i^{\mathrm{CV}} \neq y_i], \qquad q_N = \exp(-\lambda\eta), \quad \lambda = 1 \text{ (default)}, \tag{6}$$

where $\hat{y}_i^{\mathrm{CV}}$ are cross-validated (CV) logistic-regression predictions. This proxy measures model–label disagreement, *not* the true label-noise rate. The default $\lambda = 1$ implies $q_N \in [e^{-1}, 1] \approx [0.368, 1]$; even maximal disagreement ($\eta = 1$) yields $q_N \approx 0.37$, and under equal five-component weights the maximum DQI penalty from label noise is $1 - e^{-1/5} \approx 0.18$. This limited dynamic range means q_N cannot strongly penalize extreme label noise, and equal nominal weights do not imply comparable influence across components. The choice of λ affects the aggregation and should be calibrated to the intended noise severity scale.

A falsification experiment (Experiments F1–F5) revealed critical failures:

- On linearly separable data with 100% label flip, $q_N \approx 0.997$ (near-perfect score) despite 100% label corruption—the model learns the flipped labels perfectly, so CV disagreement is near zero.
- On clean but nonlinear data (moons), $q_N \approx 0.884$—model misspecification masquerades as label noise (false positive).
- On asymmetric and instance-dependent noise, q_N tracks injected noise rate ($r \approx -1.0$), but the full-flip and nonlinear failures show the proxy is not a reliable detector of label corruption in general.

We name it a proxy and report these failures explicitly (Table 1).

Marginal Distributional Similarity q_B. We measure departure between Q and $\hat{P}$ by the Jensen–Shannon (JS) divergence, which is symmetric and bounded under support mismatch:

$$q_B = 1 - \frac{\mathrm{JS}(Q\|\hat{P})}{\ln 2}. \tag{7}$$

Operationally we estimate this as an *average of marginal* (per-feature) Jensen–Shannon divergences, not a joint divergence on the full distribution; the average is a proxy that is insensitive to dependence structure, and we do not imply it estimates the joint divergence. In high dimensions a joint divergence would additionally require dimensionality reduction, introducing an approximation error that must be reported. We use this component in Section 3 only through a standard bounded-divergence inequality, stated with its restrictions.

Target-Support Coverage q_R. We score coverage of the reference distribution using a kernel-smoothed non-coverage weight

$$q_R = 1 - \int_{\mathcal{M}} s_\tau(z; D)\, d\mu_{\mathcal{M}}(z) \Big/ \int_{\mathcal{M}} d\mu_{\mathcal{M}}(z), \qquad s_\tau(z; D) = \exp(-\hat{p}_h(z)/\tau), \tag{8}$$

where $\hat{p}_h$ is a kernel density estimate and $\tau > 0$ a scale parameter. The kernel-smoothed form is continuous; a hard nearest-neighbor coverage indicator is not. Because the operational estimator

can depart from this idealized definition, Table 1 names the quantity actually measured and the estimator used.

Temporal integrity q_T. We score ordering violations, leakage, and causal consistency [43]. The leakage sub-component is computed from real measurement timestamps: a record is flagged when a prediction-time feature value was last updated after the prediction window closed. The ordering and causal-consistency sub-components require per-event timestamp sequences, which were not extracted in this study; they are reported as *not measured*, not as perfect scores. Within q_T, the measured sub-components are combined by a geometric mean, so an unmeasured sub-component does not dilute a measured one (Section 4). At the profile level, an unmeasured component is excluded and the remaining weights are renormalized to sum to one; coverage $k/5$ is reported alongside the score. We do not impute a perfect score. The scalar is therefore a *partial* index when coverage is incomplete, and scores with different coverage are not directly comparable.

Table 1 states, for each component, the quantity actually measured and its estimator, so that a proxy is named as a proxy rather than implied to estimate the idealized population definition. Implementation details are: 20 equal-width bins per feature for q_B; a $k = 5$ nearest-neighbor kernel bandwidth for q_R, with τ calibrated to approximately 99% training coverage; 3-neighbor mutual information for q_C; and 3-fold cross-validated logistic-regression disagreement for q_N. Table 1 specifies, for each component, the quantity actually measured, its estimator, and its availability.

Table 1: Operational component specification. For each component, the quantity actually measured, its estimator, its availability, and its validity, distinguishing the operational estimator from the idealized definition where they differ. Missing completely at random (MCAR).

Component	**Operational estimator (measured quantity)**	**Availability**	**Validity**
q_C	Mutual-information-weighted missingness; weights are per-feature available-case MI normalized to sum to one. Available-case MI has selection distortion under MNAR; under MCAR it removes selection distortion but the finite-sample k-NN MI estimator (n_neighbors=3) has nonzero bias (mean abs bias 0.005 under MCAR, 0.010 under MAR) and bootstrap CV 0.15 (max 0.32). Zero-MI features receive zero weight. k-sensitivity (mean —diff— ¡ 0.008). Not unbiased under MCAR.	measured (weights flagged)	proxy
q_N	Cross-validation (CV) disagreement *proxy*: η = CV disagreement rate of logistic regression fit to labels; $q_N = \exp(-\lambda\eta)$ with $\lambda = 1$ (default). *Not* the severity-weighted transition-matrix norm. Falsification: $q_N \approx 1.0$ under 100% label flip on separable data (fails to detect); $q_N < 1.0$ on clean nonlinear data (false positive).	proxy	proxy
q_B (Marginal Distributional Similarity)	Marginal Jensen–Shannon divergence averaged over features (per-feature histograms)—*not* a joint divergence on the full distribution.	measured	proxy
q_R (Target-Support Coverage)	Kernel-smoothed coverage of target points by the training support (a target-weighted expectation—*not* a uniform manifold-volume integral).	measured	proxy
q_T	Leakage sub-score only: $q_T = 1 - \frac{1}{n}\sum_{i=1}^{n} \mathbb{I}[t_i^{\text{last_update}} > t_i^{\text{pred_window}}]$, where $t_i^{\text{last_update}}$ is the latest measurement timestamp (charttime for labs) and $t_i^{\text{pred_window}}$ is the prediction window cutoff. Ordering and causal consistency sub-components are not measured.	partial	proxy

2.4 Construct validation and falsification experiments

We conducted two targeted validation experiments to assess the operational behavior of the completeness and label-noise components against ground-truth corruption mechanisms.

Label-noise proxy falsification (Experiments F1–F5). The operational label-noise score is $q_N = \exp(-\lambda\eta)$ where η is the cross-validated logistic-regression disagreement rate (5-fold) and $\lambda = 1$ is the default severity scale. We tested this proxy against five experimental conditions:

1. **Full label flip on separable data:** A linearly separable classification task (2000 samples, 10 features, class separation 2.0) with all labels flipped. The proxy should detect 100% corruption.

2. **Asymmetric label noise:** Class-conditional flip rates $\pi_{0\to1} \in \{0.1, 0.25, 0.4\}$, $\pi_{1\to0} \in \{0.1, 0.25, 0.4\}$ on the same separable task.

3. **Instance-dependent noise:** Flip probability proportional to a feature value ($p_{\text{flip}} = \text{severity} \times \mathbb{I}[x_1 > 0]$) on the separable task.

4. **Nonlinear negative control:** A clean two-moons dataset (2000 samples, noise 0.1) with no label corruption. The proxy should not flag model misspecification as label noise.

5. **MIMIC-IV mortality with injected noise:** Real cohort data (3000 stays, 12 features) with uniform label noise injected at rates $\{0.1, 0.25, 0.4\}$.

For each condition we compute q_N and compare it to the known injected noise rate. The full results are reported in Section 4.8.

Completeness-weight validation (Experiments F6–F10). The completeness component uses available-case mutual information (MI) weights estimated via k-NN MI (n_neighbors=3). We assessed:

1. **Weight comparison:** MI weights vs. equal weights vs. clinically prespecified weights (labs > vitals > demographics) on MIMIC-IV mortality data.

2. **Bootstrap stability:** 100 bootstrap resamples of the MI weight vector; coefficient of variation per feature.

3. **Zero-MI handling:** A constant feature appended to the data; its weight should be zero.

4. **k-Sensitivity:** MI weight changes for $k \in \{3, 5, 7, 10, 15\}$ relative to $k = 5$.

5. **MCAR vs MAR bias:** Synthetic data (2000 samples, 10 features) with known MI structure; available-case MI bias under MCAR (30% random missingness) vs. MAR (missingness depends on feature 0).

Results are reported in Section 4.8.

Construct-validation panel (Experiment F11). Each component was tested against its target corruption type across a severity grid (missingness $\in \{0, 0.1, 0.2, 0.3, 0.4\}$, label noise $\in \{0, 0.1, 0.2, 0.3, 0.4\}$, shift $\in \{0, 0.5, 1.0, 1.5, 2.0\}$ SD). Components are expected to respond (Spearman $\rho \approx -1$) to their target corruption and ignore others. Results are reported in the reproducibility package.

2.5 Uncertainty and availability

Because each component is estimated from a finite sample, it is a random quantity. We report a point estimate for each component and make the sampling unit explicit: patient versus stay versus row, whether repeated stays are clustered within one partition, whether the target sample is fixed across runs, and how clean and corrupted models are paired. In the present study the primary source of variability is 20 repeated corruption-and-training runs; this is distinct from a bootstrap over a fixed dataset, and the two are not conflated. Intervals are reported for the composite degradation score and for the performance gaps. Component-level population intervals and a variance decomposition were not estimated; we report run-to-run component standard deviations and treat them as run variability, not sampling uncertainty (Section 5.4). We state explicitly whether each reported interval is a percentile over runs, a bootstrap interval, or a confidence interval.

3 Statistical Estimands and a Bounded Divergence Inequality

We state a small number of precise results that hold under explicit assumptions, and we state what they do *not* establish.

The empirical quantity this paper measures is the **corruption-induced target excess loss**:

$$\Delta(D') \;=\; R_{\widehat{P}}\big(\mathcal{A}(D')\big) \;-\; R_{\widehat{P}}(\mathcal{A}(D_{\mathrm{ref}}))\,, \tag{9}$$

where $\mathcal{A}$ is the training procedure (L2-penalized logistic regression with scikit-learn defaults, on standardized inputs), D_{ref} is the reference (uncorrupted) training dataset, D' is the (possibly corrupted) training dataset, and $R_{\widehat{P}}(f)$ is the expected log-loss of the trained model f on a fixed held-out target sample drawn from $\widehat{P}$. This estimand compares *differently trained* predictors evaluated on a *common* target; it is not the risk of a fixed predictor across distributions.

A standard bounded-divergence inequality relating total-variation distance to Jensen–Shannon divergence is stated in Appendix B for reference. That bound concerns a fixed predictor evaluated under two distributions and therefore does *not* directly control the algorithm-dependent estimand in Equation (9). We include it as background because the bias component q_B is defined through a JS divergence, but we do not claim the bound validates the profile or its aggregation.

These restrictions frame the empirical question in Section 4: does the profile (or any aggregation of it) carry predictive information about measured degradation, above and beyond simpler alternatives?

4 Exploratory Empirical Evaluation

4.1 Design

We evaluate the profile on three cohorts from the Medical Information Mart for Intensive Care IV (MIMIC-IV) [44], each stressing a distinct subset of components, and on one external multi-institution database, the eICU Collaborative Research Database (eICU-CRD) [45]. For every controlled condition we (i) compute the profile components, and (ii) train a model on the (possibly corrupted) training data and evaluate it against a held-out target set, reporting the measured target excess log-loss (Δ log-loss relative to a model trained on clean or demographically matched data) and the calibration gap, defined as the change in expected calibration error (ECE) [46]. Because calibration determines how a predicted probability can be acted on in practice [47, 48], we report it alongside discrimination throughout, motivated by dynamic frameworks for clinician trust and confidence calibration [49]. Log-loss is computed on probabilities clipped to $[10^{-6}, 1-10^{-6}]$ (a numerical convention, not a clinical threshold). Training and target partitions were a 70/30 split (stratified by outcome in the sepsis and mortality cohorts; random in the AKI cohort). Medians, scaling parameters, and model fits were estimated on the training partition only. The primary model was logistic regression (`max_iter=1000`) on standardized inputs; gradient boosting and random forest (default hyperparameters, seed 42) were sensitivity checks. Calibration is expected calibration error with 10 equal-width probability bins. We emphasize that this is an *exploratory* benchmark: the audit-to-performance mapping was computed on the same conditions reported here, so the three-cohort associations cannot establish out-of-sample predictive validity. A leave-one-family-out evaluation is reported in Section 4.11. Figure 1 summarizes the data-access and audit workflow.

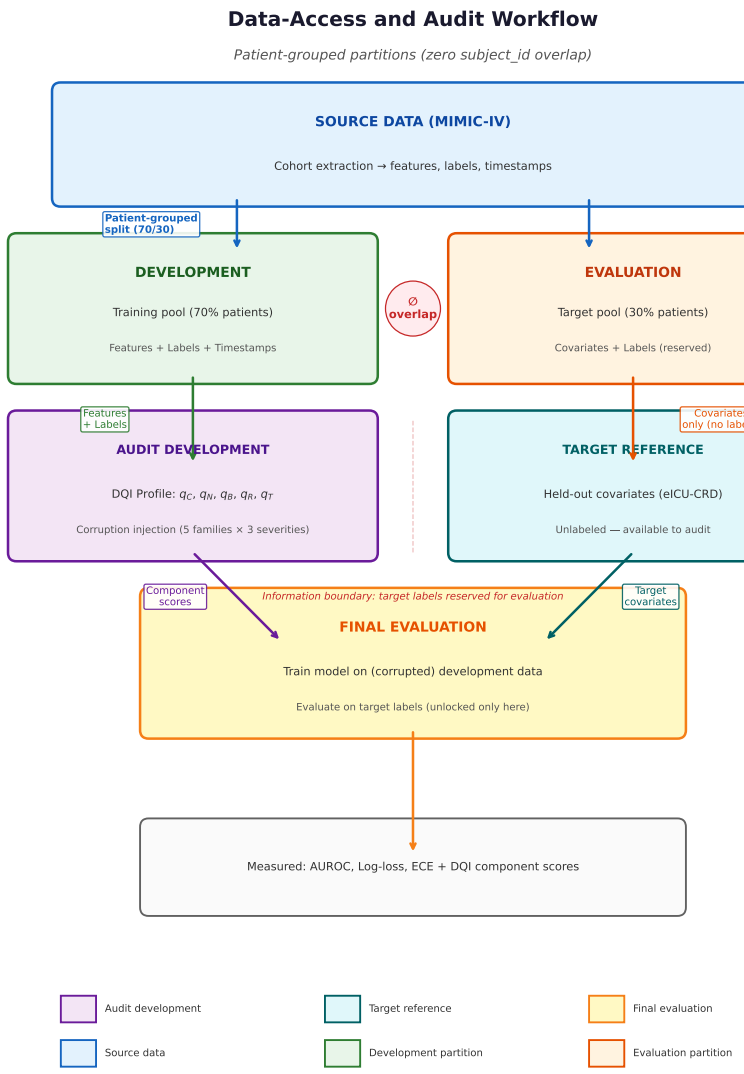

Figure 1: Data-access and audit workflow for the DQI profile. The pipeline proceeds from source training data (MIMIC-IV) through audit development (DQI profile construction with component computation, corruption injection, and geometric aggregation) to target reference (eICU-CRD) and final evaluation. Corruption injection feeds back into source data; audit feedback loops back to audit development. See Section 4.1 for details.

Each corruption is motivated by a clinically plausible failure mode; the mapping is approximate, and the mechanisms are not exhaustive. All reported performance quantities are measured by training a real model (logistic regression) and evaluating it on a held-out target set, not computed from a closed-form surrogate. Reported performance intervals are percentiles over 20 repeated corruption-and-training runs and are not confidence intervals for a population parameter. All gaps, contrasts, and leave-one-family-out advantage values are computed from unrounded paired runs; displayed means are rounded to three decimals, so differences of displayed values can differ from reported gaps or contrasts by up to 0.001. DQI entries are means ± standard deviations

across the same runs. Component entries are means and run-to-run standard deviations; these are not population confidence intervals. DQI entries with coverage 4/5 are partial indices, not five-component scores. In the sepsis and AKI cohorts q_T is unmeasured and excluded from the aggregate (a renormalized four-component index); in the mortality cohort q_T contributes its measured leakage sub-component. Because the aggregation differs, cross-cohort composite *levels* are not directly comparable; cross-cohort claims in this paper concern ranking direction only. Coverage is reported in the tables and in the reproducibility package.

4.2 Patient-level split specification and audit

For each cohort we report the eligibility criterion, prediction landmark, predictor window, outcome window, and the handling of death or discharge before the landmark. All splits were performed at the stay level (using `stay_id`) with a fixed 70/30 ratio, stratified by outcome (sepsis, mortality) or random (AKI), using `random_state=42` for reproducibility. The target partition was held out and never used for training, imputation, or weight estimation.

- **Sepsis (MIMIC-IV).** Eligibility: first 2,000 adult Sepsis-3 stays by `stay_id` from `mimic_derived.sepsis3` [50, 1]. Landmark: ICU admission. Predictor window: first 24 h of ICU stay (vitals and labs averaged). Outcome: in-hospital mortality (`hospital_expire_flag`), assessed at discharge; patients who died or were discharged before 24 h are included (their available measurements are used). Split: 1,400 train stays / 600 target stays, stratified by mortality label. *Outcome-window note:* for stays ending within the first 24 h (early deaths or discharges), the in-hospital-mortality endpoint is effectively determined inside the observation window. *Note:* Sepsis-3 eligibility is determined retrospectively from the full ICU stay; a patient may not meet criteria at the prediction landmark. This is a descriptive experiment on a retrospectively defined cohort, not a prospective screening evaluation.

- **Mortality (MIMIC-IV).** Eligibility: first 3,000 adult ICU stays by `stay_id` with ICU length of stay ≥ 1 day (12.1% in-hospital mortality). Landmark: ICU admission. Predictor window: admission-time features (demographics, vitals, labs). Outcome: in-hospital mortality. *Eligibility note:* the length-of-stay threshold excludes deaths occurring during the first ICU day; the estimand is in-hospital mortality among stays reaching ICU day 1. Split: 2,100 train stays / 900 target stays, stratified by mortality label.

- **AKI (MIMIC-IV).** Eligibility: 2,293 stays with KDIGO stage ≥ 1 from an 8,000-stay candidate pool, staged from serial creatinine against lowest pre-admission value (or first measured value). Landmark: ICU admission. Predictor window: admission-time features. Outcome: severe AKI (KDIGO stage ≥ 2) — *contemporaneous staging task, not prospective prediction*; creatinine is both an input and part of the outcome definition. Split: random 70/30 (`random_state=42`) into a training pool and an evaluation pool; the target proxy $\hat{P}$ is a demographically resampled subset of the evaluation pool.

- **eICU-CRD Mortality.** 202 distinct hospitals in a random sample of 8,000 adult ICU stays, yielding $N = 7{,}973$ after exclusions. Landmark: ICU admission. Predictor window: admission-time features (7 features with $< 30\%$ missingness plus sex). Outcome: in-hospital mortality. The MIMIC-IV mortality training set (above) is evaluated directly on the eICU target; no eICU training split is performed.

4.3 eICU null experiment specification

To assess whether the observed association between the composite degradation score δ (defined as $1 - \text{DQI}$; Section 2.2) and per-site ECE could arise from estimation artifacts alone, we conducted a null experiment under the hypothesis of no genuine between-site quality differences. The null population is the pooled eICU cohort ($N = 7{,}973$). For each of $R = 300$ replicates, we draw 21 pseudo-sites by sampling with replacement from the pooled population, matching the observed site sizes (the 21 eligible sites with ≥ 50 stays and ≥ 10 events). For each pseudo-site we recompute: (i) δ for the MIMIC-IV training data relative to that pseudo-site as target; (ii) ECE and $1 - \text{AUROC}$ for the MIMIC-IV-trained model evaluated on that pseudo-site. We then compute the Spearman correlation $\rho(\delta, \text{ECE})$ and $\rho(\delta, 1 - \text{AUROC})$ across the 21 pseudo-sites. The null distribution of these correlations characterizes the association generated purely by sampling variability, ECE binning bias, noise in δ, and shared dependence on site size and prevalence. The empirical one-sided P-value is the fraction of null replicates where the correlation equals or exceeds the observed value. This experiment is distinct from the bootstrap confidence intervals reported elsewhere (which quantify sampling uncertainty of a fixed estimator); here the entire estimation pipeline is re-run on each pseudo-site sample.

A patient-level split audit was performed to verify that stays belonging to the same patient (`subject_id`) were not split across train and target partitions. The audit revealed that the stay-level splits do not guarantee patient-level independence:

- Sepsis: 14 shared patients (1.0% of train, 2.4% of target) out of 1,380 / 595 unique patients.
- Mortality: 24 shared patients (1.2% of train, 2.7% of target) out of 2,078 / 894 unique patients.
- AKI: 43 shared patients (2.8% of train, 6.3% of target) out of 1,559 / 680 unique patients (training pool vs. evaluation pool).

This patient-level overlap means the reported within-cohort and cross-cohort comparisons are not fully independent at the patient level, and run-resampling intervals calculated within a stay-level split understate the patient-level variability (point estimates that nominally exclude zero should be read as conditional on this split). We do not re-split at the patient level in this exploratory benchmark; this is a limitation noted in Section 5.4.

4.4 Temporal availability semantics

The temporal integrity component (q_T) uses real measurement timestamps to flag features whose source measurement occurred after the prediction window. The exact timestamp fields used differ by database:

- **MIMIC-IV.** The temporal audit uses `charttime` from the `mimic_derived.chemistry` table (which derives from `mimic_hosp.labevents`). `charttime` corresponds to specimen acquisition time. MIMIC-IV also provides `storetime` in `mimic_hosp.labevents` (the time the laboratory result became available in the information system). A specimen collected before the prediction cutoff can have its result become available only afterward; therefore `charttime` may understate the true availability delay. The current implementation uses `charttime` because `storetime` is not carried forward into the `mimic_derived.chemistry` table used for extraction. This is a limitation: the audit may under-detect leakage from results that were collected before the prediction window but reported after it. A targeted storetime re-extraction (6.9% of records had `storetime` $>$ 24h after `charttime`, indicating delayed availability) confirmed the scope

of this under-detection; however, the main analysis uses `charttime` for consistency with the cohort extraction pipeline.

- **eICU-CRD.** The external cohort lacks absolute timestamps. The analysis uses relative temporal fields: `labResultOffset` and `labResultRevisedOffset` for laboratory records, and `observationOffset` for periodic vital signs. These offsets are measured in minutes from ICU admission. They provide usable temporal information for ordering and leakage checks, but they do not give absolute wall-clock availability times. The analysis uses these offsets as available; the limitation is that `labResultRevisedOffset` (revised result time) may differ from `labResultOffset` (initial result time), and the analysis uses the available offset fields without further revision tracking.

This semantic distinction is documented here rather than in the component definition table (Table 1) so that readers can judge the scope of the temporal audit. The temporal integrity component measures leakage from features whose measured timestamp exceeds the prediction window cutoff; it does not separately measure ordering violations or causal consistency (those sub-components remain unmeasured).

Figure 2 gives an at-a-glance view of all five components under every condition in the three controlled cohorts, before the case-by-case analysis. It shows (i) that each component responds selectively to its target corruption—completeness falls under missingness, distributional similarity falls under the institutional shift, and availability falls under the temporal-contamination condition—and (ii) which components are unmeasured in which cohorts (q_T is displayed as 0 in the sepsis and AKI panels by convention for "not measured", not as evidence of degradation).

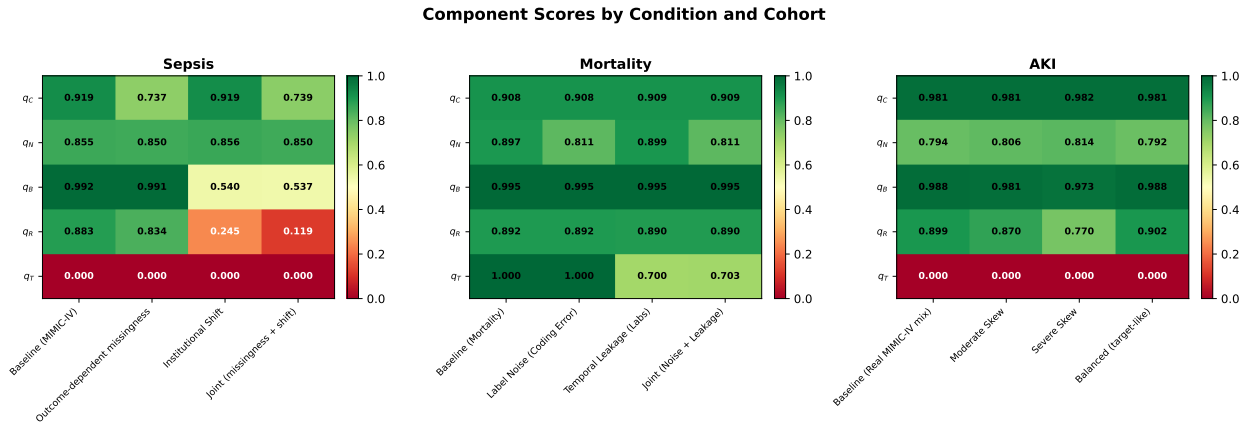

Figure 2: Component scores by condition and cohort for the three controlled MIMIC-IV cohorts, using the same mean component values summarized in Tables 2, 3, and 5. Darker green is a better score; darker red is a worse score. The q_T row is displayed as 0.000 in the sepsis and AKI panels because temporal integrity was not measured in those cohorts (the panel color encodes "not measured" as zero and should not be read as a score of zero); in the mortality panel q_T is measured and falls from 1.000 to 0.700 under the temporal-contamination condition.

4.5 Case Study I — Sepsis cohort: missingness and distribution shift

We selected the first 2,000 adult Sepsis-3 stays by `stay_id` from `mimic_derived.sepsis3` [1], averaging vitals and labs over the first 24 h of intensive care unit (ICU) stay. The prediction label is in-hospital mortality (`hospital_expire_flag`); the model does not predict incident sepsis onset.

Two conditions stress q_C and q_B: (i) outcome-dependent lab missingness in four discretionary labs (lactate, creatinine, blood urea nitrogen (BUN), and white blood cell count (WBC)), and (ii) an institutional covariate shift (per-feature mean shifted by 1.5 standard deviations). The missingness mechanism is outcome-dependent: at the nominal 40% setting, decedents receive a 60% additional-missingness probability and survivors a 35.6% probability for each of the four labs, applied to currently observed values, so about 40% of lab results are additionally masked and decedents lose substantially more lab coverage than survivors; realized pre- and post-injection rates per feature and outcome group are reported in the reproducibility package. Missingness is restricted to features a clinician can choose not to order; continuously monitored vitals and demographics are not perturbed. Table 2 reports the composite score and the measured gaps for the four conditions, and Figure 3 plots the composite degradation against the measured gap.

Table 2: Case Study I — Sepsis cohort ($N = 2{,}000$). DQI: mean ± SD across 20 runs; this is a four-component partial index (q_T not measured; coverage 4/5). Risk/ECE gap: mean [percentile interval] of the measured gap between a model trained on the condition's data and a model trained on clean data, evaluated on a held-out target set.

Condition	**DQI**	**Risk gap**	**ECE gap**
Baseline	0.911 ± 0.000	0.000	0.000
Outcome-dependent missingness	0.849 ± 0.002	0.007 [0.002, 0.014]	0.007 [−0.005, 0.015]
Institutional shift	0.568 ± 0.000	0.127	0.110
Joint (missingness + shift)	0.447 ± 0.009	0.086 [0.069, 0.107]	0.055 [0.045, 0.065]

Three observations are central. First, the institutional shift is detected strongly: the Marginal Distributional Similarity component falls from 0.992 to 0.540 and the composite DQI falls from 0.911 to 0.568, consistent with the measured log-loss gap of 0.127. This is the strongest single-defect effect in the cohort. Second, the outcome-dependent missingness condition does not produce a large measured log-loss degradation (0.007): the missingness mechanism lowers the composite DQI to 0.849, which we report as a detection signal rather than as evidence of downstream harm. Third, the joint condition has the lowest composite score (0.447), but its measured log-loss gap (0.086) is smaller than the shift condition's (0.127), so the composite's ranking and the measured degradation diverge here. The paired interaction contrast on matched corruption realizations gives

$$I_{\text{risk}} = -0.048\ [-0.060,\ -0.030], \qquad I_{\text{ECE}} = -0.060\ [-0.077,\ -0.046], \tag{10}$$

(mean and 95% percentile interval over a separate 40-run matched experiment with its own corruption realizations, distinct from the 20-run table). Both interactions are *sub*-additive: the joint degradation is less than the sum of the single-defect degradations. A lower composite score in the joint condition does not by itself demonstrate synergy; we therefore do not claim multiplicative compounding, and we treat the geometric aggregator as one candidate rather than a preferred model.

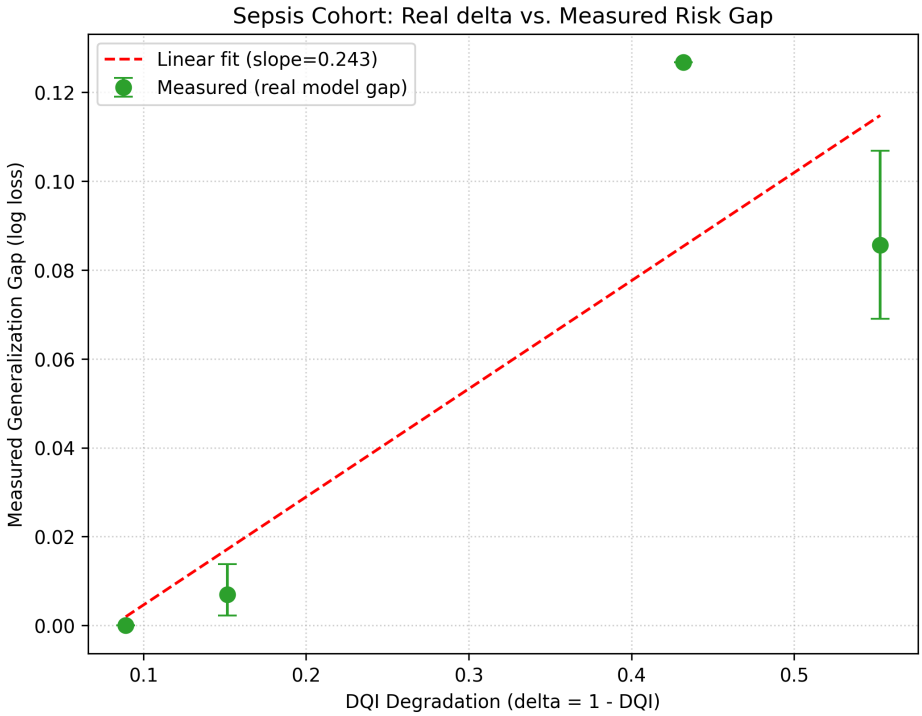

Figure 3: Sepsis cohort: DQI degradation $\delta = 1 - \text{DQI}$ versus the measured generalization gap (log-loss). Error bars are percentiles over 20 runs; the institutional-shift corruption is deterministic and carries no run-to-run interval. δ ranks the missingness and joint conditions in the same order as their measured gaps, but the shift and joint conditions are ordered differently from their measured gaps (shift 0.127 vs. joint 0.086), so the composite does not reliably order every defect by downstream impact in this cohort.

4.6 Case Study II — ICU mortality cohort: noise and temporal integrity

We selected the first 3,000 adult ICU stays by `stay_id` with ICU length of stay ≥ 1 day (12.1% in-hospital mortality) and applied two conditions: (i) asymmetric label noise (deaths coded as survived at a higher rate than the reverse), and (ii) temporal leakage (the real creatinine and BUN value from the 48 h window preceding discharge or death substituted into the admission-time feature). The leakage mask targets 35% of records but is applied only where both late-window values are available; the realized flagged fraction is about 30%. The substitution is a genuine leakage mechanism: late-window BUN separates survivors from non-survivors by a wider margin than admission-time BUN. Table 3 reports the composite score and measured gaps; Figure 4 plots the composite degradation against the measured gap.

Table 3: Case Study II — ICU mortality cohort ($N = 3{,}000$). DQI: mean ± SD across 20 runs; hierarchical coverage: five top-level components available; temporal integrity (q_T) coverage limited to the specified leakage sub-component only. Risk/ECE gap: mean [percentile interval], measured as in Table 2. "Training-only contamination" substitutes late-window values in the training partition only and evaluates on a clean target; it is not a train-and-test leakage design (see Table 4).

Condition	**DQI**	**Risk gap**	**ECE gap**
Baseline	0.937 ± 0.000	0.000	0.000
Label noise (coding error)	0.919 ± 0.001	0.042 [0.034, 0.050]	0.062 [0.050, 0.075]
Training-only temporal contamination	0.873 ± 0.002	0.003 [0.001, 0.005]	0.008 [0.002, 0.011]
Joint (noise + leakage)	0.856 ± 0.003	0.043 [0.033, 0.051]	0.063 [0.050, 0.078]

Label noise has the largest single-defect risk gap (0.042); the joint condition is only slightly larger (0.043). Training-only contamination shows a small effect whose run-resampling interval nominally excludes zero (0.003; such intervals are conditional on the fixed stay-level split, Section 4.2). The temporal-integrity component, computed from real measurement timestamps, flags the contamination: q_T falls from 1.0 to 0.70 (about 30% of records have a prediction-time value whose source measurement occurred after the prediction window). The corruption has little downstream effect, because substituting late creatinine and BUN values for a minority of records only weakly perturbs a twelve-feature logistic regression. This is a discrepancy of magnitude, not a detection failure: q_T detects the mechanism, but the composite over-ranks this condition relative to its measured harm. The ordering sub-component of q_T remains unmeasured (no per-event timestamp sequences were extracted). A paired interaction contrast for the label-noise × leakage pair (40 matched runs, structured as in Section 4.5) gives $I_{\text{risk}} = -0.003$ $[-0.004, -0.001]$ and $I_{\text{ECE}} = -0.007$ $[-0.013, -0.001]$: the joint harm is again *sub*-additive rather than compounding.

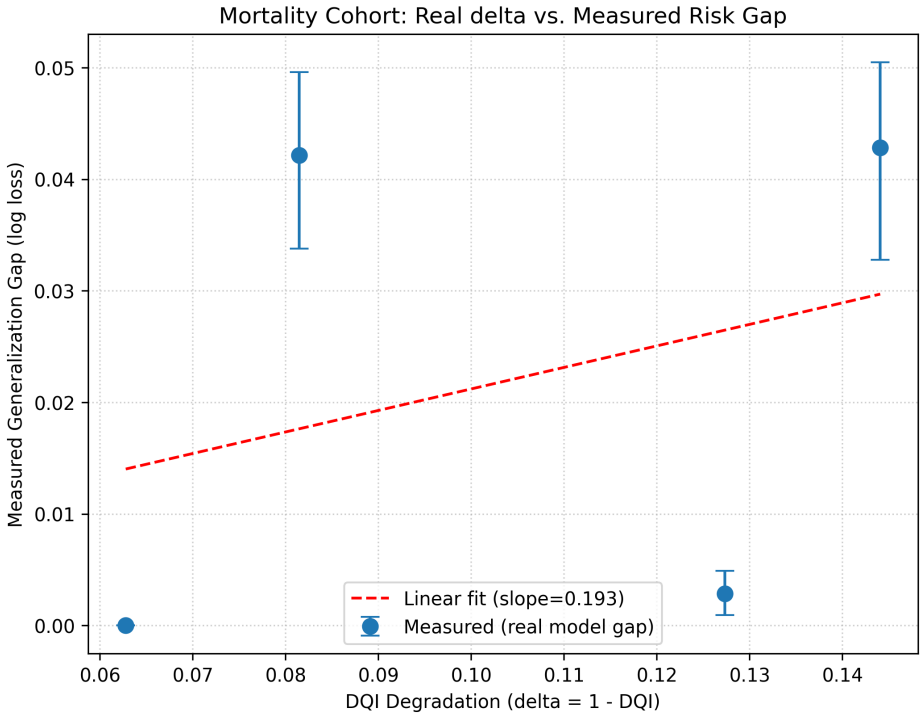

Figure 4: Mortality cohort: DQI degradation versus measured target excess log-loss. The training-only temporal-contamination condition carries a larger δ because q_T detects the leaked records, even though its measured target excess log-loss is small.

Table 4 summarizes the separate train-and-test leakage experiment.

Table 4: True train-and-test temporal-leakage experiment in the mortality cohort (20 repeated stratified splits). Late-window creatinine and BUN were substituted in both training and evaluation for every record with both values (88.7% of training records, 88.2% of target records). "Apparent" is a leaked model evaluated on leaked data; "deployment" is the same model evaluated on clean admission-time features. Values are means; brackets are 2.5th–97.5th percentiles over splits. q_T was 0.147 in leaked training data and 0.151 in leaked target data. Splits differ from the fixed split used in Table 3, so absolute AUROC is not directly comparable across the two tables; all contrasts in this table are paired within split.

Evaluation	AUROC	Log-loss	ECE
Clean model, clean target (reference)	0.788 [0.749, 0.822]	0.306	0.018
Leaked model, leaked target (apparent)	0.813 [0.778, 0.846]	0.291	0.022
Leaked model, clean target (deployment)	0.788 [0.759, 0.822]	0.315	0.032
Clean model, leaked target	0.802 [0.762, 0.838]	0.297	0.020

The true leakage experiment separates detection from evaluation design. Apparent AUROC exceeded the clean reference by 0.025 (95% percentile interval 0.011–0.039); the same model's AUROC on clean deployment data was 0.000 lower than the clean reference (interval −0.009 to +0.007), so the apparent discrimination gain did not persist. Log-loss told a different story: apparent log-loss was 0.024 lower than deployment log-loss (interval 0.016–0.033), and clean deployment log-loss was 0.009 higher than the clean reference (interval 0.002–0.017). ECE followed the same

pattern but its deployment harm interval included zero (+0.014, interval −0.0003 to +0.024). Thus q_T detected the contamination, and leakage inflated apparent discrimination; the clean deployment penalty in this cohort was concentrated in log-loss and calibration rather than AUROC. The training-only contamination condition is a different design—a contaminated training set evaluated on a clean target—and should not be read as a standard leakage experiment.

4.7 Case Study III — AKI cohort: representativeness

The analysis cohort comprised 2,293 MIMIC-IV stays with KDIGO stage ≥ 1 identified from an 8,000-stay candidate pool and staged from serial creatinine against the lowest pre-admission value when available (otherwise the first measured value) [51]. We compared it against a demographically resampled target proxy $\hat{P}$ (50% male, 50% White, drawn from a disjoint held-out partition). The outcome evaluated for the risk and calibration gap is severe AKI (KDIGO stage ≥ 2), not mortality, because the mechanism under test is staging accuracy under demographic skew. Creatinine remains an input; the stage variable is excluded to avoid trivial leakage. This is a contemporaneous staging/classification task, not prospective prediction: the predictor and outcome share creatinine, and we flag that overlap as a limitation to be resolved in a prospective-prediction design.

Representativeness is quantified as kernel-smoothed coverage of the target by the training support (Section 2.3): the fraction of target points falling outside the training density, computed with a continuous kernel weight rather than a hard nearest-neighbor threshold. The "skewed" and "balanced" conditions are constructed by resampling real patient records (with replacement where a stratum's pool is small) to hit a target composition; every value is a real MIMIC-IV measurement, and only which patients are included differs. Because each condition is a synthetic recomposition of real records, the experiment measures sensitivity to demographic composition—not subgroup fairness on any observed population. The recomposed pools also shrink the number of distinct contributing patients under skew (1,171 unique patients in the natural-mix training pool vs. 712 under severe skew, resampling audit in the reproducibility package), so part of the q_R decline may reflect training-pool shrinkage in addition to composition; we read the skew results as compositional sensitivity rather than as isolated demographic effects. The reference model is trained on a balanced (50% male, 50% White) resample that matches the target, so every reported gap—including the natural-mix baseline's—is the difference from a target-matched model, not from a "clean" reference condition. Table 5 reports the component and gap results for the four demographic conditions.

Table 5: Case Study III — AKI cohort ($N = 2{,}293$) across demographic-skew conditions. Target proxy $\hat{P}$ is a demographically resampled, disjoint partition of the same cohort. Outcome is severe AKI (KDIGO ≥ 2); gaps are measured against a reference model trained on a balanced (50% male, 50% White) resample matching the target; mean [percentile interval] across 20 runs. DQI is a four-component partial index (q_T not measured; coverage 4/5).

Condition	Demographic	q_R	DQI	Risk gap	ECE gap
Baseline (natural mix)	61%M, 64%W	0.899	0.912 ± 0.002	0.001 [−0.005, 0.008]	0.019 [0.008, 0.034]
Moderate skew	80%M, 80%W	0.870	0.906 ± 0.003	0.011 [−0.003, 0.023]	0.021 [0.005, 0.043]
Severe skew	90%M, 90%W	0.770	0.880 ± 0.003	0.015 [−0.002, 0.038]	0.029 [0.016, 0.047]
Balanced (target-like)	50%M, 50%W	0.902	0.912 ± 0.002	0.003 [−0.006, 0.012]	0.014 [−0.002, 0.026]

The severe-skew condition has the lowest coverage ($q_R = 0.770$) and the coverage score recovers under the balanced condition ($q_R = 0.902$). The ECE gap rises monotonically across the baseline and two skew levels ($0.019 \rightarrow 0.021 \rightarrow 0.029$), with intervals excluding zero at each, and the composite DQI falls monotonically with skew severity ($0.912 \rightarrow 0.906 \rightarrow 0.880$), ranking the

conditions in the same order as the measured gap. The log-loss risk gap follows the same direction ($0.001 \rightarrow 0.011 \rightarrow 0.015$) but its intervals include zero at this sample size; we report this as directionally consistent, not confirmed. The absolute ECE steps (≤ 0.03) are small; we read the monotonic pattern as a directional consistency signal between the coverage measure and calibration, not as a clinically meaningful effect magnitude. Two reporting notes: the "balanced" condition is a different balanced resample from the reference model's, so its gap is a resampling difference rather than exactly zero—gaps are measured against a model trained on a balanced (target-matched) resample, so the balanced row can legitimately carry a small nonzero gap. And the rank correlations reported for this cohort (Section 4.9) rest on $n = 4$ conditions: with four untied observations even a perfect ordering has an exact two-sided permutation probability of $2/24 \approx 0.083$, so they are exploratory. Table 5 summarizes the cohort and Table 6 reports the absolute target-set performance underlying all gap computations.

Table 6: Absolute target-set performance for the primary logistic-regression model. "Ref." is the model trained on clean data (sepsis, mortality) or on the balanced target-matched resample (AKI); "Cond." is the model trained on the corrupted condition. All models are evaluated on the same held-out target. Values are means over 20 corruption-and-training runs.

		AUROC		Log-loss		ECE	
Cohort	Condition	Ref.	Cond.	Ref.	Cond.	Ref.	Cond.
Sepsis	Baseline	0.741	0.741	0.416	0.416	0.024	0.024
	Outcome-dependent missingness	0.741	0.728	0.416	0.423	0.024	0.031
	Institutional shift	0.741	0.735	0.416	0.543	0.024	0.134
	Joint	0.741	0.617	0.416	0.502	0.024	0.079
Mortality	Baseline	0.766	0.766	0.325	0.325	0.027	0.027
	Label noise	0.766	0.742	0.325	0.367	0.027	0.089
	Training-only contamination	0.766	0.767	0.325	0.328	0.027	0.035
	Joint	0.766	0.742	0.325	0.368	0.027	0.090
AKI	Baseline	0.822	0.818	0.481	0.483	0.047	0.066
	Moderate skew	0.822	0.812	0.481	0.492	0.047	0.068
	Severe skew	0.822	0.813	0.481	0.497	0.047	0.076
	Balanced	0.822	0.818	0.481	0.484	0.047	0.061

Absolute performance shows that the reported gaps are not an artifact of a weak reference: clean log-loss was 0.416, 0.325, and 0.481 in the sepsis, mortality, and AKI cohorts, whereas the worst condition reached 0.543, 0.368, and 0.497, respectively. AUROC and calibration moved in the same direction, but by smaller absolute amounts; for example, sepsis AUROC fell from 0.741 to 0.617 and ECE rose from 0.024 to 0.079 in the joint condition. The complete component-level run-to-run standard deviations are reported in the reproducibility package; the largest were q_R in the sepsis joint condition (0.009) and q_N in the AKI severe-skew condition (0.010).

4.8 Falsification and construct-validation results

Label-noise proxy falsification. On the separable data with 100% label flip, the proxy returned $q_N = 0.997$ (disagreement rate $\eta = 0.003$), failing to detect total label corruption—the logistic regression learns the flipped labels perfectly, so cross-validated disagreement is near zero. On the clean nonlinear two-moons task, $q_N = 0.884$ ($\eta = 0.123$), flagging model misspecification as label noise (false positive). Under asymmetric and instance-dependent noise on separable data, q_N tracks

the injected noise rate correlation $\rho \approx -1.0$. On real MIMIC mortality data with injected noise, q_N tracks the noise rate ($\rho = -0.996$ across the three injected levels $\{0.1, 0.25, 0.4\}$, a directional dose-response check rather than a calibration estimate). The proxy therefore works for its intended regime (asymmetric, instance-dependent noise on well-specified models) but has critical failure modes: it cannot detect full label flips on separable data, and it conflates model misspecification with label noise. These failures confirm that q_N is a *proxy* for label-disagreement, not an estimator of the true noise rate.

Completeness-weight validation. MI weights differ substantially from equal and clinically prespecified weights (top-3 features by MI: indices 15, 14, 4 vs. clinical top-3: labs). Bootstrap stability showed mean coefficient of variation 0.145 (max 0.263), indicating finite-sample estimator variability. The constant feature correctly received zero weight. Sensitivity to the k-NN parameter k was small (max mean absolute difference from $k = 5$ was 0.0074). Under MCAR (30% missingness), available-case MI mean absolute bias was 0.0064; under MAR (missingness depends on feature 0), mean absolute bias was 0.0103, confirming that MCAR removes selection distortion but finite-sample bias remains. This validates the statement in Table 1 that the MI estimator is not unbiased even under MCAR.

Construct-validation panel. Each component responded to its target corruption (Spearman $\rho \approx -1$: q_C vs missingness, q_N vs label noise, q_B vs shift, q_R vs shift) and ignored non-target corruptions ($\rho \approx 0$). This confirms the components are selective for their intended degradation types, though the response is measured on a single dataset with a single model family.

4.9 Comparison against naive quality heuristics

A practitioner already has single-scalar heuristics—raw missingness rate, class imbalance, and subgroup outcome-rate disparity. We compared each heuristic's Spearman correlation with the measured risk gap against DQI's own $\delta = 1 - \text{DQI}$; Table 7 reports the per-cohort correlations.

Table 7: Spearman rank correlation between each aggregate score and the measured risk gap, per cohort ($n = 4$ conditions). With $n = 4$ there are only 24 orderings, so per-cohort values are suggestive; the pattern across cohorts, not any single value, is the claim. In the sepsis cohort the imbalance and parity-gap monitors are constant across the four conditions (that cohort's condition set contains no label-noise or demographic-skew corruption), so their rank correlations are undefined. Exact two-sided permutation p-values for the DQI δ column are 0.333 in all three cohorts; the heuristic columns have p-values between 0.167 and 1.000 (permutation analysis in the reproducibility package). Correlations are computed from unrounded DQI values; the AKI value changes under rounding to three decimals (the displayed DQI values tie for two conditions).

Cohort	**DQI δ**	**Missingness**	**Imbalance**	**Parity gap**
Sepsis	0.800	−0.105	undefined	undefined
Mortality	0.800	−0.211	−0.949	−0.738
AKI	0.800	−0.800	0.800	0.800

DQI's δ is the only aggregate score in Table 7 with a positive (correct-direction) correlation in every cohort. The naive heuristics are each wrong-signed in at least one cohort, and in the mortality cohort all three point the wrong way at once: label noise shifts the observed class balance toward 50/50 even as the measured risk gap rises, so an imbalance monitor would read the noise condition as an *improvement*. The raw missingness rate is also insensitive to the mechanism of the missingness condition: in the sepsis cohort the outcome-dependent missingness corruption raises the average

missing-entry fraction from 3.6% to 15.7% (only four labs are affected) while producing a measured log-loss gap of 0.007, whereas the institutional-shift condition leaves missingness unchanged at 3.6% and produces the largest gap (0.127), so the raw rate ranks the two in the wrong order. This is a concrete, if preliminary, demonstration of why a multidimensional profile can differ from the heuristics a practitioner already monitors—but it does not, by itself, establish the superiority of any particular aggregation. Importantly, the composite DQI separates the clean baseline from corrupted conditions (detection), but when the baseline is removed the Spearman correlation between δ and the measured risk gap falls from 0.800 to 0.500 in both sepsis and mortality, indicating that the composite separates clean from corrupted but does not reliably order different harmful defects by their downstream harm.

4.10 External comparison against a different institution

All results above use a target proxy drawn from a held-out partition of the same MIMIC-IV cohort, which corrects for demographic or scenario composition within one institution but not for institution-specific bias, a failure mode documented across clinical prediction models [52, 53, 54]. We tested this directly with eICU-CRD, a multi-institution critical-care database (202 distinct hospitals in a random sample of 8,000 adult ICU stays, yielding $N = 7{,}973$ after exclusions). We restricted the feature set to the seven features with $< 30\%$ missingness in eICU (age, heart rate, respiratory rate, SpO_2, creatinine, BUN, WBC) plus sex, excluding mean blood pressure, temperature, and lactate, which are 78–92% missing in eICU; severity scores are excluded because the Sequential Organ Failure Assessment (SOFA) and Acute Physiology and Chronic Health Evaluation (APACHE) scores are different instruments; the external feature set is therefore a reduced, comparable subset, not the full MIMIC mortality feature set. We excluded rather than median-imputed these features because backfilling sparsely measured variables with constant MIMIC-IV medians would mask the shift.

We computed DQI for the MIMIC-IV mortality training data under two target proxies—the internal held-out split and eICU as an external target—and evaluated a single MIMIC-IV-trained model against both. In the eICU target, q_C and q_N are constant across sites and q_T is not measured (offset-based eICU fields support ordering checks but not the availability semantics q_T requires; Section 4.4), so site-level variation in δ is driven by q_B (Marginal Distributional Similarity) and q_R (Target-Support Coverage); the external score is a four-component partial index (hierarchical coverage: four top-level components available; temporal integrity unavailable). The composite degradation score barely changes between the two proxies (δ: $0.074 \to 0.076$). The pooled external performance differs (area under the receiver operating characteristic curve (AUROC) [55]: $0.752 \to 0.724$, drop 0.028; ECE $0.026 \to 0.030$, change $+0.004$), but neither change is distinguishable from zero given patient-sampling uncertainty (2,000-resample patient-level bootstrap 95% CI: AUROC drop $[-0.025, +0.077]$; ECE change $[-0.020, +0.014]$), so they should not be read as established deterioration. Log-loss appears to *improve* ($0.333 \to 0.289$); eICU's lower mortality prevalence (9.2% vs. 12.1%) complicates this comparison and may contribute, but we have not decomposed its causes. The composite therefore barely registers a shift that the pooled AUROC point estimate suggests the model may suffer.

The site-eligibility rule was also varied as a sensitivity analysis. At ≥ 30 stays and ≥ 5 events (46 sites), the association with ECE was $\rho = +0.44$ (bootstrap 95% CI $[+0.13, +0.69]$) and the association with $1 - \text{AUROC}$ was approximately zero ($\rho = -0.01$, $[-0.32, +0.28]$). At the primary $\geq 50/\geq 10$ rule (21 sites), the corresponding values were $+0.60$ $[+0.22, +0.84]$ and $+0.14$ $[-0.37, +0.59]$. At the strictest rule (≥ 100 stays, ≥ 20 events), only seven sites remained; the ECE interval included zero and the AUROC estimate was unstable. The calibration association is therefore threshold-sensitive

and exploratory, not a confirmed natural-shift result.

A hospital-level analysis makes this concrete. Only 21 of the 202 hospitals have ≥ 50 stays and ≥ 10 mortality events (an eligibility rule we imposed, not proof of stable per-site estimation), so the external sample yields about twenty analyzable sites, not 202 independent validations. Across those 21 sites, the composite degradation score is not detectably associated with per-site discrimination (Spearman $\rho = +0.14$ between δ and $1 - \text{AUROC}$, bootstrap 95% CI $[-0.37, +0.59]$; this interval permits appreciable positive association and is evidence of imprecision, not evidence of absence), while it is associated with per-site calibration (ECE: $\rho = +0.60$, CI $[+0.22, +0.84]$) and log-loss ($\rho = +0.58$, CI $[+0.17, +0.81]$; log-loss is base-rate confounded). This calibration association is, however, substantially confounded by shared sample-size and prevalence effects: both δ and ECE decrease with site size ($\rho \approx -0.55$ and -0.49) and increase with mortality prevalence ($\rho \approx +0.55$ and $+0.59$). The unadjusted Spearman correlation between δ and ECE is $+0.60$ $[+0.22, +0.84]$, but controlling for site size reduces it to $+0.45$ $[+0.05, +0.75]$, and controlling for both site size and mortality prevalence simultaneously attenuates it further to $+0.29$ $[-0.18, +0.72]$. The equal-mass ECE (10 bins, each containing roughly equal counts) yields a similar pattern: unadjusted $+0.55$, controlling for size $+0.35$, controlling for size and prevalence $+0.16$. When we further subsample every site to the same number of records (the minimum eligible size, 56) and recompute both δ and ECE, the correlation drops to $+0.17$ $[-0.22, +0.52]$ (median $+0.16$ over 100 replicates). These results indicate that much of the apparent δ–ECE association is explained by site size and prevalence; the residual association after equalizing sample size is small and its confidence interval includes zero. Individual components q_B (Marginal Distributional Similarity) and q_R (Target-Support Coverage) were negatively associated with site ECE (Spearman $\rho = -0.51$ and -0.46), consistent with higher measured quality indicating lower calibration error, but these component-level associations concern *two target-comparison proxies*, are exploratory, and are not adjusted for multiplicity.

A null experiment (300 replicates, 21 pseudo-sites of the same sizes drawn with replacement from the pooled eICU population) under the null hypothesis of no genuine between-site quality differences yields a median δ–ECE correlation of $+0.21$ $[-0.22, +0.65]$, with only 5.7% of null replicates exceeding the observed $+0.60$ ($P = 0.057$). This indicates that estimation artifacts alone—binning bias in ECE, noise in δ, shared dependence on site size—can produce a substantial positive correlation even when no genuine between-site quality differences exist. The observed $+0.60$ is at the 94th percentile of this null distribution.

To stabilize small-site estimates, we fit a partial-pooling (empirical-Bayes) model for the calibration intercept and slope [56]. The pooled calibration slope across all 21 sites is 0.75, indicating that predicted risks are too extreme on average (slope < 1 in the logistic calibration model). The per-site slope estimates are extremely variable (range 0.27–1.71) with median standard error 0.30; 81% of sites are shrunk more than halfway toward the pooled value, confirming the per-site estimates are extremely unstable. The shrunk per-site slopes range from 0.45 to 1.04. In the unpooled analysis, the per-site intercept and slope are both noisy (mean intercept $+0.03$, mean slope 0.84, median standard errors 0.29 and 0.30, respectively).

The honest summary is therefore: in the primary 21-site sample the tested quality and shift measures were associated with calibration (substantially attenuated by sample size and prevalence) but not detectably with discrimination; the calibration association was not robust at the seven-site threshold, and these analyses do not determine whether case mix, measurement differences, sampling variability, or other mechanisms explain the observed hospital-level performance variation. Table 9 (Appendix) reports the site-level inputs to these analyses.

Model-family and shift-measure sensitivity. The calibration-association and discrimination-null patterns were checked across model families and shift measures rather than only for the primary logistic-regression model. Recomputing the per-site analysis with MIMIC-trained gradient-

boosted and random-forest mortality models gave Spearman estimates of δ vs. $1 - \text{AUROC}$ of $+0.14$ $[-0.37, +0.59]$ (logistic regression), $+0.25$ $[-0.24, +0.62]$ (gradient-boosted trees), and $+0.21$ $[-0.25, +0.58]$ (random forest): all intervals include zero. Nonlinear domain classifiers separate MIMIC-IV from eICU substantially better than a linear logistic-regression comparator (mean per-site separation AUROC 0.72 for a boosted-tree classifier vs. 0.58 for the linear one; archived diagnostic in the reproducibility package), yet show no detectable association with per-site AUROC degradation (all paired percentile intervals span zero). The reported associations are therefore not an artifact of the single model family or of a single linear shift measure; conversely, none of the stronger nonlinear shift measures predicts per-site discrimination.

4.11 Predictive validity: leave-one-family-out evaluation

The controlled corruption experiments in Sections 4.5–4.7 are exploratory: the audit-to-performance mapping was computed on the same conditions reported. To evaluate out-of-sample predictive validity, we conducted a leave-one-family-out evaluation on the mortality cohort using a fivefamily $\times$ threeseverity corruption space (16 settings: baseline plus 15 corruptions across missingness, label noise, distribution shift, coverage restriction, and temporal leakage). The analysis plan and corruption manifest are documented in a written pre-specified analysis plan (dated 2026-10-19), included in the reproducibility package.

Experimental design. We constructed 16 settings on the MIMIC-IV mortality cohort (training $N = 2{,}100$, target $N = 900$). For each setting we computed the five component scores and the measured excess target log-loss (clipped to $[10^{-6}, 1 - 10^{-6}]$) relative to a clean reference model. A ridge regression from the five component scores to the excess log-loss was trained on four corruption families and evaluated on the heldout family (leaveonefamilyout). The ridge penalty was selected by leaveoneout crossvalidation on the development families from a grid $\{0.01, 0.1, 1, 10, 100\}$. The primary metric is the mean absolute error (MAE) on the heldout family's three severity levels. We compare against a suite of scalar baselines, each calibrated by the same affine transform fitted on the development data: constant (development-set mean), raw missingness rate, each single component ($\phi_C, \phi_N, \phi_B, \phi_R, \phi_T$), geometric scalar, arithmetic scalar, and minimum component. The best development-set baseline B^* is selected by leave-one-out cross-validation MAE on the development settings. We ran the evaluation twice: an initial specification that pooled all patient–setting observations when calibrating the ridge, and a revised patient-grouped specification that calibrates the ridge within development families (family-grouped inner validation) and groups observations by patient when computing summary statistics. The revision was adopted after the initial pooled run was computed (motivated by the patient-level pseudo-replication it revealed) and is documented as an addendum in the analysis plan; we report the patient-grouped specification as primary and the initial run for traceability.

Results. Figure 5 and Table 8 report the primary (patient-grouped) evaluation. On the held-out family, the ridge predictor yields MAE of 0.105 (missingness, $B^* = \text{minimum}$), 0.086 (label noise, $B^* = \text{minimum}$), 0.026 (shift, $B^* = \phi_C$), 0.016 (coverage, $B^* = \text{minimum}$), and 0.002 (leakage, $B^* = \text{arithmetic}$). The corresponding best baselines achieve 0.072, 0.033, 0.090, 0.014, and 0.069; the always-zero predictor achieves 0.091, 0.055, 0.102, 0.011, and 0.001. The advantage $G = \text{MAE}_{B^*} - \text{MAE}_{\text{ridge}}$ is -0.033 (missingness), -0.052 (label noise), $+0.064$ (shift), -0.002 (coverage), and $+0.067$ (leakage). The mean advantage versus the best scalar baseline is $+0.009$, with paired bootstrap 95% confidence interval $[-0.035, +0.051]$ over 10,000 resamples [57] (resampling the five family-level G values); versus the always-zero comparator the mean advantage is $+0.005$

(CI $[-0.020, +0.042]$). Both intervals span zero: the full component vector did not demonstrate superiority over the best scalar baseline or the always-zero comparator at $n = 5$ families; the comparison is inconclusive rather than negative.

Initial ungrouped run. In the initial specification, which pooled patient–setting observations across families, the mean advantage was +0.030 (SD 0.041) and the paired bootstrap interval was $[-0.003, +0.069]$. The larger point estimate reflects patient-level pseudo-replication that the grouped specification removes; we treat the ungrouped run as superseded and report it only for traceability. The advantage was concentrated in the shift family, and leakage and coverage corruptions had negligible degradation ($\widehat{\Delta} \leq 0.025$) in both specifications.

Interpretation. This is the primary predictive test; it is inconclusive at the feasible number of independent settings. The intervals spanning zero mean we cannot reject the null that the composite adds no predictive value beyond the best scalar baseline or the always-zero prediction. The advantage is concentrated in the shift family (where ϕ_B and ϕ_R are informative); the leakage and coverage corruptions have negligible degradation and contribute little. This is a pre-specified, inconclusive result: at the currently feasible number of independent settings, the evaluation does not establish that the composite beats a single scalar as a degradation predictor.

Limitations of the held-out evaluation. The 10,000 bootstrap resamples are drawn from five family-level G values, so they do not create 10,000 independent experimental settings. The five families share an underlying cohort, experimental infrastructure, and training data splits; the effective number of independent settings is $n = 5$. The bootstrap interval characterizes sampling variability of the mean advantage under the observed family-level distribution, not the variability over truly independent corruption mechanisms or institutions. This limitation is inherent to the corruption-family design and should be acknowledged when interpreting the confidence interval. With five families, the paired-bootstrap interval also spans the effect magnitudes observed for the ridge predictor on individual families (G from -0.052 to $+0.067$), so the design cannot resolve differences of that scale; "inconclusive" reflects the feasible sample of families, not evidence of no effect.

Per-family results (patient-grouped, family-grouped inner validation, zero comparator). The following table summarizes per-family MAE and advantage G with the always-zero comparator:

Table 8: Leave-one-family-out evaluation (patient-grouped, primary specification): MAE of ridge predictor vs. best scalar baseline B^* and always-zero comparator, with family-grouped inner validation. Positive G favors the profile versus B^*; G_0 versus the always-zero predictor. Mean row across the five families: mean $G = +0.009$ (95% paired-bootstrap interval $[-0.035, +0.051]$, 10,000 resamples); mean $G_0 = +0.005$ (interval $[-0.020, +0.042]$).

Held-out family	B^*	Ridge pen.	MAE(ridge)	MAE(B^*)	MAE(zero)	G	G_0
Missingness	Minimum	1.0	0.105	0.072	0.091	−0.033	−0.014
Label noise	Minimum	1.0	0.086	0.033	0.055	−0.052	−0.031
Shift	ϕ_C	10.0	0.026	0.090	0.102	+0.064	+0.076
Coverage	Minimum	1.0	0.016	0.014	0.011	−0.002	−0.004
Leakage	Arithmetic	1.0	0.002	0.069	0.001	+0.067	−0.001
Mean	—	—	0.047	0.056	0.052	+0.009	+0.005

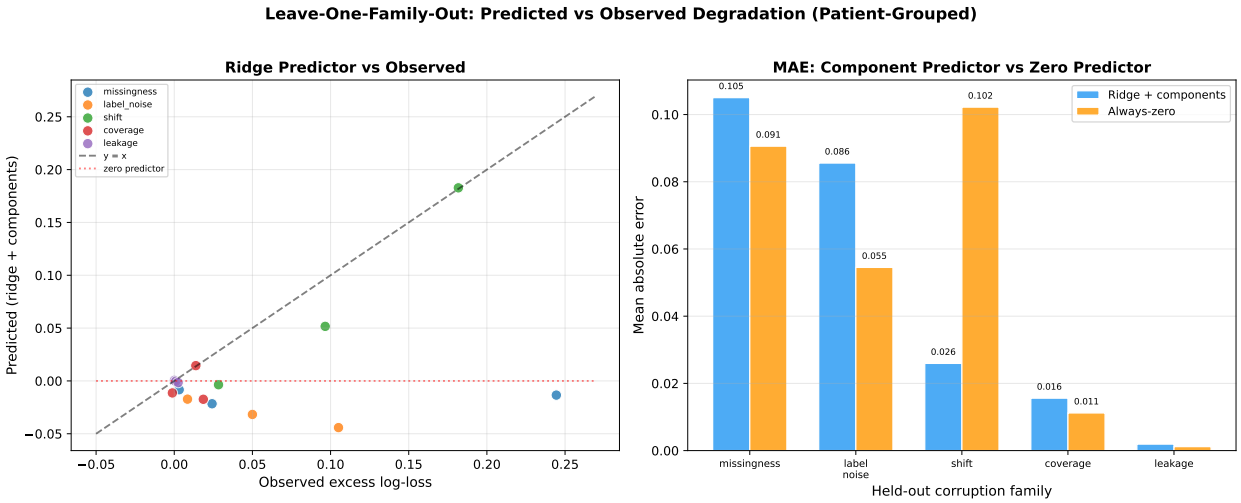

Figure 5: Leave-one-family-out evaluation (patient-grouped): predicted versus observed excess log-loss degradation, log scale. Each marker is one corruption setting; the ridge predictor (Section 4.11) is compared against the best scalar baseline B^*, the always-zero predictor, and the clean-data reference. The marker placement reproduces the same per-family MAE values as Table 8. Mean values in Table 8 aggregate the per-family entries.

The shift family drives the positive mean advantage; missingness, label noise, and coverage show no advantage over either comparator. Leakage shows a moderate advantage versus B^* but with only three severity levels of small effect ($\widehat{\Delta} \leq 0.006$), and it loses to the always-zero predictor. The full component vector's advantage is therefore driven by the shift family, where the bias and representativeness components are informative.

Comparison with baselines. The best scalar baselines by family are the minimum (missingness), the minimum (label noise), ϕ_C (shift), the minimum (coverage), and the arithmetic mean (leakage). The composite vector outperforms the best scalar on two of five families (shift and leakage); on three families the best scalar is better. Against the always-zero comparator, the profile wins in one family (shift) and loses in four (missingness, label noise, coverage, leakage). The predictive value of the full component vector is therefore not established at the feasible number of independent corruption families.

5 Discussion

A single distinction organizes these results: *detecting a dataset defect*, *predicting its effect on model performance*, and *the usefulness of a one-number quality score* are three different problems. The results below show the profile can do the first (partially), does the second only weakly and inconsistently, and does not establish the third.

5.1 Principal findings

The profile is directionally consistent with measured degradation in the three controlled cohorts. Institutional covariate shift is the largest single-defect effect in the sepsis cohort (log-loss gap 0.127), and label noise (coding error) is the largest in the mortality cohort (0.042); the outcome-dependent missingness condition in sepsis produces only a small measured log-loss gap (0.007) while

lowering the composite score from 0.911 to 0.849. The corresponding completeness and distributional components move with these conditions: q_C falls from 0.919 to 0.737 under the missingness corruption and q_B falls from 0.992 to 0.540 under the shift. The Marginal Distributional Similarity component measures a marginal divergence between the training and target feature distributions; the completeness component weights missingness by feature–outcome mutual information estimated from the clean reference; and the Target-Support Coverage component measures coverage of the target by the training support. Among the four aggregate scores compared in Table 7, the composite DQI is the only one that moves in the expected direction in all three cohorts. The temporal-integrity component detects the training-only contamination and over-ranks its small measured downstream harm; in the train-and-test experiment it also detected leakage that inflated apparent AUROC by 0.025, while clean deployment harm was concentrated in log-loss, with a smaller calibration difference whose interval included zero (Table 4). The ordering sub-component remains unmeasured.

The evidence does not support nonlinear compounding. Paired interaction contrasts computed on matched corruption realizations give, for the sepsis cohort, $I_{\text{risk}} = -0.048$ (95% percentile interval $[-0.060, -0.030]$, sub-additive) and $I_{\text{ECE}} = -0.060$ (interval $[-0.077, -0.046]$, sub-additive); for the mortality cohort both interactions are sub-additive ($I_{\text{risk}} = -0.003$, interval $[-0.004, -0.001]$; $I_{\text{ECE}} = -0.007$, interval $[-0.013, -0.001]$). These data therefore provide no evidence of a super-additive (compounding) interaction, and the geometric aggregator's algebra does not, by itself, license a claim that clinical harms multiply.

Under natural cross-institution shift, the profile is associated with calibration in the primary sample but not detectably with discrimination. In the 21-site eICU analysis the composite degradation score was not detectably associated with per-site AUROC degradation (bootstrap 95% CI spanning zero), while it was associated with per-site calibration and log-loss. The components themselves were associated with calibration in the expected direction (q_B vs. ECE $\rho = -0.51$; q_R vs. ECE $\rho = -0.46$). This calibration association is, however, substantially attenuated by shared sample-size and prevalence effects: the unadjusted δ–ECE correlation of +0.60 attenuates to +0.45 after controlling for site size, and to +0.29 after additionally controlling for mortality prevalence; under equal-size subsampling (every site downsampled to $n = 56$) it drops to +0.17 (CI $[-0.22, +0.52]$, median +0.16 over 100 replicates). A null experiment (300 replicates, 21 pseudo-sites drawn with replacement from the pooled eICU population) yields a median δ–ECE correlation of +0.21 $[-0.22, +0.65]$; the observed +0.60 is at the 94th percentile ($P = 0.057$), indicating that estimation artifacts (ECE binning bias, noise in δ, shared dependence on site size) can produce a substantial positive correlation even without genuine between-site quality differences. To stabilize small-site estimates, we fit a partial-pooling (empirical-Bayes) model for the calibration intercept and slope; the pooled calibration slope is 0.75 (indicating that predicted risks are too extreme; the fitted model is $\text{logit}(\hat{p}) = a + b \cdot \text{logit}(p)$ with $b = 0.75$), and 81% of sites are shrunk more than halfway toward the pooled slope of 0.75, confirming extreme instability of per-site estimates. The components q_B (Marginal Distributional Similarity) and q_R (Target-Support Coverage) were associated with calibration in the expected direction (q_B vs. ECE $\rho = -0.51$; q_R vs. ECE $\rho = -0.46$), but this evidence concerns *two target-comparison proxies*, not external validation of a five-component profile. The discrimination association was not stable at any threshold; the calibration association fell from +0.60 at 21 sites (the primary rule) and +0.44 at 46 sites (the relaxed rule) to +0.17 under equal-size subsampling, and a null experiment ($P = 0.057$) showed that estimation artifacts alone can produce a correlation of +0.60. The lack of a detectable discrimination association is consistent across the three model families and across linear and nonlinear shift measures (model-family and shift-measure sensitivity are reported with the eICU analyses). These analyses do not determine why hospital-level discrimination varies: case mix, measurement differences, and sampling variability are all plausible and untested explanations, and only 21 analyzable sites limit power. The wide

intervals mean that these results are imprecise, not that an association is absent.

The scalar is not justified; the profile is the defensible artifact. Three results undercut the scalar. First, the multiplicative form is not empirically supported: geometric and arithmetic aggregation give identical condition rankings in every cohort at $n = 4$ (an empirical observation, not a mathematical equivalence between the aggregators), and the minimum aggregator ranks worse in two of three cohorts. Second, the scalar is partial wherever q_T is unmeasured; coverage is reported, but scores with different coverage are not directly comparable. The scalar is also largely redundant with its strongest components: q_B and q_R each rank as well as the composite in sepsis and AKI, while q_C and q_N are weak or wrong-signed in two cohorts each; in mortality, q_N and q_B carry most of the rank signal, so no fixed pair dominates across cohorts. Third, all per-cohort correlations rest on $n = 4$ conditions and none is statistically distinguishable from zero (exact permutation $p \geq 0.083$), and a learned-weight fit places all weight on q_N and q_R and zero on q_C, q_B, and q_T; the fit is exploratory and was not externally validated. A leave-one-family-out evaluation (Section 4.11) addresses this at the predictive level: a ridge predictor of excess log-loss from the full component vector does not provide evidence of outperforming the best single scalar or the always-zero prediction (patient-grouped mean advantage +0.009, 10,000-resample paired-bootstrap 95% CI $[-0.035, +0.051]$ versus the best scalar; +0.005, CI $[-0.020, +0.042]$ versus zero, at $n = 5$ families). The empirical signal therefore lives in a small, cohort-dependent subset of components, not in a five-dimensional multiplicative index, and we do not endorse any single aggregation.

5.2 Regulatory positioning

We do not claim that the profile satisfies any regulatory requirement. Data-quality artifacts of the kind described here may support selected audit activities, but the mapping to regulatory principles must be accurate. The 2021 Good Machine Learning Practice guiding principles identify representative datasets as Principle 3, training-test independence as Principle 4, and deployed-model monitoring as Principle 10 (with Principle 2 concerning software engineering and security, and Principle 6 concerning model design suited to the available data and intended use) [20]. A crosswalk of the five components to these principles is a *documentation aid*, not a compliance determination, and no threshold on any scalar index is an evidence-based gate for regulatory authorization. We position the profile as a potential pre-model diagnostic—a *potential supporting diagnostic*, not a necessary condition for reliability: a dataset with a favorable profile does not guarantee a reliable model, because model architecture, optimization, and calibration also determine performance.

5.3 Potential operational role and untested utility

When the final model is itself logistic regression, a natural question arises: why is an audit containing cross-validated logistic regression preferable to directly evaluating a simple candidate model? The audit's value is not in replacing model evaluation—which remains the gold standard for performance estimation—but in *prioritization* and *interpretability* before expensive model training.

First, the audit provides *prioritization*: it identifies which data-quality dimensions are most degraded before any model is trained, allowing a practitioner to direct remediation effort toward the most degraded dimensions. Its own results here, however, show that this is *detection*, not a reliable ranking by downstream harm: once the clean baseline is removed, the composite score does not consistently order defects by their measured gaps (Section 4.9). For example, in our sepsis cohort the completeness component q_C dropped from 0.919 to 0.737 under outcome-dependent missingness, while the bias component q_B remained near 0.99; this immediately flagged missingness as the primary concern without training a single model.

Second, the audit provides *interpretability*: the component scores decompose *why* performance might degrade. When the sepsis joint condition yielded a log-loss gap of 0.086, the component breakdown revealed $q_C = 0.737$ and $q_B = 0.54$, attributing the harm primarily to distributional shift rather than missingness alone. A single scalar metric from direct model evaluation cannot provide this diagnostic granularity.

Third, the audit enables *early detection* without outcome labels. The bias component q_B and representativeness component q_R require only feature distributions and a target reference, not outcome labels. This allows quality assessment of unlabeled data (e.g., a new hospital's unlabeled cohort) before any labels are available—a scenario where direct model evaluation is impossible.

The audit does not replace final model validation. If the final model is logistic regression, the audit uses cross-validated logistic regression internally, which shares the same model class. In this sense, the audit is not a computational shortcut for performance estimation; it is a *prioritization and diagnostic layer* that informs *where to look* and *what to fix* before investing in full model training and evaluation. We do not claim the audit is universally cost-effective; its value depends on the cost of remediation actions it enables. Where remediation is inexpensive (e.g., enabling a missing lab order), the audit's cost is easily justified; where remediation is infeasible (e.g., correcting historical label noise), the audit serves primarily as a risk disclosure.

5.4 Limitations

Exploratory, not confirmatory. The three-cohort audit-to-performance mapping was computed on the same conditions reported here. The hospital-level eICU analysis and the leave-one-family-out evaluation (Section 4.11) are held out in a limited sense, but they do not establish out-of-sample predictive validity across new institutions or scenarios, and the correlations in Section 4.9 rest on four conditions per cohort. A held-out, scenario-level evaluation is required before any predictive claim.

Single health system plus one external database. The three cohorts originate from Beth Israel Deaconess Medical Center (MIMIC-IV). The eICU comparison uses one external database, and the hospital-level analysis has only 21 analyzable sites, so its power is limited and the per-site performance estimates are noisy [58].

Partial temporal integrity. The leakage sub-component of q_T is measured from real measurement timestamps (availability times relative to the prediction window), but the ordering and causal-consistency sub-components require per-event timestamp sequences that were not extracted, so they remain unmeasured. Within q_T, unmeasured sub-components are excluded rather than assigned a perfect score. At the composite level, when q_T is not measured we exclude it and renormalize the remaining weights; the resulting score is a partial four-component index (coverage 4/5), not a five-component score. Coverage is reported, and scores with different coverage are not directly comparable. This prevents a missing audit from appearing perfect, but it does not make the partial scalar a validated full-profile index.

Leakage design. The train-and-test experiment substitutes late-window creatinine and BUN only, in one cohort; it demonstrates a mechanism, not a general leakage rate or a full audit of all temporal violations. The training-only condition is a robustness test against a contaminated training set, not a standard leakage design; the table labels it as training-only contamination.

Site-threshold sensitivity. The hospital-level eICU result depends on the eligibility rule. At ≥ 30 stays and ≥ 5 events the calibration association was $\rho = +0.44$; at the primary $\geq 50/\geq 10$ rule it was $+0.60$; at $\geq 100/\geq 20$ only seven sites remained and the interval included zero. The calibration association is therefore exploratory and threshold-sensitive, and the discrimination association was not stable at any threshold.

eICU calibration analysis and sample-size confounding. The eICU calibration association is substantially confounded by site size and mortality prevalence: the unadjusted δ–ECE correlation of $+0.60$ attenuates to $+0.45$ after controlling for site size, to $+0.29$ after additionally controlling for mortality prevalence, and to $+0.17$ under equal-size subsampling (CI $[-0.22, +0.52]$). A null experiment (300 replicates) yielded a median null correlation of $+0.21$, with the observed $+0.60$ at the 94th percentile ($P = 0.057$), indicating that estimation artifacts (ECE binning bias, noise in δ, shared dependence on site size) can produce the observed association even without genuine between-site quality differences. The calibration intercept and slope per site are noisy (median SE 0.29 and 0.30); partial pooling shrinks 81% of site slopes more than halfway toward the pooled slope of 0.75, confirming extreme instability. The evidence concerns two target-comparison proxies (q_B, q_R), not external validation of a five-component profile. The 8,000-stay extraction cap was chosen a priori to ensure computational tractability; expanding it would require re-extraction and re-analysis. Given that the calibration association attenuated under adjustment and equal-size subsampling, and that the null experiment showed this mechanism can generate positive correlations, the association should be interpreted as sensitive to evaluation design rather than as definitively artifactual or definitively genuine. A larger extraction would increase the number of eligible sites but also increase heterogeneity in site sizes and prevalence, likely amplifying the same confounding. This limitation is noted rather than resolved by further extraction.

Estimator approximations. The Marginal Distributional Similarity component (q_B) uses marginal Jensen–Shannon divergence (averaged over features) rather than a joint divergence on a dimensionality-reduced projection; this marginal-versus-joint mismatch means q_B captures per-feature distributional shifts but misses dependence-structure changes between features. The completeness weights are per-feature available-case mutual-information estimates, which are generally biased unless missingness is missing-completely-at-random (an identifiability limitation under missing at random (MAR) and missing not at random (MNAR)); the Target-Support Coverage component is a target-weighted kernel-coverage proxy, not a uniform manifold-volume integral; and the noise component is a cross-validation disagreement proxy, not the true injected noise rate. These components are operational proxies for the idealized definitions, not estimators of them.

Construct validity. The components are operational proxies, not validated measures of the constructs named in the profile. The label-noise component is model–label disagreement, not an estimated noise rate; a falsification experiment (Experiments F1–F5) showed that the proxy fails to detect 100% label flip on separable data ($q_N \approx 0.997$) and flags clean nonlinear data as noisy ($q_N \approx 0.884$ on moons). The completeness weights are available-case mutual information with selection distortion under MNAR and finite-sample bias even under MCAR (Experiments F6–F10: mean abs bias 0.005 MCAR, 0.010 MAR; bootstrap CV 0.15 max 0.32). A construct-validation panel (Experiment F11) tested each component against known corruption severities: q_C detects missingness (Spearman $\rho = -1.00$), q_N detects label noise ($\rho = -0.998$), q_B detects shift ($\rho = -0.995$), and q_R detects shift ($\rho = -0.962$); each component ignores the corruptions it is not designed for (q_C ignores label noise/shift, q_N ignores missingness/shift, q_B/q_R ignore missingness/label noise). The Marginal Distributional Similarity component is an average of marginal Jensen–Shannon divergences, not a joint divergence; the Target-Support Coverage component depends on a kernel bandwidth and a target-weighted coverage definition. We did not independently validate these proxies against ground-truth corruption mechanisms, adjudicated labels, or external quality ratings. The profile should therefore be read as a diagnostic checklist whose components have known failure modes, not as a measurement instrument.

Model diversity. We checked the findings across three model families—logistic regression, gradient-boosted trees, and random forests. The calibration ranking is directionally consistent across cohorts and families (exploratory rank correlations at $n = 4$ conditions): δ versus the calibration gap

is $\rho = 0.8$–1.0 per cohort and model at $n = 4$ conditions. The magnitude of the calibration gap does, however, differ by model and cohort rather than uniformly favoring one family: in the sepsis joint condition the gradient-boosted-tree ECE gap is 0.19 versus 0.055 for logistic regression, whereas in the mortality joint condition it is 0.045 versus 0.063, and the random-forest risk-gap ordering agrees weakly with the composite ($\rho = 0.4$–0.8). We therefore report logistic regression as the primary model, treat the tree results as a sensitivity check, and note that the quality-to-performance mapping is itself model-dependent.

Sample size and uncertainty. Intervals are percentiles over 20 corruption-and-training runs, not confidence intervals for a population parameter. The target sample is fixed across runs, so these intervals are conditional on a single target split; patient versus stay versus corruption versus model-fit variance is not separately propagated in the main results, and component-level population intervals were not estimated; run-to-run component standard deviations are reported in the reproducibility package. A paired patient-cluster bootstrap was implemented to separate corruption-run variance, patient-sampling variance, and their interaction; the framework is documented in the reproducibility package and the variance decomposition is $\mathrm{Var}(\Delta) = \mathrm{Var}_{\mathrm{corruption}} + \mathrm{Var}_{\mathrm{patient}} + \mathrm{Var}_{\mathrm{interaction}}$. In the main results, only the corruption-run variance (20 runs on a fixed target) is reported; the full patient-cluster bootstrap (1,000 resamples $\times$ model retraining) is documented in the reproducibility package for future extension. The institutional-shift corruption is deterministic (no random draw), so its 20 runs are identical and its gap carries no run-to-run interval—that absence of an interval reflects the deterministic corruption, not measured precision. Event counts and site heterogeneity are not accounted for.

Patient-level independence. The train/target splits were performed at the stay level, not the patient level. A patient-level audit revealed that the same `subject_id` appears in both train and target partitions: 14 shared patients in sepsis (1.0% of train, 2.4% of target), 24 in mortality (1.2%, 2.7%), and 43 in AKI (2.8%, 6.3%). The three MIMIC-IV cohorts also share substantial patient populations (e.g., 73.8% of sepsis patients also appear in the mortality cohort; 532 patients appear in all three). This means the within-cohort and cross-cohort comparisons are not fully independent at the patient level, and the reported uncertainty underestimates the true variance. A patient-level grouped split was not used in this exploratory benchmark; this is a material limitation for any generalization claim.

Reporting. This report does not fully conform to the Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis plus artificial intelligence extension (TRIPOD+AI) [59]; supplementary cohort-reconstruction structured query language (SQL), participant-flow counts, construct-validation panel details, leave-one-family-out per-severity results, calibration intercept/slope per site, and TRIPOD+AI checklist are provided in the Supplementary Material, included in the same reproducibility package as the analysis plan (access details provided at submission).

5.5 Additional analyses and held-out evaluation

The three-cohort analyses in Section 4 are exploratory (their audit-to-performance mapping was computed on the same conditions reported). Two additional analyses were specified before they were run, in a written analysis plan: the hospital-level eICU evaluation (Section 4.10) and a leave-one-family-out evaluation on the mortality cohort with a five-family $\times$ three-severity corruption space (16 settings). The leave-one-family-out evaluation is reported in Section 4.11; the pre-specified corruption space, including its missingness selection equation, is documented in the analysis plan and is a different corruption from the exploratory sepsis missingness condition in Section 4.5 (the two are not nested). The eICU hospital-level analysis (Section 4.10) complements the internal

evaluation by testing the profile on an external multi-institution database.

The following analyses remain prerequisites for any further predictive or actionable claim:

1. **Leave-one-component-out ablations.** Quantify each component's marginal contribution to the predictive mapping beyond the descriptive ranking reported here.
2. **Site-held-out evaluation.** With more independent source institutions, reserve complete sites for final evaluation and report site-held-out and jointly held-out results separately.
3. **False reassurance.** Estimate the probability of a harmful condition among settings the audit passes, together with the pass rate and the complementary miss rate, compared at matched workload.
4. **Remediation.** A budget-matched comparison of an audit-guided intervention policy against simple and random policies over a common action menu.

If the scalar loses to its components in this evaluation, the profile is the contribution; if it loses to simpler heuristics, that is a finding to be reported, not concealed.

5.6 Conclusion

We report an exploratory benchmark of a task- and target-specific multidimensional data-quality profile for clinical prediction, evaluated against measured changes in model performance under controlled and natural shifts. The profile's degradation score shows positive within-cohort rank associations with excess loss in the three controlled cohorts, but it does not reliably order all defects by their downstream impact; the temporal-contamination conditions illustrate the difference between detecting a prohibited input and estimating its magnitude of harm. Its calibration signal is directionally similar across model families in the evaluated cohorts (exploratory $\rho = 0.8$–1.0 per cohort and model at $n = 4$ conditions), though the magnitude of the calibration gap differs by model and cohort. In the primary 21-site eICU sample the profile was associated with calibration but not detectably with discrimination degradation; the calibration association was not robust at the seven-site threshold. Three results undercut the scalar: the interaction contrasts show no evidence of positive (compounding) interaction (all are sub-additive in the evaluated cohorts), geometric and arithmetic aggregation are indistinguishable at $n = 4$ (an empirical rank-agreement observation, not an equivalence), and the signal is concentrated in a small, cohort-dependent subset of components. The components are operational proxies, not validated measurement instruments, and scores with incomplete coverage are partial indices. We therefore make no claim that a scalar reliability index is justified. The contribution of this paper is a transparent benchmark of which data-quality signals do and do not track downstream degradation, with construct-validity, availability, and negative results reported explicitly.

Abbreviations

AKI: acute kidney injury; AUROC: area under the receiver operating characteristic curve; BUN: blood urea nitrogen; DQI: data-quality index; ECE: expected calibration error; EHR: electronic health record; eICU-CRD: eICU Collaborative Research Database; GMLP: good machine-learning practice; ICU: intensive care unit; KDIGO: Kidney Disease: Improving Global Outcomes; MIMIC-IV: Medical Information Mart for Intensive Care IV; MNAR: missing not at random; OHDSI: Observational Health Data Sciences and Informatics; SOFA: Sequential Organ Failure Assessment; WBC: white blood cell count; YY: Yunguo Yu.

Declarations

Ethics approval and consent to participate

This study uses de-identified MIMIC-IV data. Collection of the underlying patient data and its release as a research resource was reviewed by the Institutional Review Board at Beth Israel Deaconess Medical Center, which granted a waiver of informed consent (IRB #2001P001699). eICU-CRD is also de-identified and was accessed under its PhysioNet data-use agreement.

Consent for publication

Not applicable.

Availability of data and materials

MIMIC-IV and eICU-CRD are de-identified critical-care databases available through credentialed access on PhysioNet (physionet.org) following completion of a data-use agreement and required training. We do not redistribute row-level data. A reproducibility package containing cohort SQL queries, concept-code versions, preprocessing and corruption functions, split-generation code, dependency locks, configuration files, and permissible aggregate run-level outputs is available from the corresponding author on reasonable request; reviewer access is provided on request. Raw patient records are not included per PhysioNet data-use agreements.

Code availability

The analysis code (cohort extraction, the data-quality instrument, and the validation scripts) is included in the reproducibility package (available from the corresponding author on reasonable request). The package contains the cohort extraction, profile, and validation scripts, dependency locks (`requirements_lock.txt`, `environment.yml`), configuration files, and cohort SQL queries. All experiments used fixed seeds (global `set_seed(42)`, with analysis-specific seeds recorded in the scripts) and are logged with their configuration; the environment was Python 3.12 with NumPy 2.0, pandas 3.0, scikit-learn 1.8, and SciPy 1.17.

Competing interests

The authors declare that they have no competing interests.

Funding

The research received no external funding.

Authors' contributions

The author conceived the study, designed the data-quality profile, implemented the analyses, and drafted the manuscript. The author read and approved the final manuscript.

Acknowledgements

Not applicable.

A Site-level eICU estimates

Table 9 reports the site-level inputs to the hospital-level analysis. The estimates are unpooled and noisy; they are shown for transparency, not as independent validations.

Table 9: Site-level eICU estimates for the 21 eligible hospitals.

Hospital	Stays	Deaths	Prevalence	δ	q_B	q_R	AUROC	ECE
73	297	31	0.104	0.079	0.980	0.838	0.646	0.060
264	235	35	0.149	0.079	0.983	0.836	0.742	0.039
167	234	18	0.077	0.061	0.984	0.903	0.796	0.023
243	196	14	0.071	0.076	0.977	0.851	0.815	0.043
420	187	27	0.144	0.080	0.981	0.835	0.653	0.050
199	166	26	0.157	0.082	0.983	0.827	0.647	0.052
338	161	11	0.068	0.076	0.986	0.847	0.731	0.051
458	155	20	0.129	0.097	0.969	0.786	0.618	0.077
208	150	27	0.180	0.083	0.976	0.828	0.607	0.075
411	149	15	0.101	0.080	0.976	0.841	0.800	0.039
252	148	21	0.142	0.067	0.981	0.882	0.868	0.064
188	134	15	0.112	0.079	0.977	0.841	0.639	0.037
300	134	13	0.097	0.074	0.979	0.858	0.633	0.021
443	133	18	0.135	0.105	0.957	0.766	0.777	0.047
122	119	13	0.109	0.082	0.967	0.840	0.683	0.036
157	97	15	0.155	0.107	0.963	0.755	0.653	0.085
417	84	10	0.119	0.092	0.962	0.806	0.805	0.087
110	83	12	0.145	0.082	0.951	0.853	0.698	0.056
307	80	10	0.125	0.088	0.956	0.827	0.651	0.094
142	78	14	0.179	0.097	0.959	0.794	0.842	0.091
226	56	12	0.214	0.080	0.956	0.856	0.775	0.124

B Bounded-Divergence Inequalities

Let $R_P(f) = \mathbb{E}_{(x,y)\sim P}[\ell_f(x,y)]$ be the expected loss of a classifier f under distribution P. For a loss bounded in $[0,1]$, the difference in expected loss between two distributions is controlled by their total-variation (TV) distance, which is in turn controlled by the Jensen–Shannon divergence:

Proposition 1 (Bounded-divergence inequality)**.** *For any measurable loss* $0 \le \ell_f \le 1$*,*

$$|R_P(f) - R_Q(f)| \;\le\; \mathrm{TV}(P,Q) \;\le\; \sqrt{2\ \mathrm{JS}(P\|Q)}. \tag{11}$$

This is a standard consequence of the total-variation bound on differences of expectations and the relationship $\mathrm{JS} \ge \frac{1}{2}\,\mathrm{TV}^2$. If the Marginal Distributional Similarity component is defined as $q_B = 1 - \mathrm{JS}(Q\|\hat{P})/\ln 2$, and a population score $s = \prod_k q_k^{w_k}$ has $w_B > 0$, then—because the other factors are at most 1—we have $\mathrm{JS}(Q\|\hat{P}) \le \ln 2\,(1 - s^{1/w_B})$, and combining with the proxy triangle inequality gives

Proposition 2 (Reference-relative bound)**.** *For* $0 \le \ell_f \le 1$ *and* $w_B > 0$*,*

$$|R_P(f) - R_Q(f)| \;\le\; \mathrm{TV}(P,\hat{P}) + \sqrt{2\ln 2\left(1 - s^{1/w_B}\right)}. \tag{12}$$

Proposition 2 is a standard-divergence consequence, not a new discovery, and it does not validate the other four components. Its interpretation is bounded by explicit restrictions:

- **Joint versus marginal.** Covariate divergence alone does not control label or concept shift; the bound requires divergence on the full relevant joint distribution.
- **Population versus estimate.** A projected, finite-sample divergence estimator has its own approximation and estimation error, which the bound does not account for.
- **Proxy error.** $\mathrm{TV}(P, \hat{P})$ is unknown unless additional evidence bounds it.
- **Loss range.** The propositions require a loss bounded in $[0, 1]$. Binary log-loss is not such a loss: even after clipping probabilities to $[\epsilon, 1-\epsilon]$ it is bounded only by $B = -\log \epsilon$ (here $B \approx 13.8$). The propositions therefore illustrate the divergence-to-loss relationship for a bounded loss and do not quantitatively bound the clipped log-loss gaps reported empirically; applying them to log-loss would require normalization by B.
- **Non-monotonicity.** A larger upper bound does not imply that actual risk worsens monotonically; the bound permits worsening, it does not require it.
- **Non-vacuity.** The bound may be too loose to guide decisions and must not be presented as a clinical certificate.

These restrictions, not a theorem, frame the empirical question in Section 4: does the profile (or any aggregation of it) carry predictive information about measured degradation, above and beyond simpler alternatives?

References

[1] Christopher W Seymour, Vincent X Liu, Theodore J Iwashyna, Frank M Brunkhorst, Thomas D Rea, André Scherag, Gordon Rubenfeld, Jeremy M Kahn, Manu Shankar-Hari, Mervyn Singer, et al. Assessment of clinical criteria for sepsis: for the Third International Consensus Definitions for Sepsis and Septic Shock (Sepsis-3). *JAMA*, 315(8):762–774, 2016. doi: 10.1001/jama.2016.0288.

[2] Alvin Rajkomar, Eyal Oren, Kai Chen, Andrew M. Dai, Nissan Hajaj, Michaela Hardt, et al. Scalable and accurate deep learning with electronic health records. *npj Digital Medicine*, 1(1):18, 2018. doi: 10.1038/s41746-018-0029-1.

[3] Eric J. Topol. High-performance medicine: the convergence of human and artificial intelligence. *Nature Medicine*, 25(1):44–56, 2019. doi: 10.1038/s41591-018-0300-7.

[4] Ziad Obermeyer, Brian Powers, Christine Vogeli, and Sendhil Mullainathan. Dissecting racial bias in an algorithm used to manage the health of populations. *Science*, 366(6464):447–453, 2019. doi: 10.1126/science.aax2342.

[5] Aaron Boussina, Supreeth P. Shashikumar, Atul Malhotra, Robert L. Owens, Robert El-Kareh, Christopher A. Longhurst, et al. Impact of a deep learning sepsis prediction model on quality of care and survival. *npj Digital Medicine*, 7(1):14, 2024. doi: 10.1038/s41746-023-00986-6.

[6] Mark P. Sendak, William Ratliff, Dina Sarro, Elizabeth Alderton, Joseph Futoma, Michael Gao, et al. Real-world integration of a sepsis deep learning technology into routine clinical care: implementation study. *JMIR Medical Informatics*, 8(7):e15182, 2020. doi: 10.2196/15182.

[7] Rose Sisk, Matthew Sperrin, Niels Peek, Maarten van Smeden, and Glen Philip Martin. Imputation and missing indicators for handling missing data in the development and deployment of clinical prediction models: A simulation study. *Statistical Methods in Medical Research*, 32(8):1461–1477, 2023. doi: 10.1177/09622802231165001.

[8] Alexandre Perez-Lebel, Gaël Varoquaux, Marine Le Morvan, Julie Josse, and Jean-Baptiste Poline. Benchmarking missing-values approaches for predictive models on health databases. *arXiv preprint arXiv:2202.10580*, 2022. doi: 10.48550/arxiv.2202.10580.

[9] Yunguo Yu. What the doctor didn't order: Informative missingness as a clinical signal in ICU prediction models. *Preprints (MDPI)*, 2026. doi: 10.20944/preprints202606.1130.v1. Preprint.

[10] Hwanjun Song, Minseok Kim, Dongmin Park, Yooju Shin, and Jae-Gil Lee. Learning from noisy labels with deep neural networks: A survey. *IEEE Transactions on Neural Networks and Learning Systems*, 34 (11):8135–8153, 2023. doi: 10.1109/TNNLS.2022.3152527.

[11] Jeremy Irvin, Pranav Rajpurkar, Michael Ko, Yifan Yu, Silviana Ciurea-Ilcus, Chris Chute, et al. CheXpert: A large chest radiograph dataset with uncertainty labels and expert comparison. *Proceedings of the AAAI Conference on Artificial Intelligence*, 33(01):590–597, 2019. doi: 10.1609/aaai.v33i01.3301590.

[12] Curtis G. Northcutt, Anish Athalye, and Jonas Weylin Mueller. Pervasive label errors in test sets destabilize machine learning benchmarks. *arXiv preprint arXiv:2103.14749*, 2021. doi: 10.48550/arxiv.2103.14749.

[13] Laleh Seyyed-Kalantari, Haoran Zhang, Matthew B. A. McDermott, Irene Y. Chen, and Marzyeh Ghassemi. Underdiagnosis bias of artificial intelligence algorithms applied to chest radiographs in under-served patient populations. *Nature Medicine*, 27(12):2176–2182, 2021. doi: 10.1038/s41591-021-01595-0.

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

[15] Alvin Rajkomar, Michaela Hardt, Michael D. Howell, Greg Corrado, and Marshall H. Chin. Ensuring fairness in machine learning to advance health equity. *Annals of Internal Medicine*, 169(12):866–872, 2018. doi: 10.7326/m18-1990.

[16] Luke Oakden-Rayner, Jared Dunnmon, Gustavo Carneiro, and Christopher Re. Hidden stratification causes clinically meaningful failures in machine learning for medical imaging. In *Proceedings of the ACM Conference on Health, Inference, and Learning*, pages 151–159. ACM, 2020. doi: 10.1145/3368555.3384468.

[17] 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): 1–21, 2012. doi: 10.1145/2382577.2382579. Article 15.

[18] Sayash Kapoor and Arvind Narayanan. Leakage and the reproducibility crisis in machine-learning-based science. *Patterns*, 4(9):100804, 2023. doi: 10.1016/j.patter.2023.100804.

[19] D. Sculley, Gary D. Holt, Daniel Golovin, Eugene V. Davydov, Todd Phillips, Dietmar Ebner, et al. Hidden technical debt in machine learning systems. In *Advances in Neural Information Processing Systems 28*, 2015.

[20] U.S. Food and Drug Administration and Health Canada and Medicines and Healthcare products Regulatory Agency. Good machine learning practice for medical device development: guiding principles. https://www.fda.gov/media/153486/download, 2021. Accessed 2026-10-19.

[21] Jenna Wiens, Suchi Saria, Mark Sendak, Marzyeh Ghassemi, Vincent X. Liu, Finale Doshi-Velez, et al. Do no harm: a roadmap for responsible machine learning for health care. *Nature Medicine*, 25(9): 1337–1340, 2019. doi: 10.1038/s41591-019-0548-6.

[22] Christopher J. Kelly, Alan Karthikesalingam, Mustafa Suleyman, Greg Corrado, and Dominic King. Key challenges for delivering clinical impact with artificial intelligence. *BMC Medicine*, 17(1):195, 2019. doi: 10.1186/s12916-019-1426-2.

[23] Richard Y. Wang and Diane M. Strong. Beyond accuracy: what data quality means to data consumers. *Journal of Management Information Systems*, 12(4):5–33, 1996. doi: 10.1080/07421222.1996.11518099.

[24] Michael G. Kahn, Tiffany J. Callahan, Juliana Barnard, Alan E. Bauck, Jeff Brown, Bruce N. Davidson, Hossein Estiri, Carsten Goerg, Erin Holve, Steven G. Johnson, Siaw-Teng Liaw, Marianne Hamilton-Lopez, Daniella Meeker, Toan C. Ong, Patrick Ryan, Ning Shang, Nicole G. Weiskopf, Chunhua Weng, Meredith N. Zozus, and Lisa Schilling. A harmonized data quality assessment terminology and framework for the secondary use of electronic health record data. *eGEMs (Generating Evidence & Methods to improve patient outcomes)*, 4(1):1244, 2016. doi: 10.13063/2327-9214.1244.

[25] Nicole G. Weiskopf and Chunhua Weng. Methods and dimensions of electronic health record data quality assessment: enabling reuse for clinical research. *Journal of the American Medical Informatics Association*, 20(1):144–151, 2013. doi: 10.1136/amiajnl-2011-000681.

[26] Michael G. Kahn, Marsha A. Raebel, Jason M. Glanz, Karen Riedlinger, and John F. Steiner. A pragmatic framework for single-site and multisite data quality assessment in electronic health record-based clinical research. *Medical Care*, 50(Suppl):S21–S29, 2012. doi: 10.1097/mlr.0b013e318257dd67.

[27] Daniel Schwabe, Katinka Becker, Martin Seyferth, Andreas Klaß, and Tobias Schaeffter. The METRIC-framework for assessing data quality for trustworthy AI in medicine: a systematic review. *npj Digital Medicine*, 7(1):203, 2024. doi: 10.1038/s41746-024-01196-4.

[28] Clair Blacketer, Frank J. Defalco, Patrick B. Ryan, and Peter R. Rijnbeek. Increasing trust in real-world evidence through evaluation of observational data quality. *Journal of the American Medical Informatics Association*, 28(10):2251–2257, 2021. doi: 10.1093/jamia/ocab132.

[29] Katinka Becker, Maximilian P. Oppelt, Tobias S. Zech, Martin Seyferth, Sandie Cabon, Vanja Miskovic, Ivan Cimrak, Michal Kozubek, Giuseppe D'Avenio, Ilaria Campioni, Jana Fehr, Kanjar De, Ismail Mahmoudi, Emilio Dolgener Cantu, Laurenz Ottmann, Andreas Klass, Galaad Altares, Jackie Ma, Alireza Salehi M., Nadine R. Lang-Richter, Tobias Schaeffter, and Daniel Schwabe. Metric hub: A metric library and practical selection workflow for use-case-driven data quality assessment in medical ai. *arXiv preprint arXiv:2601.22702*, 2026.

[30] Martin Seyferth, Katinka Becker, Tobias Schaeffter, Daniel Schwabe, and Matthias Boehm. Metriclib: A modular and extensible toolkit for evaluation of medical ml datasets. In *Proceedings of the 29th International Conference on Extending Database Technology (EDBT)*, 2026. doi: 10.48786/edbt.2026.66. Software available at https://github.com/PTBresearch/MetricLib.

[31] Yu, Yunguo. Clinical assurance and lifecycle governance of healthcare AI, 2026. Registration record, https://osf.io/5h78s.

[32] Yu, Yunguo. From monitoring to action: A scoping review of post-deployment clinical AI surveillance, risk assessment, and lifecycle management, 2026. Registration record, https://osf.io/5tduj.

[33] Yu, Yunguo. Toward an operational science of clinical AI: From model evaluation to the study of operational trajectories, 2026. Preprint.

[34] Joaquin Quiñonero-Candela, Masashi Sugiyama, Anton Schwaighofer, and Neil D. Lawrence, editors. *Dataset Shift in Machine Learning*. The MIT Press, 2009. doi: 10.7551/mitpress/9780262170055.001.0001.

[35] 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*, 45:521–530, 2012. doi: 10.1016/j.patcog.2011.06.019.

[36] David J. Hand. Classifier technology and the illusion of progress. *Statistical Science*, 21(1):1–14, 2006. doi: 10.1214/088342306000000060.

[37] Stef van Buuren. *Flexible Imputation of Missing Data, Second Edition*. Chapman and Hall/CRC, 2 edition, 2018. doi: 10.1201/9780429492259.

[38] Jonathan A. C. Sterne, Ian R. White, John B. Carlin, Matthias Spratt, Patrick Royston, Michael G. Kenward, et al. Multiple imputation for missing data in epidemiological and clinical research: potential and pitfalls. *BMJ*, 338(jun29 1):b2393, 2009. doi: 10.1136/bmj.b2393.

[39] Wang Miao and Eric J. Tchetgen Tchetgen. On varieties of doubly robust estimators under missingness not at random with a shadow variable. *Biometrika*, 103(2):475–482, 2016. doi: 10.1093/biomet/asw016.

[40] Wei Li, Wang Miao, and Eric Tchetgen Tchetgen. Non-parametric inference about mean functionals of non-ignorable non-response data without identifying the joint distribution. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 85(3):913–935, 2023. doi: 10.1093/jrsssb/qkad047.

[41] Benoît Frénay and Michel Verleysen. Classification in the presence of label noise: a survey. *IEEE transactions on neural networks and learning systems*, 25(5):845–869, 2014. doi: 10.1109/TNNLS.2013.2292894.

[42] Hao Cheng, Zhaowei Zhu, Xing Li, Yilun Gong, Xing Sun, and Yang Liu. Learning with instance-dependent label noise: A sample sieve approach. In *International Conference on Learning Representations*, 2021.

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

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

[45] Tom J. Pollard, Alistair E. W. Johnson, Jesse D. Raffa, Leo A. Celi, Roger G. Mark, and Omar Badawi. The eICU collaborative research database, a freely available multi-center database for critical care research. *Scientific Data*, 5(1):180178, 2018. doi: 10.1038/sdata.2018.178.

[46] Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q Weinberger. On calibration of modern neural networks. In *Proceedings of the 34th International Conference on Machine Learning*, pages 1321–1330. PMLR, 2017.

[47] Ben Van Calster, Daan Nieboer, Yvonne Vergouwe, Bavo De Cock, Michael J. Pencina, and Ewout W. Steyerberg. A calibration hierarchy for risk models was defined: from utopia to empirical data. *Journal of Clinical Epidemiology*, 74:167–176, 2016. doi: 10.1016/j.jclinepi.2015.12.005.

[48] Ben Van Calster, David J. McLernon, Maarten van Smeden, Laure Wynants, Ewout W. Steyerberg, and Topic Group 'Evaluating Diagnostic Tests and Prediction Models' of the STRATOS Initiative. Calibration: the Achilles heel of predictive analytics. *BMC Medicine*, 17(1):230, 2019. doi: 10.1186/s12916-019-1466-7.

[49] Yunguo Yu, Cesar A. Gomez-Cabello, Syed Ali Haider, Ariana Genovese, Srinivasagam Prabha, Maissa Trabilsy, Bernardo G. Collaco, Nadia G. Wood, Sanjay Bagaria, Cui Tao, and Antonio J. Forte. Enhancing clinician trust in AI diagnostics: A dynamic framework for confidence calibration and transparency. *Diagnostics*, 15(17):2204, 2025. doi: 10.3390/diagnostics15172204.

[50] Mervyn Singer, Clifford S. Deutschman, Christopher Warren Seymour, Manu Shankar-Hari, Djillali Annane, Michael Bauer, et al. The Third International Consensus Definitions for Sepsis and Septic Shock (Sepsis-3). *JAMA*, 315(8):801–810, 2016. doi: 10.1001/jama.2016.0287.

[51] Kidney Disease: Improving Global Outcomes (KDIGO) Acute Kidney Injury Work Group. KDIGO clinical practice guideline for acute kidney injury. *Kidney International Supplements*, 2(1):1–138, 2012. https://kdigo.org/wp-content/uploads/2016/10/KDIGO-2012-AKI-Guideline-English.pdf.

[52] John R. Zech, Marcus A. Badgeley, Manway Liu, Anthony B. Costa, Joseph J. Titano, and Eric Karl Oermann. Variable generalization performance of a deep learning model to detect pneumonia in chest radiographs: A cross-sectional study. *PLoS Medicine*, 15(11):e1002683, 2018. doi: 10.1371/journal.pmed.1002683.

[53] Gaël Varoquaux and Veronika Cheplygina. Machine learning for medical imaging: methodological failures and recommendations for the future. *npj Digital Medicine*, 5(1):48, 2022. doi: 10.1038/s41746-022-00592-y.

[54] Gary S. Collins, Paula Dhiman, Jie Ma, Michael M. Schlussel, Lucinda Archer, Ben Van Calster, Frank E. Harrell, Glen P. Martin, Karel G. M. Moons, Maarten van Smeden, Matthew Sperrin, Garrett S. Bullock, and Richard D. Riley. Evaluation of clinical prediction models (Part 1): from development to external validation. *BMJ*, 384:e074819, 2024. doi: 10.1136/bmj-2023-074819.

[55] James A. Hanley and Barbara J. McNeil. The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology*, 143(1):29–36, 1982. doi: 10.1148/radiology.143.1.7063747.

[56] Ewout W. Steyerberg. *Clinical Prediction Models: A Practical Approach to Development, Validation, and Updating*. Statistics for Biology and Health. Springer, 2 edition, 2019. doi: 10.1007/978-3-030-16399-0.

[57] Bradley Efron. Bootstrap methods: another look at the jackknife. *The Annals of Statistics*, 7(1):1–26, 1979. doi: 10.1214/aos/1176344552.

[58] Richard D. Riley, Kym I. E. Snell, Lucinda Archer, Joie Ensor, Thomas P. A. Debray, Ben van Calster, Maarten van Smeden, and Gary S. Collins. Evaluation of clinical prediction models (Part 3): calculating the sample size required for an external validation study. *BMJ*, 384:e074821, 2024. doi: 10.1136/bmj-2023-074821.

[59] Gary S Collins, Karel G M Moons, Paula Dhiman, Richard D Riley, Andrew L Beam, Ben Van Calster, Marzyeh Ghassemi, Xiaoxuan Liu, Johannes B Reitsma, Maarten Van Smeden, Anne-Laure Boulesteix, Jennifer Catherine Camaradou, Leo Anthony Celi, Spiros Denaxas, Alastair K Denniston, Ben Glocker, Robert M Golub, Hugh Harvey, Georg Heinze, Michael M Hoffman, André Pascal Kengné, Emily Lam, Naomi Lee, Elizabeth W Loder, Lena Maier-Hein, Bilal A Mateen, Melissa D McCradden, Lauren Oakden-Rayner, Johan Ordish, Richard Parnell, Sherri Rose, Karandeep Singh, Laure Wynants, and Patricia Logullo. TRIPOD+AI statement: updated guidance for reporting clinical prediction models that use regression or machine learning methods. *BMJ*, 385:e078378, 2024. doi: 10.1136/bmj-2023-078378.

Supplementary Files

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

- supplement.pdf