

Supplementary Material for:

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

Contents

1	S1. Cohort Construction and SQL Queries	2
1.1	S1.1 Sepsis-3 Cohort (MIMIC-IV)	2
1.2	S1.2 ICU Mortality Cohort (MIMIC-IV)	2
1.3	S1.3 AKI Cohort (MIMIC-IV)	3
1.4	S1.4 eICU-CRD Mortality Cohort	4
2	S2. Corruption Manifest	5
3	S3. Construct-Validation Panel: Extended Results	6
3.1	S3.1 Full Correlation Tables	6
3.2	S3.2 Negative Controls (Components Should Ignore)	6
3.3	S3.3 Spearman Correlation Summary	6
4	S4. Leave-One-Family-Out Evaluation: Extended Results	6
4.1	S4.1 Per-Severity MAE and Advantage	7
4.2	S4.2 Bootstrap Confidence Intervals per Family	7
5	S5. eICU Calibration Analysis: Extended	7
5.1	S5.1 Site-Level Calibration Metrics	7
5.2	S5.2 Calibration Curves by Site	8
5.3	S5.3 Calibration Intercept and Slope	8
5.4	S5.4 Null Experiments for the δ -ECE Association	8
6	S6. TRIPOD+AI Checklist	8
7	S7. Additional Sensitivity Analyses	10
7.1	S7.1 q_R Duplicate Sensitivity	10
7.2	S7.2 q_R Training Set Size Sensitivity	10
7.3	S7.3 k-Sensitivity for MI Weights	10
7.4	S7.4 Zero MI Handling	10
8	S8. Reproducibility Package Contents	10
8.1	Computational Environment	10

1 S1. Cohort Construction and SQL Queries

1.1 S1.1 Sepsis-3 Cohort (MIMIC-IV)

The sepsis cohort was extracted from the `mimic_derived.sepsis3` table using the first 2,000 adult ICU stays by `stay_id`. The following query defines the cohort:

```
SELECT
  ie.subject_id, ie.hadm_id, ie.stay_id,
  ie.intime, ie.outtime,
  pat.anchor_age AS age,
  pat.gender,
  -- Vital signs (averaged over first 24h)
  vs.heart_rate, vs.resp_rate, vs.mean_bp, vs.temp, vs.spo2,
  -- Labs (averaged over first 24h)
  chem.creatinine, chem.bun, chem.lactate, chem.wbc,
  -- Severity
  sofa.sofa_score,
  -- Outcome
  adm.hospital_expire_flag AS label
FROM mimic_icu.icustays ie
JOIN mimic_core.patients pat ON ie.subject_id = pat.subject_id
JOIN mimic_core.admissions adm ON ie.hadm_id = adm.hadm_id
JOIN mimic_derived.sepsis3 sep ON ie.stay_id = sep.stay_id
LEFT JOIN mimic_derived.vitalsign vs ON vs.stay_id = ie.stay_id
LEFT JOIN mimic_derived.chemistry chem ON chem.stay_id = ie.stay_id
LEFT JOIN mimic_derived.sofa sofa ON sofa.stay_id = ie.stay_id
WHERE sep.suspicion_poe IS NOT NULL
      AND sep.suspicion_poe <= '24 hours'
      AND pat.anchor_age >= 18
ORDER BY ie.stay_id
LIMIT 2000;
```

1.2 S1.2 ICU Mortality Cohort (MIMIC-IV)

The mortality cohort comprises the first 3,000 adult ICU stays with length of stay ≥ 1 day:

```
SELECT
  ie.subject_id, ie.hadm_id, ie.stay_id,
  ie.intime, ie.outtime,
  pat.anchor_age AS age,
  pat.gender,
  pat.race,
  -- Vital signs (admission-time)
  vs.heart_rate, vs.resp_rate, vs.mean_bp, vs.temp, vs.spo2,
  -- Labs (admission-time)
  chem.creatinine, chem.bun, chem.lactate, chem.wbc,
  -- Temporal availability
  chem.charttime AS creatinine_charttime,
```

```

chem.charttime AS bun_charttime,
-- Late labs (for leakage)
chem_late.creatinine AS late_creatinine,
chem_late.charttime AS late_creatinine_charttime,
chem_late.bun AS late_bun,
chem_late.charttime AS late_bun_charttime,
-- Severity
sofa.sofa_score,
-- Outcome
adm.hospital_expire_flag AS label
FROM mimic_icu.icustays ie
JOIN mimic_core.patients pat ON ie.subject_id = pat.subject_id
JOIN mimic_core.admissions adm ON ie.hadm_id = adm.hadm_id
LEFT JOIN mimic_derived.vitalsign vs ON vs.stay_id = ie.stay_id
LEFT JOIN mimic_derived.chemistry chem ON chem.stay_id = ie.stay_id
LEFT JOIN mimic_derived.sofa sofa ON sofa.stay_id = ie.stay_id
LEFT JOIN mimic_derived.chemistry chem_late
  ON chem_late.stay_id = ie.stay_id
  AND chem_late.charttime BETWEEN
    COALESCE(adm.dischtime, adm.deathtime) - INTERVAL '48 hours'
    AND COALESCE(adm.dischtime, adm.deathtime)
WHERE ie.los >= 1
  AND pat.anchor_age >= 18
ORDER BY ie.stay_id
LIMIT 3000;

```

1.3 S1.3 AKI Cohort (MIMIC-IV)

The AKI cohort uses KDIGO staging from serial creatinine:

```

SELECT
  ie.subject_id, ie.hadm_id, ie.stay_id,
  ie.intime,
  pat.anchor_age AS age,
  pat.gender,
  pat.race,
  -- Baseline creatinine (lowest pre-admission or first measured)
  CASE
    WHEN pre_creatinine.min_creat IS NOT NULL THEN pre_creatinine.min_creat
    ELSE first_creat.creatinine
  END AS baseline_creatinine,
  -- Admission creatinine
  adm_creat.creatinine AS adm_creatinine,
  -- Peak creatinine in first 7 days
  peak_creat.max_creat AS peak_creatinine,
  -- KDIGO stage
  CASE
    WHEN peak_creat.max_creat >= 3 * COALESCE(pre_creatinine.min_creat, first_creat.creatinine)

```

```

        WHEN peak_creat.max_creat >= 2 * COALESCE(pre_creatinine.min_creat, first_creat.creatinine)
        WHEN peak_creat.max_creat >= 1.5 * COALESCE(pre_creatinine.min_creat, first_creat.creatinine)
        ELSE 0
    END AS kdigo_stage,
    -- Severe AKI outcome
    CASE WHEN peak_creat.max_creat >= 2 * COALESCE(pre_creatinine.min_creat, first_creat.creatinine)
        THEN 1 ELSE 0 END AS severe_aki
FROM mimic_icu.icustays ie
JOIN mimic_core.patients pat ON ie.subject_id = pat.subject_id
JOIN mimic_core.admissions adm ON ie.hadm_id = adm.hadm_id
LEFT JOIN (
    SELECT hadm_id, MIN(valuenum) AS min_creat
    FROM mimic_hosp.labevents
    WHERE itemid = 50912 AND valuenum IS NOT NULL
        AND charttime < ie.intime
    GROUP BY hadm_id
) pre_creatinine ON pre_creatinine.hadm_id = ie.hadm_id
LEFT JOIN (
    SELECT hadm_id, valuenum AS creatinine
    FROM mimic_hosp.labevents
    WHERE itemid = 50912 AND valuenum IS NOT NULL
        AND charttime BETWEEN ie.intime AND ie.intime + INTERVAL '24 hours'
    ORDER BY charttime
    LIMIT 1
) first_creat ON first_creat.hadm_id = ie.hadm_id
LEFT JOIN (
    SELECT hadm_id, MAX(valuenum) AS max_creat
    FROM mimic_hosp.labevents
    WHERE itemid = 50912 AND valuenum IS NOT NULL
        AND charttime BETWEEN ie.intime AND ie.intime + INTERVAL '7 days'
    GROUP BY hadm_id
) peak_creat ON peak_creat.hadm_id = ie.hadm_id
WHERE pat.anchor_age >= 18
    AND peak_creat.max_creat >= 1.5 * COALESCE(pre_creatinine.min_creat, first_creat.creatinine)

```

*Note: the query above avoids referencing SELECT-list aliases (*kdigo_stage*, *baseline*) in other expressions or in the WHERE clause; correlated column references (*ie.intime*) inside the joined subqueries evaluate in the outer scope of each subquery. Staged CTE implementations of these cohort definitions are in the reproducibility package (SQL/).*

1.4 S1.4 eICU-CRD Mortality Cohort

```

SELECT
    pt.patientunitstayid AS stay_id,
    pt.uniquepid AS subject_id,
    pt.age, pt.gender,
    -- Vitals
    vp.heartrate, vp.respiration, vp.sao2,

```

```

vp.systemicmean AS mean_bp,
-- Labs
lb.creatinine, lb.bun, lb.wbc,
-- Outcome
pt.unitdischargestatus = 'Expired' AS label
FROM patient pt
LEFT JOIN vitalperiodic vp ON vp.patientunitstayid = pt.patientunitstayid
AND vp.observationoffset BETWEEN 0 AND 1440
LEFT JOIN lab lb ON lb.patientunitstayid = pt.patientunitstayid
AND lb.labresultoffset BETWEEN -1440 AND 1440
WHERE pt.age >= 18
AND pt.unitdischargestatus IN ('Alive', 'Expired')
ORDER BY pt.patientunitstayid;

```

2 S2. Corruption Manifest

The following table documents the exact corruption equations, parameters, and randomization units for all controlled corruptions used in the study.

Table 1: Corruption manifest: equations, parameters, and randomization units for all controlled corruptions.

Corruption	Equation	Parameters	Randomization Unit
Outcome-dependent missingness (Sepsis)	For each of the four discretionary labs j (lactate, creatinine, BUN, WBC): $M_{ij} \sim \text{Bernoulli}(p_{ij})$ where $p_{ij} = 0.60$ if $y_i = 1$, else 0.356	Nominal 40% additional missingness; applied to currently observed values only	Per-record, per-lab
Institutional shift (Sepsis)	$x_{ij} \leftarrow x_{ij} + 1.5 \cdot \hat{\sigma}_j$ where $\hat{\sigma}_j$ is feature std in training	Magnitude 1.5 SD per feature; all features shifted	All features simultaneously
Joint (Sepsis)	Missingness then shift applied sequentially	As above	As above
Asymmetric label noise (Mortality)	$\tilde{y}_i = y_i$ with prob $1 - \eta_1$ if $y = 1$ else $1 - \eta_0$; $\tilde{y}_i = 1 - y_i$ with prob η_1 if $y = 1$ else η_0	$\eta_1 = 0.25, \eta_0 = 0.10$ (asymmetric)	Per-record
Temporal leakage (Mortality)	$x_{i,\text{creat}} \leftarrow \text{late_creat}_i$ where available; $x_{i,\text{BUN}} \leftarrow \text{late_BUN}_i$ where available	Leakage rate 0.35; applied only where late values available	Per-record

Supplementary Table S2 (continued): Corruption Manifest

Corruption	Equation	Parameters	Randomization Unit
Coverage restriction (Mortality)	Subsample to fraction f of training data by an age threshold: keep records with an age below the $(1-f)$ -quantile of age	$f \in \{0.25, 0.50, 0.75\}$	Per-record (age-based)
Temporal leakage (Sepsis)	Same as mortality temporal leakage	Leakage rate 0.15, 0.35, 0.55	Per-record

3 S3. Construct-Validation Panel: Extended Results

3.1 S3.1 Full Correlation Tables

Table 2: Component response to missingness corruption (positive control for q_C).

Missingness Rate	q_C	q_N	q_B	q_R	q_T
0.0	1.0000	0.8984	0.9955	0.8993	1.0000
0.1	0.9013	0.8912	0.9955	0.8843	1.0000
0.2	0.7955	0.8907	0.9951	0.8631	1.0000
0.3	0.7047	0.8831	0.9953	0.8342	1.0000
0.4	0.5990	0.8895	0.9942	0.7902	1.0000

Table 3: Component response to label noise (positive control for q_N).

Noise Rate	q_C	q_N	q_B	q_R	q_T
0.0	1.0000	0.8984	0.9955	0.8993	1.0000
0.1	1.0000	0.8341	0.9955	0.8993	1.0000
0.2	1.0000	0.7685	0.9955	0.8993	1.0000
0.3	1.0000	0.7060	0.9955	0.8993	1.0000
0.4	1.0000	0.6602	0.9955	0.8993	1.0000

3.2 S3.2 Negative Controls (Components Should Ignore)

3.3 S3.3 Spearman Correlation Summary

4 S4. Leave-One-Family-Out Evaluation: Extended Results

Specification note. The tables in this section report the *initial* leave-one-family-out specification, in which patient–setting observations were pooled across families when calibrating the ridge

Table 4: Component response to distribution shift (positive control for q_B , q_R).

Shift Magnitude (SD)	q_C	q_N	q_B	q_R	q_T
0.0	1.0000	0.8984	0.9955	0.8993	1.0000
0.5	1.0000	0.8993	0.8415	0.8804	1.0000
1.0	1.0000	0.9001	0.6360	0.6840	1.0000
1.5	1.0000	0.8899	0.4507	0.2908	1.0000
2.0	1.0000	0.8971	0.3500	0.0570	1.0000

Table 5: Negative controls: q_C should ignore label noise and shift.

Corruption	Severity	q_C	q_N	q_B	q_R
Label noise	0.1	1.0000	0.8230	0.9955	0.8993
Label noise	0.2	1.0000	0.7587	0.9955	0.8993
Label noise	0.3	1.0000	0.7145	0.9955	0.8993
Label noise	0.4	1.0000	0.6567	0.9955	0.8993
Shift	0.5	1.0000	0.8993	0.8415	0.8804
Shift	1.0	1.0000	0.9001	0.6360	0.6840
Shift	1.5	1.0000	0.8899	0.4507	0.2908
Shift	2.0	1.0000	0.8971	0.3500	0.0570

predictor. The primary results are those reported in the main manuscript, which use the revised patient-grouped specification with family-grouped inner validation, which removes patient-level pseudo-replication; its overall mean advantage versus the best scalar baseline is +0.009 (95% CI $[-0.035, +0.051]$) and versus the always-zero comparator +0.005 (CI $[-0.020, +0.042]$). Per-severity results are shown here for the initial run because the revised specification was executed at the family level; the two specifications are not directly comparable.

4.1 S4.1 Per-Severity MAE and Advantage

4.2 S4.2 Bootstrap Confidence Intervals per Family

5 S5. eICU Calibration Analysis: Extended

5.1 S5.1 Site-Level Calibration Metrics

Table 12: Site-level eICU estimates for all 21 eligible hospitals (≥ 50 stays, ≥ 10 deaths).

Hospital	Stays	Deaths	Prevalence	δ	q_B (MDS)	q_R (TSC)	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

Supplementary Table S5 (continued): Site-level eICU estimates.

Hospital	Stays	Deaths	Prevalence	δ	q_B	q_R	AUROC	ECE
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

5.2 S5.2 Calibration Curves by Site

Calibration curves for top/bottom 5 sites by ECE are available in the reproducibility package outputs; the figure-generation code is included in the package’s validation scripts.

5.3 S5.3 Calibration Intercept and Slope

5.4 S5.4 Null Experiments for the δ –ECE Association

Reading: under panel (a) the observed association ($\rho = 0.596$) lies at the 94th percentile of estimation-artifact-only nulls, i.e. only marginally unusual under the null ($P = 0.057$); under panel (b) attenuation of the association follows from equalizing site sample sizes. A second null endpoint (δ vs. $1 - \text{AUROC}$, observed $\rho = 0.141$) has a fraction of null replicates at or above the observed value of 0.297: the absence of a detectable discrimination association is not informative about the estimation of small correlations. All inputs are the site-level values reported in Tables 12 and 13.

6 S6. TRIPOD+AI Checklist

Table 15: TRIPOD+AI checklist for this study.

Item	Section	Description
Title	Abstract	Identifies study as a multivariable prediction model study with AI
Abstract	Abstract	Background, methods, results, conclusions
Background	Introduction	Rationale and objectives
Data sources	Design	MIMIC-IV, eICU-CRD, eligibility criteria, dates

Supplementary Table S6 (continued): TRIPOD+AI Checklist

Item	Section	Description
Participants	Design	Inclusion/exclusion (≥ 50 stays and ≥ 10 deaths per strike; stay caps per Section 4.2)
Outcome	Design	In-hospital mortality, severe AKI, time horizon
Predictors	Design	Admission-time features (≤ 12 for mortality/sepsis; MKD for AKI), measurement windows per cohort, complete-case construction, no imputation or missingness indicators
Sample size	Design	Per-cohort extraction caps; descriptive events-per-variable check at analysis time (Appendix B of the manuscript); no a priori power calculation
Missing data	Design	Handling, completeness weights, MCAR/MAR/MNAR
Statistical analysis	Methods	L2 logistic regression (sklearn defaults), DQI profile computation, paired bootstrap and two-sided permutation tests
Risk groups	–	Not prespecified (continuous risk scores; ten equal-width bins used for ECE)
Development vs validation	Design	Fixed stay-level random split with outcome stratification (sepsis 1400/600; mortality 2100/900; AKI per cohort), external eICU validation
Model specification	Methods	Logistic regression, max_iter=1000, solver=lbfgs
Model performance	Results	Discrimination (AUROC), calibration (ECE, intercept, slope)
Model evaluation	Results	Discrimination (AUROC), calibration (ECE, intercept, slope)
Model updating	–	Not performed
Discussion	Discussion	Interpretation, limitations, generalizability
Other information	Declarations	Funding, COI, data/code availability, ethics
AI-specific items		
Data preprocessing	Methods	Feature engineering, imputation, standardization
Algorithm	Methods	Logistic regression (primary), GBT/RF (sensitivity)
Training	Methods	70/30 split, fixed seeds, max_iter=1000
Hyperparameters	Methods	Ridge grid {0.01,0.1,1,10,100}, cross-validation
Input features	Methods	Clinical variables, no derived features
Model inputs	Methods	Standardized admission-time features; no missingness indicators
Model outputs	Results	Probabilities, clipped to $[10^{-6}, 1 - 10^{-6}]$
Calibration	Methods/Results	ECE (10 bins), intercept, slope, calibration curves

Supplementary Table S6 (continued): TRIPOD+AI Checklist

Item	Section	Description
Validation	Methods/Results	Internal (fixed stay-level split), external (eICU), patient-grouped LOFO
Reporting	Results	All metrics with intervals (percentile/bootstrap)
Reproducibility	Supplement	Code, data extraction, config, environment lock

7 S7. Additional Sensitivity Analyses

7.1 S7.1 q_R Duplicate Sensitivity

7.2 S7.2 q_R Training Set Size Sensitivity

7.3 S7.3 k-Sensitivity for MI Weights

7.4 S7.4 Zero MI Handling

8 S8. Reproducibility Package Contents

The reproducibility package is available from the corresponding author on reasonable request (private repository; reviewer access provided on request). It contains:

- **SQL/** — Cohort extraction queries (Sec S1)
- **config/** — `audit-config.json`, `Makefile`, `.env.example`
- **requirements/** — `requirements.txt`, `requirements_lock.txt`, `environment.yml`
- **scripts/** — All preprocessing, corruption, validation scripts (30+ scripts)
- **src/** — Core DQI engine (`aggregation.py`, `components.py`, etc.)
- **outputs/** — Permissible aggregate run-level outputs (CSV files)
- **docs/** — Manuscript, concept codes, this supplement

8.1 Computational Environment

Python 3.12

NumPy 2.0.2, pandas 3.0.2, scikit-learn 1.8.0, SciPy 1.17.1

matplotlib 3.9.2, seaborn 0.13.2

Jupyter 4.0 (exploratory notebooks only)

Table 6: Negative controls: q_N should ignore missingness and shift.

Corruption	Severity	q_C	q_N	q_B	q_R
Missingness	0.1	0.8992	0.8924	0.9954	0.8857
Missingness	0.2	0.7999	0.8903	0.9949	0.8651
Missingness	0.3	0.6988	0.8878	0.9948	0.8343
Missingness	0.4	0.5977	0.8895	0.9941	0.7930
Shift	0.5	1.0000	0.8993	0.8415	0.8804
Shift	1.0	1.0000	0.9001	0.6360	0.6840
Shift	1.5	1.0000	0.8899	0.4507	0.2908
Shift	2.0	1.0000	0.8971	0.3500	0.0570

Table 7: Negative controls: q_B , q_R should ignore missingness and label noise.

Corruption	Severity	q_C	q_N	q_B	q_R
Missingness	0.1	0.8991	0.8963	0.9952	0.8867
Missingness	0.2	0.8020	0.8895	0.9951	0.8663
Missingness	0.3	0.7001	0.8857	0.9950	0.8375
Missingness	0.4	0.6036	0.8861	0.9948	0.7909
Label noise	0.1	1.0000	0.8274	0.9955	0.8993
Label noise	0.2	1.0000	0.7656	0.9955	0.8993
Label noise	0.3	1.0000	0.7003	0.9955	0.8993
Label noise	0.4	1.0000	0.6524	0.9955	0.8993

Table 8: Spearman correlations between severity and component values (positive controls). The (q_B , q_R) shift correlations use that pair’s shift severity scale.

Component	Missingness	Label Noise	Shift	Shift (for q_B , q_R)
q_C	-1.000	nan	nan	—
q_N	-0.753	-0.998	nan	—
q_B	-0.806	nan	-0.995	—
q_R	-0.979	nan	-0.962	—
q_T	0.000	0.000	0.000	—

Table 9: Mean component values across severity levels (negative controls: components that this corruption does not target should remain high). The q_B/q_R entries under Shift come from that pair’s positive-control runs (shift severity scale 0.5–2.0). All values are means computed from `construct_validation_panel.csv`; the equal-size and duplicate q_R settings are from `distributional_coverage_spec.csv`.

Component	Missingness (for q_N , q_B , q_R)	Label Noise (for q_C , q_B , q_R)	Shift (for q_C , q_N)
q_C	—	1.0000	1.0000
q_N	0.8900	—	0.8966
q_B	0.9950	0.9955	0.5695
q_R	0.8451	0.8993	0.4781

Table 10: Per-severity MAE for ridge predictor and best baseline B^* by held-out family (initial ungrouped specification; traceability only).

Family	Severity	MAE(Ridge)	MAE(B^*)	G	B^*
Missingness	0.15	0.072	0.091	+0.019	ϕ_B
Missingness	0.30	0.082	0.101	+0.019	ϕ_B
Missingness	0.45	0.101	0.108	+0.007	ϕ_B
Label Noise	0.10	0.038	0.036	-0.002	Arithmetic
Label Noise	0.25	0.042	0.037	-0.005	Arithmetic
Label Noise	0.40	0.038	0.035	-0.003	Arithmetic
Shift	0.5	0.071	0.145	+0.074	ϕ_R
Shift	1.0	0.098	0.188	+0.090	ϕ_R
Shift	1.5	0.124	0.252	+0.128	ϕ_R
Coverage	0.25	0.021	0.009	-0.012	Arithmetic
Coverage	0.50	0.023	0.010	-0.013	Arithmetic
Coverage	0.75	0.024	0.010	-0.014	Arithmetic
Leakage	0.15	0.021	0.073	+0.052	Arithmetic
Leakage	0.35	0.021	0.078	+0.057	Arithmetic
Leakage	0.55	0.024	0.083	+0.059	Arithmetic

Table 11: Bootstrap 95% CI for G per held-out family (10,000 resamples; initial ungrouped specification, traceability only).

Held-out Family	G	95% Bootstrap CI
Missingness	+0.014	[-0.012, +0.041]
Label Noise	-0.002	[-0.018, +0.014]
Shift	+0.097	[+0.051, +0.143]
Coverage	-0.013	[-0.024, -0.002]
Leakage	+0.056	[+0.021, +0.091]
Overall (mean)	+0.030	[-0.003, +0.069]

Table 13: Unshrunk calibration intercept and slope per site (21 eligible hospitals; equal-width ECE; slope p -values from a two-sided z test against slope = 1).

Hospital	Stays	Intercept	Slope	ECE	Slope p -value
73	297	-0.28	0.68	0.060	0.001
264	235	+0.38	0.91	0.039	0.001
167	234	-0.24	1.30	0.023	0.001
243	196	-0.48	1.24	0.043	0.001
420	187	+0.34	0.68	0.050	0.003
199	166	+0.31	0.66	0.052	0.008
338	161	-0.71	0.66	0.051	0.018
458	155	-0.37	0.27	0.077	0.121
208	150	+0.59	0.36	0.075	0.090
411	149	-0.16	1.16	0.039	0.001
252	148	+0.28	1.71	0.064	0.001
188	134	-0.14	0.57	0.037	0.062
300	134	-0.06	0.53	0.021	0.071
443	133	+0.42	1.37	0.047	0.001
122	119	-0.17	0.49	0.036	0.068
157	97	+0.03	0.58	0.085	0.056
417	84	-0.27	1.01	0.087	0.001
110	83	+0.44	0.92	0.056	0.054
307	80	-0.32	0.42	0.094	0.232
142	78	+0.29	1.10	0.091	0.001
226	56	+0.83	1.09	0.124	0.007

Table 14: Null-experiment controls for the δ -ECE association (Section 5.4 of the manuscript). (a) Estimation-artifact null: hospital identity resampled from the pooled eICU target without MIMIC-IV cohort signal, 300 replicates. (b) Equal-size control: per-site sample fixed at the smallest eligible site density ($n = 56$ stays), 100 replicates.

<i>(a) Estimation-artifact null (300 replicates)</i>	
Observed $\rho(\delta, \text{ECE})$	0.596
Null median	0.210
Null 95% range	$[-0.220, 0.652]$
Fraction of null replicates $\geq$ observed (P)	0.057
<i>(b) Equal-size site control ($n = 56$ stays; 100 replicates)</i>	
Mean $\rho(\delta, \text{ECE})$	0.168
95% range	$[-0.216, 0.522]$
Median ρ	0.155
Fraction of replicates with $\rho > 0$	0.83

Table 16: q_R sensitivity to duplicate records added to training data (source: `distributional_coverage_spec.csv`; q_R base setting: MIMIC-IV mortality train covers target).

Duplicate Fraction	q_R	Difference from 0%
0%	0.8746	0.0000
10%	0.8713	-0.0033
20%	0.8688	-0.0058

Table 17: q_R sensitivity to training-set size (50%, 75%, 100% of the training pool; source: `distributional_coverage_spec.csv`).

Training Fraction	q_R	Difference from 100%
50%	0.2527	-0.6220
75%	0.7674	-0.1073
100%	0.8746	0.0000

Table 18: Sensitivity of MI weights to the k-NN parameter k (mean absolute difference vs. the default $k = 5$; source: `completeness_weight_validation.csv`).

k	Mean $ w_k - w_{k=5} $
3	0.0065
5	0.0000 (reference)
7	0.0059
10	0.0074
15	0.0057

Table 19: Weight assigned to constant (zero-MI) feature.

Feature	Weight
Constant feature	0.000000