Auditing Measurement Processes in Clinical Registries: Detecting Label–Environment Confounding Before Model Development

Apurv Shukla APURV3894@GMAIL.COM
University of Michigan, Dearborn

Devin L. McCaslin
University of Michigan Medicine

Pravansu Mohanty
University of Michigan, Dearborn

Abstract

Clinical prediction models trained on registry data routinely achieve high in-distribution accuracy, yet their transportability across institutions, clinicians, and time periods remains poorly understood. We formalize this problem through a *measurement process model* in which the observed label Y^{obs} depends on both a true clinical state Y^* and a reporting indicator $R(e,t)$ that varies with environment e and era t: when reporting is inactive, the true label is unobserved and may be imputed by registry convention, introducing structured bias. We propose a *measurement audit*—a battery of three permutation-calibrated statistical tests that detects label–environment entanglement *before* any predictive model is built—and an *evaluation-based attribution* framework that estimates how much of a model's apparent discrimination is due to environment confounding versus genuine clinical signal. Applied to a cochlear implant registry of 3,584 patients, the audit identifies severe, degenerate confounding (Cramér's $V = 0.39$ for label–era association, 4 of 8 eras with zero negative labels, permutation $p < 0.002$). The attribution reveals that approximately 10% of the IID AUROC (0.978, 95% CI $[0.955, 0.991]$) traces to era confounding; the era-blocked AUROC of 0.875 ($[0.668, 0.925]$) better estimates transportable clinical discrimination. External validation on the MIMIC-IV ICU mortality cohort ($n = 67{,}224$; 7 care-unit environments; 5 eras) shows that the audit distinguishes qualitatively different confounding regimes: it detects diffuse, unit-driven confounding ($V = 0.145$, permutation $p < 0.002$; confounding share 3.4% of IID AUROC) while correctly passing era-level tests on a registry with no zero-negative strata. A semi-synthetic calibration study confirms favorable operating properties (power rising monotonically with confounding strength; 4% false-positive rate; zero empirical family-wise error under label permutation). We discuss concrete remediation strategies for registries that fail the audit, and propose environment-aware evaluation as a candidate pre-modeling diagnostic for registry-trained clinical ML.

1. Introduction

Clinical registries are a primary data source for developing prediction models in healthcare. Cochlear implantation, the standard of care for severe-to-profound sensorineural hearing loss, is a representative example: candidacy decisions remain heterogeneous across clinicians

and institutions (Gifford et al., 2010; Sorkin, 2013), and registries capturing demographics, audiological data, and outcomes are natural substrates for ML (Crowson et al., 2020).

However, registries are not random samples from a fixed data-generating process: they are products of *measurement processes* that vary across clinicians, institutions, and eras. Which patients are entered, which fields are completed, and which outcomes are recorded all depend on who is documenting, when, and under what protocols—a fundamental mismatch with the IID assumption underlying standard cross-validation. Despite growing awareness of dataset shift in clinical ML (Subbaswamy and Saria, 2020; Finlayson et al., 2021; Nestor et al., 2019), there is no standard procedure for auditing whether a registry's labels are entangled with its measurement process *before* building prediction models; researchers typically discover the problem post-hoc, when temporal validation reveals unexpected performance drops—by which point substantial effort has already been invested.

In this work, we make three contributions:

1. **A measurement process model** (Section 3) that formalizes how documentation practices generate observed labels, distinguishing between unobserved outcomes and registry-imputed defaults, and identifies conditions under which IID evaluation can overstate clinical utility.

2. **A measurement audit algorithm** (Section 4) comprising three complementary permutation-calibrated statistical tests that detect label–environment confounding before any predictive model is trained, producing an automated diagnostic flag whose aggregation rule is grounded in Fisher's combined-probability test.

3. **An evaluation-based attribution framework** (Section 5) that estimates how much of IID AUROC is attributable to era confounding, clinician confounding, and genuine clinical discrimination.

The individual statistics composing the audit are deliberately classical. The contribution lies in problem formulation and packaging: (a) identifying the measurement process as a distinct and testable source of label corruption, (b) composing complementary tests into a calibrated battery with a transparent, statistically grounded decision rule, and (c) connecting the audit verdict to an actionable evaluation hierarchy.

Beyond proposing an audit battery, we characterize its operating properties under a semi-synthetic measurement-process model, showing that detection power increases monotonically with confounding strength while maintaining a low false-positive rate (Appendix I). We validate these tools on a CI registry of 3,584 patients, where a candidacy model achieves AUROC = 0.978 under IID evaluation but the audit correctly identifies that approximately 10% of this performance is attributable to documentation artifacts (Section 7); and on the publicly available MIMIC-IV ICU mortality cohort ($n = 67{,}224$), where the audit detects a qualitatively different regime—diffuse, unit-driven confounding with a much smaller share of IID AUROC (3.4%)—demonstrating severity-graded, regime-aware assessment rather than a binary alarm (Section 8). We also discuss concrete remediation strategies for registries that fail the audit (Section 9).

Generalizable Insights about Machine Learning in the Context of Healthcare

- Clinical registries are products of measurement processes that vary across clinicians, institutions, and eras; treating them as IID samples can produce models whose apparent accuracy is substantially attributable to documentation artifacts rather than clinical signal.
- Environment–label confounding can and should be audited *before* model development with a small battery of cheap statistical tests, avoiding wasted modeling effort on registries whose labels are entangled with their reporting process.
- Evaluation-based attribution—decomposing IID performance into environment-confounded and environment-blocked components—gives a more honest estimate of transportable clinical discrimination and is a strong candidate for inclusion in reporting standards.

2. Related Work

Our work intersects several literatures; this section summarizes the positioning, and Appendix A gives an extended discussion of each thread. *Dataset shift and temporal drift*: distribution shift is well documented in medical AI (Subbaswamy and Saria, 2020; Finlayson et al., 2021; Nestor et al., 2019; Wong et al., 2021; Moons et al., 2014), and models degrade over time even without detectable covariate drift (Davis et al., 2017; Sáez et al., 2020; Vela et al., 2022); we differ by tracing confounding to the *label-generating process itself* and providing a pre-modeling diagnostic rather than a post-hoc finding, with a quantitative comparison against generic two-sample shift-detection baselines (Rabanser et al., 2019; Gretton et al., 2012; Lopez-Paz and Oquab, 2017) in Section 8. *Missing data, selection, and measurement error*: our reporting indicator induces a missing-not-at-random pattern in Rubin's taxonomy (Rubin, 1976; Little and Rubin, 2019; Seaman and White, 2013), a discrete analogue of Heckman selection (Heckman, 1979), with EHR data-quality taxonomies (Carroll et al., 2006; Hernán and Robins, 2020; Goldstein et al., 2017; Weiskopf et al., 2013; Hripcsak and Albers, 2013) providing the vocabulary our reporting function operationalizes; the crucial distinction is that missingness here is not a nuisance to impute over—combined with a default label d, it *inflates* apparent discrimination. *Label quality*: label-noise methods (Oakden-Rayner et al., 2020; Northcutt et al., 2021a,b; Karimi et al., 2020; Menon et al., 2015) address labels that are noisy; ours are *systematically absent* in specific environment–era strata. *Invariance and transportability*: IRM, DRO, and domain-generalization benchmarks (Arjovsky et al., 2019; Sagawa et al., 2020; Gulrajani and Lopez-Paz, 2021; Koh et al., 2021; Creager et al., 2021) assume trustworthy labels, and causal transportability (Pearl and Bareinboim, 2011; Bareinboim and Pearl, 2016; Subbaswamy et al., 2019) assumes a known graph; our audit is a diagnostic prerequisite that empirically detects when the label itself depends on environment. *Shortcut learning, fairness, and deployment failures*: post-hoc analyses (Zech et al., 2018; DeGrave et al., 2021; Wynants et al., 2020; Roberts et al., 2021; Obermeyer et al., 2019; Seyyed-Kalantari et al., 2021; Rajkomar et al., 2018; Suresh and Guttag, 2021) document exactly the failure modes we formalize; the audit surfaces them *before* modeling effort is invested. *Registries and reporting standards*: prediction-model methodology (Crowson et al., 2020; Steyerberg, 2019;

Riley et al., 2020) and TRIPOD/TRIPOD+AI and PROBAST (Collins et al., 2015, 2024; Wolff et al., 2019) standardize validation and reporting, but none currently require auditing the measurement process that generates labels; our audit could serve as a concrete addition.

3. Measurement Process Model

3.1. Generative Framework

We model the data-generating process for a clinical registry as follows. For patient i evaluated in environment e_i (clinician, institution) at time t_i (era):

$$X_i \sim P_X(X \mid e_i, t_i) \tag{1}$$
$$Y_i^* \sim P_{Y^*}(Y^* \mid X_i) \tag{2}$$
$$R_i \sim \text{Bernoulli}\big(\rho(e_i, t_i)\big) \tag{3}$$

Here X_i are clinical features, Y_i^* is the true clinical status (1 = implant-appropriate, 0 = not), and R_i is a *reporting indicator* that determines whether the true label is observed. The reporting probability $\rho(e,t) = \mathbb{P}(R = 1 \mid e, t)$ captures the likelihood that the documentation process in environment e at era t records the patient's true status.

The observed label in the registry is then:

$$Y_i^{\text{obs}} = \begin{cases} Y_i^* & \text{if } R_i = 1 \\ \text{missing} & \text{if } R_i = 0 \end{cases} \tag{4}$$

Crucially, the *operational label* used in practice depends on how the registry handles missing outcomes. We define a registry-specific imputation function g:

$$Y_i = g(Y_i^{\text{obs}}) = \begin{cases} Y_i^{\text{obs}} & \text{if } R_i = 1 \\ d & \text{if } R_i = 0 \end{cases} \tag{5}$$

where d is the default label assigned to unreported outcomes. In many procedural registries, patients who are entered but whose negative disposition is never documented default to the positive class ($d = 1$), because the procedure of interest (implantation, surgery, treatment initiation) is the expected pathway and its non-occurrence must be *actively* recorded. Other registries may set $d = \text{NA}$ (exclusion) or $d = 0$ (conservative default). The choice of g determines the nature and direction of the resulting bias. Note the deterministic collapse hidden in Eq. 5: in any stratum where reporting has ceased ($\rho(e,t) = 0$) and $d = 1$, *every* patient receives $Y_i = 1$ by construction. This deterministic collapse is the structural source of the bias studied in this paper.

This formulation—features and true label drawn from a joint process (Eqs. 1–2), passed through environment-dependent reporting (Eq. 3), outcome observation (Eq. 4), and registry imputation (Eq. 5)—generalizes across registries; the specific pathology depends on whether $\rho(e,t)$ varies across environments, on the choice of d, and on the correlation between X and (e,t).

3.2. Conditions for Evaluation Bias

We now state the conditions under which IID evaluation on Y can mislead. Throughout, *discrimination* refers to AUROC, which we define formally to fix notation.

Definition 1 (AUROC) *For a score function f and binary label Z, with (X_i, Z_i) and (X_j, Z_j) independent draws,*

$$AUROC(f, Z) = \mathbb{P}\big(f(X_i) > f(X_j) \mid Z_i = 1, Z_j = 0\big) + \tfrac{1}{2}\,\mathbb{P}\big(f(X_i) = f(X_j) \mid Z_i = 1, Z_j = 0\big). \tag{6}$$

IID AUROC *denotes the estimate of $AUROC(f, Y)$ obtained by stratified cross-validation that ignores environment structure.*

Proposition 2 (Existence of AUROC inflation) *Let $f : \mathcal{X} \to [0, 1]$ be a predictor evaluated on operational labels Y with default $d = 1$. Suppose (i) $\rho(e, t) < 1$ for some (e, t), (ii) X is not independent of (e, t), and (iii) f captures information about (e, t) through X. Then there* exist *measurement processes and predictors satisfying (i)–(iii) for which*

$$AUROC(f, Y) > AUROC(f, Y^*). \tag{7}$$

Conditions (i)–(iii) are necessary for such inflation but not sufficient: the sign of $AUROC(f, Y) - AUROC(f, Y^*)$ depends on the score distribution of the unreported true negatives, which (i)–(iii) do not constrain. A complete proof appears in Appendix G.

Remark 3 (Non-universality) *The inequality is an existence claim, not a universal one. Consider $X = E \in \{0, 1\}$, $f(X) = E$, $Y^* = E$, $\rho(E{=}0) = 0.5$, $d = 1$: conditions (i)–(iii) hold, yet $AUROC(f, Y) < AUROC(f, Y^*) = 1$, because under $d = 1$ the unreported true negatives move into the observed positive class. This is consistent with the AUC-invariance results of Menon et al. (2015) under class-conditional corruption, and with the deflation observed in our semi-synthetic study under stochastic reporting (Appendix I); if $d = 0$, the bias direction reverses analogously. Corollary 5 (Appendix G) characterizes the bias direction exactly via a mixture decomposition, separating the deterministic-collapse regime of our case study (inflation) from the class-conditional-corruption regime (deflation).*

The key practical insight survives the weakened statement: any environment-dependent reporting combined with feature–environment correlation introduces a confound that IID evaluation cannot detect, and whose direction is governed by the mixture decomposition of Corollary 5.

Corollary 4 (Temporal split diagnostic) *Under a temporal split at time t_0, if $\rho(e, t) = 0$ for all $t > t_0$ and $d = 1$, then $Y_i = 1$ for all patients in the test set and AUROC is undefined. More generally, if ρ is much smaller post-t_0 than pre-t_0, temporal AUROC will collapse even when true clinical discrimination $AUROC(f, Y^*)$ is stable.*

3.3. Estimating the Reporting Function

We estimate $\hat{\rho}(e, t)$ nonparametrically as the observed rate of negative labels within each environment–era stratum:

$$\hat{\rho}(e, t) = \frac{|\{i : Y_i = 0,\, e_i = e,\, t_i \in \mathcal{T}\}|}{|\{i : e_i = e,\, t_i \in \mathcal{T}\}|} \tag{8}$$

where $\mathcal{T}$ is an era bin. Because per-cell estimation is unreliable in small strata, we impose a minimum-count threshold: strata with fewer than 20 patients are pooled into their parent environment before estimation, and we report sensitivity to this threshold. Strata where $\hat{\rho}(e, t) = 0$ indicate inactive reporting—the measurement process has ceased recording negative outcomes, making it impossible to distinguish true positives from unreported negatives. We caution that $\hat{\rho} = 0$ is ambiguous on its own: a stratum can lack negative labels because reporting stopped, because every patient evaluated there was genuinely a candidate, or because the evaluated patient mix shifted. The audit therefore treats zero-negative strata as flags for investigation rather than definitive diagnoses; disambiguation strategies include (a) clinical review of the plausibility of the stratum's patient mix, (b) cross-referencing external prevalence estimates for the outcome, and (c) mask-only analyses of documentation patterns.

4. Measurement Audit Algorithm

We propose a three-test battery that can be applied to any clinical registry *before* building prediction models. The audit requires only environment metadata (era, clinician/institution ID) and outcome labels; no feature engineering or model training is needed for Tests A and C, and Test B requires only simple baseline models.

4.1. Test A: Label-Prevalence Instability (Permutation-Calibrated)

For each environment variable $E \in \{e_{\text{clinician}}, e_{\text{era}}, e_{\text{institution}}\}$, compute the chi-squared statistic of independence between Y and E, along with Cramér's V as an effect-size measure:

$$V = \sqrt{\frac{\chi^2}{n \cdot (\min(r, c) - 1)}} \tag{9}$$

where r, c are the dimensions of the contingency table. Significance is calibrated by permutation rather than chi-squared asymptotics, which are unreliable precisely where measurement-process confounding is most likely—under extreme label imbalance (94.5% positive rate in our case study) (Agresti, 2002). We permute Y 500 times and report the permutation p-value (fraction of permuted statistics exceeding the observed value); with 500 permutations the minimum reportable value is $1/501 \approx 0.002$, so we write $p < 0.002$ rather than smaller nominal values. Earlier versions of this framework reported the asymptotic test and its permutation-calibrated counterpart as two separate tests (A and D); we consolidate them into this single permutation-calibrated Test A, deprecate the asymptotic variant, and report the asymptotic χ^2 statistic only for continuity with standard contingency-table practice.

4.2. Test B: Conditional Predictive Gain

Compare two models: f_X trained on clinical features X alone, and $f_{X,E}$ trained on X plus environment variables E, both via 5-fold cross-validation. Under the null that $Y \perp E \mid X$, adding E should not improve prediction:

$$\Delta\text{AUROC} = \text{AUROC}(f_{X,E}) - \text{AUROC}(f_X) \approx 0 \tag{10}$$

Because ΔAUROC can depend on the model family used for f_X and $f_{X,E}$—flexible learners may absorb environment-correlated signal through features, shrinking the measured gap—we report Test B under both logistic regression and gradient boosting, and recommend the simpler model as the default (Appendix H).

4.3. Test C: Label–Era Monotonicity

Count the fraction of era bins with zero negative labels; if negatives disappear entirely after a specific era, the label is temporally bounded rather than prospectively observed:

$$\text{Score}_C = \frac{|\{\mathcal{T} : \hat{\rho}(\cdot, \mathcal{T}) = 0\}|}{|\text{all era bins}|} \tag{11}$$

4.4. Decision Rule: Fisher-Calibrated Aggregation

Each test uses a *dual criterion*: a permutation-derived p-value *and* a minimum effect-size floor, ensuring that the audit flags only associations that are both statistically significant and practically meaningful.

- **Test A** (era/clinician): FAIL if permutation $p < 0.002$ *and* $V > 0.1$; CAUTION if $p < 0.01$ and $V > 0.1$. The $V > 0.1$ floor corresponds to Cohen's "small" effect-size threshold (Cohen, 1988); labels are permuted 500 times to obtain the null distribution, so $p < 0.002$ ($= 1/501$) is the smallest reportable value.
- **Test B**: FAIL if permutation $p < 0.01$ and $\Delta\text{AUROC} > 0$; CAUTION if $p < 0.05$. Environment columns are shuffled within cross-validation folds (200 permutations) to obtain the null ΔAUROC distribution.
- **Test C**: FAIL if binomial tail $p < 0.001$; CAUTION if $p < 0.01$. The baseline negative rate $\bar{p}$ is estimated from active strata only (those with ≥ 1 negative label), and $P(\geq k$ zero-negative eras) is computed via Poisson approximation.

Overall diagnostic: FAIL if ≥ 2 individual FAILs; CAUTION if ≥ 1 FAIL or ≥ 2 CAUTIONs; PASS otherwise. In this tally, two CAUTIONs are treated as equivalent to one FAIL. A FAIL recommends: "Do not report IID AUROC as an estimate of clinical utility. Report environment-aware evaluations as primary results."

Statistical grounding of the aggregation rule. The equivalence "two CAUTIONs = one FAIL" is Fisher's combined-probability test (Fisher, 1932) calibrated to the FAIL level: under the global null, two independent permutation p-values at the CAUTION threshold ($p_1, p_2 \leq 0.01$) combine via $-2[\ln p_1 + \ln p_2] \sim \chi^2_4$ to a combined $p \approx 0.001$, matching the single-test FAIL threshold. Tests B and C are partially dependent through the shared era partition; Brown's adjustment for correlated p-values (Brown, 1975) would inflate the combined p slightly, making the rule conservative (fewer FAILs), not liberal. We also validate the rule empirically: under label permutation on MIMIC-IV ($\delta = 0$, $N = 100$ replicates), the family-wise error rate of the overall FAIL verdict is 0.000, and varying the (conventional, not optimized) V cutoff from 0.05 to 0.20 leaves the FAIL verdict unchanged on both case-study registries with empirical family-wise error ≤ 0.02 throughout (Appendix I). The thresholds are further validated by a semi-synthetic calibration study (Appendix I): detection power increases monotonically with confounding strength δ while the false-positive rate at $\delta = 0$ remains at 4%.

5. Evaluation-Based Attribution

To estimate *how much* of IID performance is attributable to environment confounding versus genuine clinical discrimination, we propose a sequential evaluation framework—termed *evaluation-based attribution* rather than "decomposition" to emphasize that these are protocol-dependent estimates of environment contribution, not causal components, and are not guaranteed to be additive.

5.1. Evaluation Hierarchy

We evaluate the same model under five regimes of increasing distributional stringency:

1. **IID CV**: standard stratified 5-fold cross-validation.
2. **Era-blocked CV**: each fold holds out one era block (eras with ≥ 2 negative labels are evaluable).
3. **Clinician-blocked CV**: leave-one-clinician-out (clinicians with ≥ 1 negative test-set label are evaluable).
4. **Doubly-controlled**: within each era, leave one clinician out—removing both confounding dimensions at once.
5. **Residualized**: train a clinical model f_X and an environment model g_E via 5-fold CV; the score $s_i = \text{logit}(f_X(x_i)) - \text{logit}(g_E(e_i, t_i))$ subtracts the environment-predictable component.

5.2. Attribution Estimates

Define:

$$\hat{\Delta}_{\text{era}} = \text{AUROC}_{\text{IID}} - \text{AUROC}_{\text{era-blocked}} \tag{12}$$

$$\hat{\Delta}_{\text{clin}} = \text{AUROC}_{\text{IID}} - \text{AUROC}_{\text{clin-blocked}} \tag{13}$$

$$\hat{\Delta}_{\text{env}} = \text{AUROC}_{\text{IID}} - \text{AUROC}_{\text{residualized}} \tag{14}$$

These are estimates of the AUROC attributable to era-specific patterns ($\hat{\Delta}_{\text{era}}$), clinician-specific patterns ($\hat{\Delta}_{\text{clin}}$), and total environment signal ($\hat{\Delta}_{\text{env}}$). The residual $\hat{\Delta}_{\text{int}} = \hat{\Delta}_{\text{env}} - \hat{\Delta}_{\text{era}} - \hat{\Delta}_{\text{clin}}$ captures any interaction, though in practice this quantity may be small or negative due to the non-additivity of evaluation-regime effects. We residualize on the logit scale because logits are the natural additive scale on which independent sources of evidence combine; alternative schemes (partial-correlation adjustment of scores, doubly-robust environment adjustment) are possible and worth comparing in future work. The attribution is *approximate*: $\hat{\Delta}_{\text{era}}$, $\hat{\Delta}_{\text{clin}}$, and $\hat{\Delta}_{\text{env}}$ are not orthogonal contributions and need not be additive. When the environment model is sufficiently expressive, the residualized AUROC may fall below 0.5, indicating that the subtraction over-corrects by removing shared variance that is both clinically meaningful and environment-correlated. Rather than treating this as a defect, we propose the residualized AUROC as a *regime diagnostic*: collapse far below 0.5 signals a degenerate regime in which label structure is entangled with environment at the imputation rule itself, whereas a moderate residualized value indicates diffuse, non-pathological confounding. Our two case studies realize both regimes (residualized AUROC of 0.334 on the CI registry, Section 7, versus 0.753 on MIMIC-IV, Section 8); the recommendation is to report the residualized AUROC alongside—never instead of—the blocked AUROCs.

6. Data: Cochlear Implant Registry

We first validate our framework on a single-center cochlear implant registry containing 3,584 patient records collected over approximately two decades. (Throughout this paper, "CI" abbreviates *cochlear implant* only; confidence intervals are always written "95% CI.") The registry contains 10 clinician identifiers; the 8 with $n \geq 100$ patients define the primary clinician environments used throughout, and the remaining 2 low-volume identifiers are retained only as additional clusters in the bootstrap analysis (Appendix J).

Eligibility criteria. We include every patient evaluated for cochlear implant candidacy at the contributing institution between the first and last year of the registry, as entered by the clinicians with primary responsibility for candidacy evaluations (the 8 primary-environment clinicians defined above). We exclude: (i) records with missing or malformed entry date, preventing era assignment ($n < 20$); (ii) records with no implantation-status field populated at all (a data-entry artifact, excluded as missing—distinct from the implicit $d = 1$ default of Section 7, which concerns patients recorded as implanted because non-implantation was never actively documented); and (iii) revision-surgery encounters. No patients were excluded on the basis of age, demographics, etiology, or comorbidity, so the analytic cohort matches the deployment population; all identifying information was removed before analysis, with clinician identifiers SHA-256 hashed to preserve grouping structure.

Table 1: Reporting function $\hat{\rho}(e,t)$: fraction of patients labeled "No CI" (not implanted) by era and clinician, with per-clinician totals n in the row labels. Dashes denote strata with no patients. The 8 chronological audit era bins are aggregated into 4 columns for display. After 2012, no negative labels are recorded in any environment.

Clinician	**≤2008**	**2009–12**	**2013–16**	**2017+**
clin_F ($n = 997$)	0.088	0.412	0.000	0.000
clin_E ($n = 337$)	0.140	—	—	—
clin_A ($n = 1149$)	0.021	0.127	0.000	0.000
clin_D ($n = 665$)	0.083	0.250	0.000	0.000
All others	0.000	0.000	0.000	0.000

Features. After cleaning, the analytic dataset contains 20 clinical features including patient type (adult/pediatric), age at entry and implantation, hearing loss onset type and etiology, cochlear anatomy, additional disabilities, and communication mode. Continuous features are standardized with median imputation for missing values; categorical features are one-hot encoded with an explicit missing-indicator column; no feature selection is applied before the audit; imputation and scaling parameters are fit on training folds only within each cross-validation split. Patient type and age are strongly collinear (pediatric patients are by definition young at entry), and the pediatric/adult split is plausibly a stronger driver of label patterns than clinician identity; we return to this collinearity when interpreting feature importance (Section 7). Age at implantation is defined only for implanted patients and is therefore not available at the time a candidacy prediction would be made; Section 7 reports an ablation excluding it.

Outcome. We study candidacy prediction: whether a patient received a cochlear implant (implanted = 1, $n = 3{,}386$, 94.5%) versus not implanted (0, $n = 198$, 5.5%). Implantation status reflects both clinical suitability and evolving practice guidelines—candidacy criteria have expanded substantially over two decades, with audiological thresholds shifting and indications broadening (Gifford et al., 2010)—so a patient deemed "not a candidate" in 2005 might meet criteria under 2020 guidelines. The label thus encodes a mixture of clinical state, clinician judgment, and era-specific policy, making it a natural testbed for our framework.

Models. We train XGBoost (200 trees, max depth 5, learning rate 0.05, class-weighted) as the primary model, with logistic regression and random forest as comparisons.

7. Results

7.1. Measurement Process Characterization

Estimating the reporting function $\hat{\rho}(e,t)$ reveals severe measurement heterogeneity (Table 1): all 198 negative labels originate from the pre-2012 era, two clinicians account for 88% of them (clinician F: 64.1%; clinician E: 23.7%), and after 2012 $\hat{\rho}(e,t) = 0$ in every environment—the measurement process ceased recording negative outcomes entirely.

Table 2: Consolidated permutation-calibrated measurement audit on the CI registry. Each test uses a dual criterion (permutation p-value *and* effect-size floor); Test A is reported for both environment dimensions. All rows produce FAIL flags; overall diagnostic is FAIL.

Test	**Statistic**	**Value**	**Perm.** p	**Flag**
A: Label–era	Cramér's V	0.390	< 0.002	FAIL
A: Label–clinician	Cramér's V	0.260	< 0.002	FAIL
B: Env. predictive gain	ΔAUROC	$+0.013$	< 0.005	FAIL
C: Era monotonicity	Tail p	4/8 zero-neg	1.4×10^{-15}	FAIL
Overall				**FAIL**

In terms of the measurement model (Section 3), the registry uses an implicit default of $d = 1$: patients whose negative disposition is never documented are recorded as implanted—not an explicit design choice, but a consequence of workflow in which implantation is the expected outcome and non-implantation requires active documentation.

7.2. Measurement Audit Results

The audit produces a clear FAIL diagnostic (Table 2).

The strongest signal comes from Test C (era monotonicity): negative labels disappear entirely after era 4—i.e., after 2012; eras are indexed chronologically, with era 1 the earliest of the 8 audit bins (Table 1 aggregates them into 4 columns for display)—meaning the label is temporally bounded rather than prospectively observed. Test A confirms a strong label–era association ($V = 0.39$; asymptotic $\chi^2 = 531.6$; permutation $p < 0.002$), far exceeding what would arise by chance: the permutation 95th percentile of χ^2 is 13.4 versus the observed 531.6.

Test B reveals a subtler signal: adding environment variables (year + clinician) to clinical features increases AUROC by 0.013 (permutation $p < 0.005$). Part of this gain is trivially attributable to zero-negative strata, where environment perfectly predicts the observed label; the MIMIC-IV validation (Section 8) shows the test also detects non-trivial conditional gain in registries with no degenerate strata. While small in absolute terms, this confirms that environment carries predictive information *beyond* clinical features—precisely what the measurement model predicts when R correlates with X through (e, t).

7.3. Evaluation-Based Attribution

The attribution framework estimates the contribution of each confounding source (Table 3).

Attribution estimates. The IID AUROC of 0.978 (95% CI $[0.955, 0.991]$) partitions (approximately) into an era-attributable component $\hat{\Delta}_{\text{era}} = 0.102$ ($[0.058, 0.297]$; $\approx$10% of IID AUROC), a clinician-attributable component $\hat{\Delta}_{\text{clin}} = 0.011$ ($[-0.006, 0.235]$; $\approx$1.1%), and a total environment component $\hat{\Delta}_{\text{env}} = 0.643$ ($[0.412, 0.872]$; residualized AUROC $0.334 < 0.5$, indicating over-correction); these do not sum exactly due to non-additivity of

Table 3: AUROC under increasingly stringent evaluation regimes with cluster-bootstrap 95% CIs (1,000 replicates, clinician-level clusters). The era-blocked AUROC of 0.875 estimates transportable clinical discrimination after removing era confounding. These are protocol-dependent estimates, not causal effects.

Evaluation Regime	**AUROC**	**95% CI**	**$\hat{\Delta}$ from IID**
IID 5-fold CV (baseline)	0.978	$[0.955, 0.991]$	—
Era-blocked CV	0.875	$[0.668, 0.925]$	-0.102
Clinician-blocked CV	0.967	$[0.747, 0.990]$	-0.011
Doubly-controlled (era $\times$ clin)	0.912	—	-0.066
Residualized (env removed)	0.334	$[0.088, 0.578]$	-0.643
Environment alone (year + clin)	0.751	—	-0.227

evaluation-regime effects. The era effect dominates: negative labels are concentrated pre-2012 and pediatric patients are overrepresented in that era, so any model capturing age or patient-type variation implicitly discriminates era. The clinician effect is small with a 95% CI including zero, suggesting clinician identity contributes little beyond what era captures; and the residualized AUROC of 0.334 falls below 0.5, indicating that the environment model absorbs clinically meaningful, environment-correlated signal—consistent with the deep entanglement diagnosed by the audit and with the degenerate-regime diagnostic of Section 5.

Environment alone and doubly-controlled evaluation. A model using *only* entry year and clinician ID achieves AUROC = 0.751: this is the "free" discrimination any clinical model inherits by proxy through features correlated with (e, t). The most stringent regime—within-era, leave-one-clinician-out—yields AUROC $= 0.912 \pm 0.082$ with substantial variance across strata (range 0.785–1.000), suggesting meaningful but heterogeneous clinical signal remains after controlling for both environment dimensions.

7.4. IID Performance (for reference)

For completeness, IID results across model classes are reported in Appendix C (Table 5); all models achieve near-perfect discrimination under the IID assumption, and in light of the audit these numbers overstate clinical utility by the attribution estimates above.

7.5. Feature-Availability Ablation

Age at implantation is determined by—and undefined without—the outcome, so it would not be available at the time a candidacy prediction is made. Re-running the full pipeline without `age_at_implant`, the IID AUROC drops from 0.978 to 0.964, the era-attributable confounding share is 9.2% (versus 10.2% with the feature), and the audit verdict is unchanged. Temporal feature leakage therefore accounts for at most a small fraction of the apparent performance; the measurement-process confound persists independently of it.

7.6. Feature Importance Stability

SHAP feature-importance rankings are remarkably stable across clinician environments, but the top features (patient type, cochlear anatomy) are exactly those entangled with the documentation confound; Appendix D reports the full analysis, including a value-of-information study showing substantial redundancy in the registry's data dictionary.

8. External Validation: MIMIC-IV

To assess whether the audit's operating properties extend beyond a single registry with degenerate confounding, we apply the identical battery and attribution hierarchy to the MIMIC-IV ICU mortality cohort (Johnson et al., 2023) (Beth Israel Deaconess Medical Center; available via PhysioNet). This dataset differs from the CI registry in every dimension one would want varied: registry type (EHR-derived ICU data vs. procedural audiology registry), outcome prevalence (11.0% vs. 5.5%), environment structure (care units vs. clinicians), and—critically—confounding structure (diffuse and unit-driven, with no zero-negative strata, vs. degenerate and era-driven).

Cohort and features. The cohort comprises $n = 67{,}224$ ICU stays with in-hospital mortality (11.0%) as the outcome and 48 first-24-hour features (demographics, insurance, admission type, length of stay, and 42 vital-sign and laboratory summaries). Environments are the 7 care units remaining after consolidating units with fewer than 5,000 stays into an "Other" category; eras are 5 admission-year groups. Continuous features are median-imputed per environment (imputers fit on training folds only inside each cross-validation split); binary features are indicator-augmented; no features are excluded a priori, so environment-entangled "shortcut" variables are flagged by the analysis itself rather than by curation.

Audit results. Table 4 summarizes the audit. Permutation Test A detects significant unit-level confounding ($V = 0.145$, permutation $p < 0.002$; the maximum V across 500 permuted replicates is 0.017), while the era-level association falls below the effect-size floor ($V = 0.042$): the audit FAILs on the unit dimension and PASSes on era. Test B detects a conditional predictive gain under logistic regression (ΔAUROC $= +0.014$; the gap shrinks to $+0.002$ under XGBoost, which absorbs unit-correlated signal through features—Appendix H), and Test C passes (0/5 zero-negative eras): mortality is reliably recorded across units and eras, so the reporting indicator is effectively constant and the detected confounding arises through environment-conditional label prevalence rather than differential reporting. Unlike the CI registry, no stratum is degenerate, so Test B's gain cannot be explained by zero-negative label structure.

Attribution. The attribution hierarchy yields IID AUROC $= 0.815$; unit-blocked $= 0.787$; era-blocked $= 0.811$; residualized $= 0.753$; environment-only $= 0.626$—an environment-attributable confounding share of 3.4% of IID AUROC, versus roughly 10% in the CI registry. Notably, the residualized AUROC does not collapse below 0.5 as on the CI registry (0.753 vs. 0.334): per the regime diagnostic of Section 5, the two registries are distinguished cleanly—degenerate confounding (10% share, residualized 0.334, 4/8 zero-negative eras) versus diffuse confounding (3.4% share, residualized 0.753, 0/5 zero-negative eras)—

Table 4: Measurement audit on MIMIC-IV (n = 67,224). The audit FAILs on the care-unit dimension while correctly passing era-level tests—the mirror image of the CI registry's degenerate, era-driven confounding.

Test	**Statistic**	**Value**	**Perm.** p	**Flag**
A: Label–unit	Cramér's V	0.145	< 0.002	FAIL
A: Label–era	Cramér's V	0.042	—	PASS
B: Env. predictive gain (LR)	ΔAUROC	+0.014	< 0.005	FAIL
C: Era monotonicity	zero-neg. eras	0/5	—	PASS
Overall (unit dimension)				**FAIL**

demonstrating that the audit provides a severity-graded, regime-aware assessment rather than a single binary verdict.

Comparison to generic two-sample tests and simple baselines. On the CI registry, an MMD two-sample test on the feature matrix (Gretton et al., 2012) returns $p = 0.23$ (non-significant) and a classifier two-sample test (Lopez-Paz and Oquab, 2017) likewise fails to flag the registry, while Test A returns $p < 0.002$: the confounding operates through the label-generating process, which feature-space tests do not examine. Conversely, simple diagnostics (prevalence-by-era plots, environment-only classifiers, temporal splits) do flag the CI registry's degenerate structure on their own—but on MIMIC-IV the prevalence gradient is modest and visually dismissible, while the audit's calibrated thresholds provide an unambiguous quantitative flag. The audit's value is standardized detection with calibrated thresholds, not any single novel statistic.

Attribution reliability guidance. The CI registry's wide era-blocked interval ($[0.668, 0.925]$) reflects its small number of clinician clusters; MIMIC-IV's per-unit blocked AUROCs are substantially tighter. As practical guidance, reliable attribution requires at least 4–5 environment clusters with ≥ 100 observations each; below this, report the audit verdict (which remains valid) without attribution point estimates.

9. Remediation After a FAIL Audit

What should researchers do after a FAIL? We outline five strategies, ordered roughly by increasing effort, distinguishing those that *evaluate around* the bias (1, 2, 4) from those that address its structural source—the imputation rule g itself (3, 5).

1. Report environment-aware evaluations as primary results. Replace IID AUROC with era-blocked or clinician-blocked AUROC as the headline metric—no additional data collection, only a change in evaluation protocol. In our registry, this shifts the reported AUROC from 0.978 to 0.875 (era-blocked) or 0.967 (clinician-blocked).

2. Model the reporting function explicitly. If $\hat{\rho}(e, t)$ can be estimated reliably, reweight or restrict the analysis to strata where reporting is active ($\hat{\rho} > 0$). In our registry,

restricting to the pre-2012 era would reduce the sample from 3,584 to 1,859 but yield a more interpretable prediction target.

3. Reconstruct labels via chart review. Where the true status Y^* is determinable but undocumented, targeted chart review of patients in low-$\hat{\rho}$ strata can recover missing negative labels; the reporting function itself supplies the sampling strategy (review strata where $\hat{\rho}(e,t) = 0$). This presupposes that negative outcomes are documented somewhere in the record.

4. Apply invariant learning methods. Where label reconstruction is infeasible, IRM (Arjovsky et al., 2019) or DRO (Sagawa et al., 2020) can optimize for cross-environment stability; the audit supplies the environment definitions, and the attribution estimates indicate whether the potential gain justifies the complexity. A caution: environment-correlated features such as age, patient type, or entry era are often genuinely prognostic; down-weighting them trades clinical signal for stability without fixing the structural source of the bias, which is why Strategies 3 and 5 are preferable when feasible.

5. Restrict scope to actively-reporting strata. The most direct structural remedy: restrict *both* training and evaluation to strata where negative labels are actively recorded, and declare inactive-reporting regimes out of scope for prediction. In our registry the pre-2012 regime is where the prediction problem is well-posed; the post-2012 regime is not a prediction task but a documentation artifact. A pre-2012-only model with a sharply defined scope statement is methodologically cleaner than a mixed-regime AUROC, at the cost of sample size and of leaving the recent regime unmodeled until documentation practice changes or chart review (Strategy 3) restores label variance.

10. Discussion

10.1. From Case Study to Diagnostic Toolkit

Our central contribution is not the finding that a specific CI registry has confounded labels—though this is clinically relevant—but rather the *toolkit* for detecting and quantifying such confounding in any registry: the measurement process model (Section 3) identifies the conditions under which IID evaluation can mislead, the audit (Section 4) operationalizes detection through three tests requiring only environment metadata and labels, and the attribution framework (Section 5) provides protocol-dependent estimates of confounding magnitude. We recommend applying the toolkit as a *preprocessing step* before any registry-based prediction study: Tests A and C require no model training, and Test B needs only a baseline comparison.

10.2. Implications for Clinical ML Reporting

Our results suggest three reporting practices. (1) *Report the measurement audit*: before presenting model performance, report the audit results (Table 2); a FAIL indicates that IID metrics should not be treated as estimates of clinical utility. (2) *Report attribution estimates*: Table 3 makes the magnitude of confounding transparent—here the headline

number drops from 0.978 to 0.875 (era-blocked), a meaningful difference for clinical decision-making. (3) *Investigate label provenance*: report $\hat{\rho}(e, t)$ (Table 1) to make the measurement process visible before any modeling.

10.3. What the Audit Can and Cannot Detect

The audit detects statistical dependence between labels and *observed* environment variables; two boundaries follow. First, legitimate changes in clinical practice are observationally indistinguishable from documentation artifacts at this level: CI candidacy criteria genuinely expanded over the study period, so part of the era–label association may reflect true policy change rather than reporting failure—a failing registry should undergo clinical review to separate the two, since the audit localizes *where* labels depend on environment, not *why*. Second, latent confounders recorded nowhere are out of scope—the clinically more dangerous case, requiring design-based remedies rather than retrospective auditing. Finally, the audit assumes environments are observed at meaningful granularity: when only institution-level identifiers are available, Tests A and C remain applicable but Test B's power degrades, so registries should record era metadata alongside institution identifiers.

10.4. Limitations

Our validation spans two registries with contrasting confounding structures—a single-center procedural registry with degenerate, era-driven confounding and a large public EHR-derived cohort (MIMIC-IV) with diffuse, unit-driven confounding—but both are single-institution datasets; further multi-domain validation (e.g., eICU, multi-center registries, continuous environments) is needed before the audit could anchor a formal reporting standard. The attribution estimates are protocol-dependent and should not be interpreted as causal effects. The reporting function $\hat{\rho}(e, t)$ is estimated nonparametrically and may be unreliable in small strata; we mitigate this with the minimum-count pooling threshold of Section 3. The semi-synthetic calibration study (Appendix I) and the empirical family-wise-error check on MIMIC-IV demonstrate favorable operating properties; calibration across a broader range of registry structures remains future work.

11. Conclusion

We have presented a measurement process model, audit algorithm, and evaluation-based attribution framework for detecting and quantifying label–environment confounding in clinical registries. On a cochlear implant registry, the audit correctly identifies that roughly 10% of IID AUROC traces to documentation artifacts; the era-blocked AUROC of 0.875 (95% CI $[0.668, 0.925]$) better estimates transportable discrimination than the headline 0.978, and the FAIL verdict is reproduced in 81% of 1,000 cluster-bootstrap replicates with 0% producing a PASS (Appendix J). Applied to MIMIC-IV, the same battery detects diffuse, unit-driven confounding (3.4% of IID AUROC; residualized AUROC 0.753) while correctly passing era-level tests—severity-graded, regime-aware assessment across qualitatively different registries. The framework is lightweight, requires only environment metadata, and applies to any registry-based study; we propose it as a candidate pre-modeling diagnostic and provide remediation strategies for registries that fail it.

Ethics Statement

This study used a de-identified clinical registry. All personally identifiable information was removed prior to analysis; patient and clinician identifiers were replaced with SHA-256 hashes that preserve grouping structure for environment-aware evaluation without exposing identity. The study protocol was reviewed and approved by the institutional review board of the contributing institution under an exempt / minimal-risk protocol, consistent with the use of de-identified retrospective registry data. The contributing institution is University of Michigan Medicine. No patient-identifiable imagery, audio, or free-text clinical notes were used in any analysis reported in this paper. The MIMIC-IV analyses use the publicly available, de-identified MIMIC-IV database accessed under the PhysioNet credentialed-access data use agreement (Johnson et al., 2023).

Reproducibility

Code for data cleaning, the measurement audit algorithm (packaged as the standalone `registry-audit` toolkit), and the attribution framework is included in the supplementary materials and available at `https://github.com/TODO/measurement-audit-mlhc2026`. The MIMIC-IV validation additionally includes the complete extraction pipeline (first-24-hour labs and vitals, environment definition, cohort construction) and the full audit outputs (all test statistics, attribution metrics, and per-environment and per-era AUROCs). The audit algorithm is implemented as a standalone module that can be applied to any dataset with environment metadata and binary labels. All reported experiments use fixed random seeds (`numpy` and `xgboost` seeds set to 42 for the primary model; 500 permutations for Test A and 1,000 cluster-bootstrap replicates for attribution CIs, each with independent per-replicate seeds derived from a single master seed). Primary models were trained on a single workstation CPU (no GPU required); the full audit and attribution pipeline runs in under 10 minutes on 3,584 patients.

References

R. H. Gifford, M. F. Dorman, A. J. Spahr, and S. A. Bacon. Auditory function and speech understanding in listeners who qualify for EAS surgery. *Ear and Hearing*, 31(1):2–12, 2010.

D. L. Sorkin. Cochlear implantation in the world's largest medical device market. *Cochlear Implants International*, 14(sup1):S4–S12, 2013.

M. G. Crowson, V. Lin, J. M. Chen, and T. D. Chan. Machine learning and cochlear implantation. *Otology & Neurotology*, 41(1):e36–e45, 2020.

A. Subbaswamy and S. Saria. From development to deployment: dataset shift, causality, and shift-stable models in health AI. *Biostatistics*, 21(2):345–352, 2020.

S. G. Finlayson, A. Subbaswamy, K. Singh, et al. The clinician and dataset shift in artificial intelligence. *New England Journal of Medicine*, 385(3):283–286, 2021.

K. G. M. Moons, J. A. H. de Groot, W. Bouwmeester, et al. Critical appraisal and data extraction for systematic reviews of prediction modelling studies: The CHARMS checklist. *PLoS Medicine*, 11(10):e1001744, 2014.

B. Nestor, M. McDermott, W. Boag, et al. Feature robustness in non-stationary health records. In *Proceedings of MLHC*, 2019.

A. Wong, E. Otles, J. P. Donnelly, et al. External validation of a widely implemented proprietary sepsis prediction model. *JAMA Internal Medicine*, 181(8):1065–1070, 2021.

M. Arjovsky, L. Bottou, I. Gulchan, and D. Lopez-Paz. Invariant risk minimization. *arXiv preprint arXiv:1907.02893*, 2019.

S. Sagawa, P. W. Koh, T. B. Hashimoto, and P. Liang. Distributionally robust neural networks for group shifts. In *ICLR*, 2020.

L. Oakden-Rayner, J. Dunnmon, G. Carneiro, and C. Ré. Hidden stratification causes clinically meaningful failures in machine learning for medical imaging. In *Proceedings of ACM CHIL*, 2020.

S. M. Lundberg and S.-I. Lee. A unified approach to interpreting model predictions. In *NeurIPS*, 2017.

A. Agresti. *Categorical Data Analysis*. Wiley, 2nd edition, 2002.

J. Cohen. *Statistical Power Analysis for the Behavioral Sciences*. Routledge, 2nd edition, 1988.

S. Rabanser, S. Günnemann, and Z. Lipton. Failing loudly: An empirical study of methods for detecting dataset shift. In *NeurIPS*, 2019.

S. E. Davis, T. A. Lasko, G. Chen, E. D. Siew, and M. E. Matheny. Calibration drift in regression and machine learning models for acute kidney injury. *Journal of the American Medical Informatics Association*, 24(6):1052–1061, 2017.

C. Sáez, A. Gutiérrez-Sacristán, I. Kohane, J. M. García-Gómez, and P. Avillach. EHRtemporalVariability: delineating temporal data-set shifts in electronic health records. *GigaScience*, 9(8):giaa079, 2020.

D. Vela, A. Sharp, R. Zhang, T. Nguyen, A. Hoang, and O. S. Pianykh. Temporal quality degradation in AI models. *Scientific Reports*, 12:11654, 2022.

C. G. Northcutt, A. Athalye, and J. Mueller. Pervasive label errors in test sets destabilize machine learning benchmarks. In *NeurIPS Track on Datasets and Benchmarks*, 2021.

C. G. Northcutt, L. Jiang, and I. Chuang. Confident learning: Estimating uncertainty in dataset labels. *Journal of Artificial Intelligence Research*, 70:1373–1411, 2021.

D. Karimi, H. Dou, S. K. Warfield, and A. Gholipour. Deep learning with noisy labels: exploring techniques and remedies in medical image analysis. *Medical Image Analysis*, 65:101759, 2020.

R. J. Carroll, D. Ruppert, L. A. Stefanski, and C. M. Crainiceanu. *Measurement Error in Nonlinear Models: A Modern Perspective*. Chapman & Hall/CRC, 2nd edition, 2006.

M. A. Hernán and J. M. Robins. *Causal Inference: What If*. Chapman & Hall/CRC, 2020.

B. A. Goldstein, A. M. Navar, M. J. Pencina, and J. P. A. Ioannidis. Opportunities and challenges in developing risk prediction models with electronic health records data: a systematic review. *Journal of the American Medical Informatics Association*, 24(1):198–208, 2017.

N. G. Weiskopf, G. Hripcsak, A. Swaminathan, and C. Weng. Defining and measuring completeness of electronic health records for secondary use. *Journal of Biomedical Informatics*, 46(3):424–430, 2013.

G. Hripcsak and D. J. Albers. Next-generation phenotyping of electronic health records. *Journal of the American Medical Informatics Association*, 20(1):117–121, 2013.

I. Gulrajani and D. Lopez-Paz. In search of lost domain generalization. In *ICLR*, 2021.

P. W. Koh, S. Sagawa, H. Marklund, et al. WILDS: A benchmark of in-the-wild distribution shifts. In *ICML*, volume 139, pages 5637–5664, 2021.

E. Creager, J.-H. Jacobsen, and R. Zemel. Environment inference for invariant learning. In *ICML*, volume 139, pages 2189–2199, 2021.

E. W. Steyerberg. *Clinical Prediction Models: A Practical Approach to Development, Validation, and Updating*. Springer, 2nd edition, 2019.

R. D. Riley, J. P. Ensor, K. I. E. Snell, et al. Calculating the sample size required for developing a clinical prediction model. *BMJ*, 368:m441, 2020.

G. S. Collins, J. B. Reitsma, D. G. Altman, and K. G. M. Moons. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): The TRIPOD statement. *Annals of Internal Medicine*, 162(1):55–63, 2015.

G. S. Collins, K. G. M. Moons, P. Dhiman, R. D. Riley, A. L. Beam, B. Van Calster, et al. TRIPOD+AI statement: Updated guidance for reporting clinical prediction models that use regression or machine learning methods. *BMJ*, 385:e076411, 2024.

R. F. Wolff, K. G. M. Moons, R. D. Riley, et al. PROBAST: A tool to assess the risk of bias and applicability of prediction model studies. *Annals of Internal Medicine*, 170(1):51–58, 2019.

D. B. Rubin. Inference and missing data. *Biometrika*, 63(3):581–592, 1976.

R. J. A. Little and D. B. Rubin. *Statistical Analysis with Missing Data*. Wiley, 3rd edition, 2019.

S. R. Seaman and I. R. White. Review of inverse probability weighting for dealing with missing data. *Statistical Methods in Medical Research*, 22(3):278–295, 2013.

J. J. Heckman. Sample selection bias as a specification error. *Econometrica*, 47(1):153–161, 1979.

J. Pearl and E. Bareinboim. Transportability of causal and statistical relations: A formal approach. In *Proceedings of the 25th AAAI Conference on Artificial Intelligence*, pages 247–254, 2011.

E. Bareinboim and J. Pearl. Causal inference and the data-fusion problem. *Proceedings of the National Academy of Sciences*, 113(27):7345–7352, 2016.

A. Subbaswamy, P. Schulam, and S. Saria. Preventing failures due to dataset shift: Learning predictive models that transport. In *AISTATS*, PMLR 89, pages 3118–3127, 2019.

J. R. Zech, M. A. Badgeley, M. Liu, A. B. Costa, J. J. Titano, and E. K. 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.

A. J. DeGrave, J. D. Janizek, and S.-I. Lee. AI for radiographic COVID-19 detection selects shortcuts over signal. *Nature Machine Intelligence*, 3(7):610–619, 2021.

L. Wynants, B. Van Calster, G. S. Collins, et al. Prediction models for diagnosis and prognosis of covid-19: systematic review and critical appraisal. *BMJ*, 369:m1328, 2020.

M. Roberts, D. Driggs, M. Thorpe, et al. Common pitfalls and recommendations for using machine learning to detect and prognosticate for COVID-19 using chest radiographs and CT scans. *Nature Machine Intelligence*, 3(3):199–217, 2021.

Z. Obermeyer, B. Powers, C. Vogeli, and S. Mullainathan. Dissecting racial bias in an algorithm used to manage the health of populations. *Science*, 366(6464):447–453, 2019.

L. Seyyed-Kalantari, H. Zhang, M. B. A. McDermott, I. Y. Chen, and M. Ghassemi. Underdiagnosis bias of artificial intelligence algorithms applied to chest radiographs in underserved patient populations. *Nature Medicine*, 27(12):2176–2182, 2021.

A. Rajkomar, M. Hardt, M. D. Howell, G. Corrado, and M. H. Chin. Ensuring fairness in machine learning to advance health equity. *Annals of Internal Medicine*, 169(12):866–872, 2018.

A. K. Menon, B. van Rooyen, C. S. Ong, and R. C. Williamson. Learning from corrupted binary labels via class-probability estimation. In *ICML*, volume 37, pages 125–134, 2015.

R. A. Fisher. *Statistical Methods for Research Workers*. Oliver & Boyd, 4th edition, 1932.

M. B. Brown. A method for combining non-independent, one-sided tests of significance. *Biometrics*, 31(4):987–992, 1975.

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

D. Lopez-Paz and M. Oquab. Revisiting classifier two-sample tests. In *ICLR*, 2017.

A. E. W. Johnson, L. Bulgarelli, L. Shen, et al. MIMIC-IV, a freely accessible electronic health record dataset. *Scientific Data*, 10(1):1, 2023.

H. Suresh and J. Guttag. A framework for understanding sources of harm throughout the machine learning life cycle. In *ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization (EAAMO)*, 2021.

Appendix A. Extended Related Work

This appendix expands the condensed discussion of Section 2.

Dataset shift in clinical ML. Distribution shift between training and deployment is well-documented in medical AI (Subbaswamy and Saria, 2020). Temporal validation has been recommended as a stronger alternative to random cross-validation (Moons et al., 2014), and several studies have shown performance degradation under temporal or geographic shifts (Nestor et al., 2019; Wong et al., 2021). Finlayson et al. (2021) demonstrated how clinician ordering patterns confound diagnostic models. Rabanser et al. (2019) provided an empirical comparison of dataset shift detection methods, finding that two-sample testing approaches are most effective across domains including healthcare. Our work differs from this literature by tracing confounding to the *label-generating process itself*—not just covariate shift in features—and providing a pre-modeling diagnostic framework rather than a post-hoc finding. In Section 8 we also compare the audit quantitatively against generic two-sample shift-detection baselines—an MMD test on features (Gretton et al., 2012) and a classifier two-sample test (Lopez-Paz and Oquab, 2017)—and find that they fail to flag label–environment confounding that operates through the label-generating process rather than through covariate shift.

Temporal drift and calibration. Clinical prediction models are known to degrade over time even when underlying disease prevalence is stable. Davis et al. (2017) showed that calibration drift in acute kidney injury models accumulates over months, requiring periodic recalibration. Sáez et al. (2020) developed tools for delineating temporal variability in EHR data, highlighting how institutional workflow changes generate non-stationary data streams. Vela et al. (2022) demonstrated that 91% of ML models degrade over time independent of detectable data drift, suggesting that unmeasured process changes—of the kind our measurement model formalizes—may be responsible. These findings motivate our audit: if degradation can occur without detectable covariate shift, the label-generating process itself may be changing.

Selection bias and missing data. Our framework draws directly on the classical missing-data literature. Rubin's taxonomy of missing-data mechanisms (1976)—MCAR, MAR, and MNAR—provides the formal vocabulary for our reporting indicator $R(e, t)$, which induces a missing-not-at-random (MNAR) pattern on Y^* whenever ρ depends on (e, t). Little and Rubin's canonical treatment (2019) catalogs likelihood-based, imputation, and weighting methods for principled inference under missingness, while Seaman and White (2013) review inverse-probability weighting approaches that are closely related to the reporting-function reweighting we propose as a remediation strategy (Section 9). Heckman's selection

model (1979) gave an early econometric account of how sample-inclusion processes correlated with outcomes bias downstream estimates; our setting is a discrete analogue in which inclusion in the "documented" stratum depends jointly on environment and era. A crucial distinction from prior missing-data work is that missingness here is not simply a nuisance to impute over, but an artifact that *inflates* apparent discriminative performance when the registry imputes a default label d.

Label quality and annotation. Label quality has been studied in medical imaging, where Oakden-Rayner et al. (2020) showed that hidden stratification within nominally correct labels produces clinically meaningful model failures. Northcutt et al. (2021a) identified pervasive label errors ($\geq$3.3%) across standard ML benchmarks, demonstrating that label noise destabilizes model evaluation even in curated test sets. The confident learning framework (Northcutt et al., 2021b) provides principled methods for estimating and correcting label errors by jointly reasoning about data and label uncertainty. Karimi et al. (2020) reviewed strategies for handling label noise in medical image analysis, categorizing six families of approaches from data cleaning to noise-robust loss functions. Our setting is distinct from all of these: the labels are not "noisy" in the classical sense but rather *systematically absent* in certain environment–era strata, creating structured bias that standard label-noise methods do not address.

Measurement error and data completeness. The statistical theory of measurement error is well-developed in epidemiology and biostatistics (Carroll et al., 2006; Hernán and Robins, 2020), with particular attention to how misclassification biases regression coefficients and causal estimates. In the EHR context, Goldstein et al. (2017) systematically reviewed 107 EHR-based prediction studies and identified data quality—including completeness, coding variability, and documentation heterogeneity—as a pervasive challenge. Weiskopf and Weng (2013) defined four types of EHR completeness (documentation, breadth, density, and predictive) and proposed measurement methods, establishing a taxonomy that our reporting function $\hat{\rho}(e, t)$ operationalizes for registry labels specifically. Hripcsak and Albers (2013) discussed how automated phenotyping from EHR data is complicated by incomplete and inconsistent documentation practices. Our measurement process model extends this line of work by formalizing the label-generating process as a function of environment and era, providing testable conditions under which documentation artifacts inflate evaluation metrics.

Invariance, robustness, and domain generalization. The invariant risk minimization (IRM) framework (Arjovsky et al., 2019) learns predictors stable across environments; distributionally robust optimization (DRO) (Sagawa et al., 2020) optimizes for worst-case group performance. Gulrajani and Lopez-Paz (2021) introduced DomainBed, a comprehensive testbed for domain generalization algorithms, and found that careful hyperparameter tuning of empirical risk minimization (ERM) often matches or exceeds specialized algorithms—suggesting that identifying the right evaluation protocol may matter more than the training algorithm. Koh et al. (2021) extended this to real-world settings with the WILDS benchmark, including healthcare tasks (tumor identification across hospitals), documenting substantial performance gaps under distribution shift. Creager et al. (2021) addressed a practical bottleneck: when environment labels are unavailable, they proposed

methods for inferring latent environment structure from data, a complement to our approach where environments are observed but their effect on *labels* must be diagnosed. Our attribution framework identifies the specific failure modes these methods are designed to address, serving as a diagnostic prerequisite: one should audit whether the label–environment entanglement is in the measurement process before applying invariant learning methods that assume the labels are trustworthy.

Causal transportability and target validity. A parallel literature frames distribution shift causally. Pearl and Bareinboim (2011) formalized *transportability*—when causal or statistical conclusions from one population can be transferred to another—using selection diagrams that make source–target differences explicit. Bareinboim and Pearl (2016) generalized this to the data-fusion problem, in which heterogeneous studies must be combined under known structural assumptions. Subbaswamy et al. (2019) applied these ideas to prediction, learning models that are provably transportable across specified shift variables. Our audit is complementary: the selection diagrams that enable transportability analysis assume a known causal graph, while our tests *empirically* detect when the label itself depends on environment regardless of the graph, serving as a pre-modeling sanity check before transportability claims are made.

Shortcut learning and deployment failures. Several high-profile post-hoc analyses have shown how confounding in medical ML surfaces only at deployment. Zech et al. (2018) found that a pneumonia CNN achieving strong in-hospital AUC transferred poorly across hospitals because it had learned hospital-specific radiographic artifacts rather than pathological signal. DeGrave et al. (2021) used saliency analysis to show that several COVID-19 chest-radiograph classifiers relied on shortcuts—laterality markers, text tokens, patient positioning—rather than disease features. The systematic reviews of COVID-19 prognostic models by Wynants et al. (2020) and Roberts et al. (2021) concluded that essentially none of the several hundred reviewed models were clinically deployable, citing exactly the measurement-process failures we formalize: biased sampling, label leakage, and opaque cohort definitions. Our audit aims to surface these failure modes *before* such effort is invested rather than as a post-deployment forensic finding.

Fairness and algorithmic bias in health ML. Obermeyer et al. (2019) demonstrated that a widely deployed care-management algorithm systematically under-allocated resources to Black patients because its proxy label (healthcare cost) was differentially recorded across groups—a measurement-process confound in the label itself. Seyyed-Kalantari et al. (2021) showed that chest-X-ray classifiers systematically under-diagnose historically under-served populations across three public datasets. Rajkomar et al. (2018) surveyed fairness considerations for health ML and argued that equity requires attention to who is represented in training data and how their outcomes are measured. Suresh and Guttag (2021) formalized a taxonomy of harms across the ML lifecycle in which measurement bias is a distinct and remediable source. These works share a throughline with our framing: when the observed label is a function of a documentation process that varies across subgroups or environments, metrics computed on it cannot be read off as clinical utility.

Clinical prediction models and registries. Cochlear implantation is a well-studied domain for clinical ML: Crowson et al. (2020) provide a structured review of opportunities and

challenges for ML applied to CI data, surveying prior efforts at CI outcome prediction and identifying candidacy prediction as a core application. More broadly, Steyerberg (2019) provides a comprehensive treatment of clinical prediction model development, validation, and updating, including guidance on handling heterogeneous data sources. Riley et al. (2020) addressed the statistical requirements for external validation, emphasizing that adequate validation requires both sufficient sample size and appropriate distributional coverage—a concern amplified when labels are environment-dependent.

Reporting standards. The TRIPOD statement (Collins et al., 2015) established a 22-item checklist for transparent reporting of prediction model studies, recently extended to 27 items covering AI and ML methods in TRIPOD+AI (Collins et al., 2024). The PROBAST tool (Wolff et al., 2019) provides a complementary framework for assessing risk of bias and applicability in prediction model studies. While these guidelines address many aspects of model development and reporting, none currently require auditing the measurement process that generates labels. Our audit algorithm could serve as a concrete addition to such reporting frameworks: a pre-modeling diagnostic that determines whether IID evaluation metrics can be meaningfully interpreted as estimates of clinical utility.

Appendix B. Measurement Audit Pseudocode

Algorithm 1 summarizes the full measurement audit procedure described in Section 4. The three tests require only labels, environment metadata, and (for Test B) clinical features; the routine returns a single PASS / CAUTION / FAIL diagnostic based on the decision rule of Section 4.4.

Algorithm 1 Measurement Audit for Clinical Registries

Require: Labels Y, environment $E = (e, t)$, features X
Ensure: Diagnostic flag $\in \{\text{PASS}, \text{CAUTION}, \text{FAIL}\}$

Test A: For Y vs. e and Y vs. t: compute χ^2 and Cramér's V; permute Y 500 times to obtain the null distribution and permutation p-value
Test B: $\Delta\text{AUROC} \leftarrow \text{AUROC}(f_{X,E}) - \text{AUROC}(f_X)$ via 5-fold CV; permutation null via within-fold shuffling of E (200 permutations)
Test C: Count era bins with zero negative labels; binomial tail probability via Poisson approximation
Score each test as PASS / CAUTION / FAIL per the thresholds of Section 4.4
if n_FAIL ≥ 2 **then**
  return FAIL
else if n_FAIL ≥ 1 or n_CAUTION ≥ 2 **then**
  return CAUTION
else
  return PASS
end if

Appendix C. IID Performance Across Model Classes

Table 5: IID 5-fold cross-validation (mean ± SD across folds, per model class). The attribution analyses (Table 3) use the primary class-weighted XGBoost with a fixed seed, whose pooled IID AUROC is 0.978; small differences between the two tables reflect the fold-aggregation method (pooled vs. per-fold mean), not a discrepancy. These results overstate clinical utility; see Table 3 for environment-corrected estimates.

Model	AUROC	AUPRC	Brier
XGBoost	0.983 ± 0.005	0.999 ± 0.000	0.040 ± 0.007
Logistic	0.974 ± 0.005	0.998 ± 0.000	0.059 ± 0.009
Random Forest	0.980 ± 0.006	0.999 ± 0.000	0.033 ± 0.006

Appendix D. Feature Importance and Value of Information

SHAP analysis (Lundberg and Lee, 2017) reveals that feature importance rankings are remarkably stable across clinician environments, despite the confounded evaluation: four features appear in every clinician's top-5—`patient_type` (mean |SHAP| = 0.549), `normal_cochlea` (0.213), `onset_type` (0.019), and `age_at_entry` (0.016)—all aligned with established clinical factors for CI candidacy. This stability merits careful interpretation: the model *is* capturing genuine clinical patterns, but these same features correlate with the documentation artifact (recall the patient-type/age collinearity of Section 6—pediatric patients are overrepresented in the pre-2012 era when negative labels were recorded). Stability across clinicians therefore reflects both real clinical signal and the consistency of the documentation confound across environments; the attribution framework (Table 3) is needed to disentangle the two.

Value-of-information analysis shows that only `patient_type` (ΔAUC = +2.6%) and `age_at_implant` (+1.0%) provide meaningful marginal information. All other registry fields individually contribute $< 0.5\%$, suggesting substantial redundancy in the data dictionary. We note that this analysis excludes entry year but retains features (such as patient type) that remain entangled with era; a fully principled distinction between excludable and non-excludable environment-correlated features is an open problem.

Appendix E. Full Clinician-Level Results

Table 6 reports the per-clinician results underlying the aggregate clinician-blocked AUROC reported in Table 3. Of the eight clinicians in the registry, only the four shown below contribute at least one negative label in a held-out fold and are therefore evaluable; the remaining four clinicians have ≈100% implantation rates, yielding undefined AUROC under leave-one-clinician-out. The mean across evaluable clinicians (0.972) is close to the IID AUROC, consistent with the small clinician-attributable attribution $\hat{\Delta}_{\text{clin}} = 0.011$.

Table 6: Clinician-blocked cross-validation. Only clinicians with ≥ 1 negative label in the test set are evaluable. Four clinicians with ≈100% implantation rates yield undefined AUROC.

Clinician	n	**P(neg)**	**AUROC**
clin_D	665	0.005	0.999
clin_A	1,135	0.014	0.972
clin_E	333	0.138	0.954
clin_F	994	0.128	0.961
Mean			0.972

Appendix F. Doubly-Controlled Results

Table 7 reports the stratum-level results for the doubly-controlled regime (within-era, leave-one-clinician-out) summarized in Table 3. Only six (era, clinician) strata have ≥ 10 patients and ≥ 1 negative label in both the training and held-out partitions; all six lie in the pre-2012 era, since negative labels are absent thereafter. Within-stratum AUROC ranges from 0.785 to 1.000, with a mean of 0.912, suggesting that meaningful but heterogeneous clinical signal remains after controlling for both environment dimensions.

Table 7: Doubly-controlled evaluation (within-era, leave-one-clinician-out). Only strata with ≥ 10 patients and ≥ 1 negative label in both train and test are evaluable.

Era	**Test Clinician**	n	**AUROC**
≤2008	clin_D	12	1.000
≤2008	clin_A	280	0.960
≤2008	clin_E	329	0.934
≤2008	clin_F	728	0.975
2009–2012	clin_A	79	0.816
2009–2012	clin_F	153	0.785
Mean			0.912

Appendix G. Proof of Proposition 2

Notation. Let $P = \{i : Y_i^* = 1\}$ and partition true negatives into reported N_R ($R_i = 1$) and unreported N_U ($R_i = 0$). Under $d = 1$, the operational positive class is $P \cup N_U$ and the operational negative class is N_R.

We first state the mixture decomposition referenced in Section 3.

Corollary 5 (Predictable bias direction) *Let $P = \{i : Y_i^* = 1\}$, and partition the true negatives into reported (N_R: $Y_i^* = 0, R_i = 1$) and unreported (N_U: $Y_i^* = 0, R_i = 0$). Write A_{PR}, A_{UR}, A_{PU} for the AUROC of f separating P from N_R, N_U from N_R, and P from*

N_U respectively, and let $\pi = |N_U|/(|P| + |N_U|)$ and $\nu = |N_U|/(|N_R| + |N_U|)$. Under $d = 1$, exact mixture decompositions give

$$AUROC(f, Y) = (1 - \pi)A_{PR} + \pi A_{UR}, \qquad AUROC(f, Y^*) = (1 - \nu)A_{PR} + \nu A_{PU}, \quad (15)$$

so inflation occurs if and only if $\pi(A_{UR} - A_{PR}) > \nu(A_{PU} - A_{PR})$. Two regimes follow. (a) Deterministic collapse*: if reporting is deterministic within environments and f separates reporting-inactive from reporting-active environments ($A_{UR} \to 1$) while $A_{PU} \leq A_{PR}$, inflation follows—the regime of our CI case study. (b)* Class-conditional corruption*: if unreported negatives have the same score distribution as reported negatives, then $A_{UR} = \frac{1}{2}$ and $A_{PU} = A_{PR}$, so the observed AUROC is deflated toward $\frac{1}{2}$ whenever $A_{PR} > \frac{1}{2}$—the regime of the stochastic semi-synthetic simulation (Appendix I).*

Step 1: mixture identity. AUROC is a mean over positive–negative pairs. Partitioning the operational positive class,

$$\text{AUROC}(f, Y) = \frac{|P|}{|P| + |N_U|} A_{PR} + \frac{|N_U|}{|P| + |N_U|} A_{UR} = (1 - \pi)A_{PR} + \pi A_{UR},$$

where A_{PR} (A_{UR}) is the AUROC of f separating P (resp. N_U) from N_R, and $\pi = |N_U|/(|P| + |N_U|)$. Partitioning the *true* negative class analogously gives $\text{AUROC}(f, Y^*) = (1 - \nu)A_{PR} + \nu A_{PU}$ with $\nu = |N_U|/(|N_R| + |N_U|)$. This proves Eq. 15 of Corollary 5, and inflation holds iff $\pi(A_{UR} - A_{PR}) > \nu(A_{PU} - A_{PR})$.

Step 2: existence. Construct: $E \in \{0, 1\}$ with $\mathbb{P}(E = 1) = \frac{1}{2}$; $Y^* \sim$ Bernoulli($\frac{1}{2}$) independent of E; $X = E$; $\rho(E{=}0) = 1$, $\rho(E{=}1) = 0$; $d = 1$; $f(X) = X$. Conditions (i)–(iii) hold. Then $|P| = \frac{n}{2}$, $|N_R| = |N_U| = \frac{n}{4}$, so $\pi = \frac{1}{3}$, $\nu = \frac{1}{2}$. With $f = E$: $A_{PR} = \mathbb{P}(E_i = 1) \cdot 1 + \mathbb{P}(E_i = 0) \cdot \frac{1}{2} = \frac{3}{4}$ (positives are half $E{=}1$, half $E{=}0$; reported negatives all have $E = 0$); $A_{UR} = 1$ (unreported negatives all have $E = 1$); $A_{PU} = \frac{1}{4}$. Hence $\text{AUROC}(f, Y) = \frac{2}{3} \cdot \frac{3}{4} + \frac{1}{3} \cdot 1 = \frac{5}{6} > \frac{1}{2} = \frac{1}{2} \cdot \frac{3}{4} + \frac{1}{2} \cdot \frac{1}{4} = \text{AUROC}(f, Y^*)$. □

Step 3: necessity of (i)–(iii). If (i) fails ($\rho \equiv 1$), then $N_U = \emptyset$, $Y = Y^*$, and the two AUROCs coincide. If (ii) or (iii) fails, the score $f(X)$ is independent of the reporting indicator given the true class: by Eqs. 2–3, R depends only on (e, t), so if $f(X) \perp (e, t)$ within the true-negative class, then N_U and N_R have identical score distributions. Then $A_{UR} = \frac{1}{2}$ and $A_{PU} = A_{PR}$, giving $\text{AUROC}(f, Y) = (1 - \pi)A_{PR} + \frac{\pi}{2} \leq A_{PR} = \text{AUROC}(f, Y^*)$ whenever $A_{PR} \geq \frac{1}{2}$, with strict deficit for $A_{PR} > \frac{1}{2}$, $\pi > 0$: deflation, not inflation. Thus each condition is necessary for inflation.

Step 4: insufficiency. The counterexample of Remark 3 ($Y^* = E$, $f = E$, $\rho(E{=}0) = 0.5$) satisfies (i)–(iii) yet yields deflation, so the conditions are not sufficient; the sign is governed by the score distribution of N_U through the inequality of Step 1. This is consistent with Menon et al. (2015) (Corollary 3), who show AUC is invariant under class-conditional corruption: it is precisely the *environment-dependence* of the corruption—which makes the N_U score distribution differ from that of N_R—that creates room for inflation.

Appendix H. Test B Model Sensitivity

Test B's statistic ΔAUROC depends on the model family used for f_X and $f_{X,E}$. On MIMIC-IV, logistic regression yields $\Delta\text{AUROC} = +0.014$ (FAIL), while XGBoost yields $+0.002$: the flexible learner already absorbs unit-correlated signal through the features, shrinking the measured conditional gain. Both model families detect the confounding on the CI registry. We therefore recommend a simple, less flexible base model (logistic regression) as the Test B default—it maximizes the test's sensitivity by preventing f_X from absorbing environment proxies—and recommend reporting the model family alongside ΔAUROC. We also note an interpretive caveat raised by the CI registry: where zero-negative strata exist, environment trivially predicts the observed label and part of Test B's gain is mechanical; on MIMIC-IV, no unit has zero negative labels, so the measured gain is non-trivial conditional signal.

Appendix I. Semi-Synthetic Calibration Study

To characterize the audit's operating properties, we conduct a semi-synthetic calibration study that preserves the real registry's clinical feature distribution $P(X)$ and environment structure E while generating synthetic outcomes Y^* and reporting indicators R with a controllable confounding parameter δ.

I.1. Simulation Design

We fit a logistic model β on pre-2012 data (where label variance exists) to define the clinical data-generating process. For all n patients, we compute clinical scores $s_i = \beta^\top X_i + b$, where b is a shift calibrated so that $P(Y^* = 0) \approx 8\%$ (matching the active-era negative rate).

The reporting function is a mixture:

$$\rho_\delta(E_i) = (1 - \delta) \cdot \rho_0 + \delta \cdot \hat{\rho}(e_i, t_i) \tag{16}$$

where ρ_0 is the baseline reporting rate (estimated from active strata) and $\hat{\rho}(e_i, t_i)$ is the real registry's estimated reporting function. At $\delta = 0$, reporting is uniform (no environment confounding); at $\delta = 1$, the simulation recovers the real registry's measurement process.

For each $\delta \in \{0, 0.1, \ldots, 1.0\}$, we generate 50 replicates of (Y^*, R, Y^{obs}) and run the full audit battery on Y^{obs}.

I.2. Results

Operating characteristic (Figure 1). At $\delta = 0$ (no confounding), the audit produces a FAIL diagnostic in 4% of replicates (false positive rate). Detection power increases monotonically with δ: P(FAIL) reaches 56% by $\delta = 0.8$ and 88% by $\delta = 1$. Importantly, even at moderate confounding ($\delta = 0.5$), P(PASS) falls to 6%, meaning the audit rarely gives a clean bill of health when meaningful confounding is present.

AUROC under stochastic reporting (Figure 2). Under the stochastic simulation, the gap between AUROC(Y^{obs}) and AUROC(Y^*) is consistently negative (mean ≈ -0.03 to -0.05): the operational AUROC is *lower* than the oracle AUROC. This is expected and reflects a key asymmetry: the stochastic $d = 1$ default converts some true negatives into positives uniformly at random, adding label noise that *reduces* discriminability. This

contrasts with the real registry, where the default operates *deterministically*—all patients in post-2012 eras receive $Y = 1$, creating a spurious feature–label correlation that *inflates* apparent performance. The stochastic simulation does not reproduce the deterministic collapse, and therefore the inflation direction differs.

Crucially, this does not undermine the calibration study's primary purpose: validating the *audit's detection power*. The audit correctly flags the environment–label association introduced by $\rho_\delta(E)$, regardless of whether that association inflates or deflates AUROC. The monotonic increase in P(FAIL) with δ confirms that the audit is sensitive to the confounding mechanism itself, not merely to its downstream effect on one performance metric.

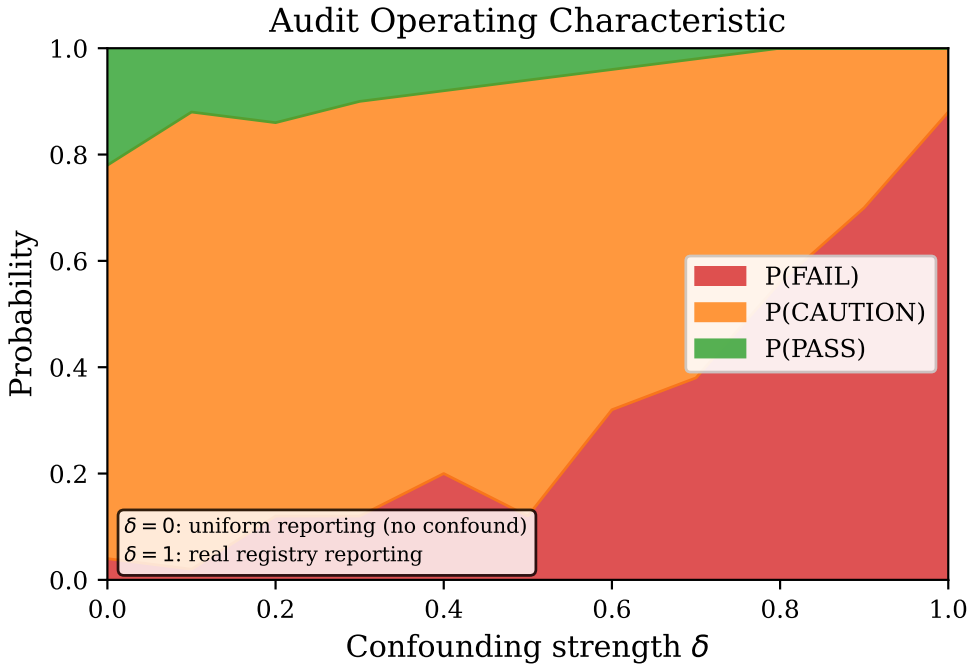

Figure 1: Audit operating characteristic. Stacked areas show $P(\text{FAIL})$, $P(\text{CAUTION})$, and $P(\text{PASS})$ as a function of confounding strength δ. The false-positive rate at $\delta = 0$ is 4%; detection power reaches 88% at $\delta = 1$.

Sensitivity to thresholds. We note that the dual-criterion design (p-value *and* effect-size floor) is the primary source of robustness: the p-value controls the false-positive rate while the effect-size floor prevents flagging of small, clinically irrelevant associations. At $\delta = 0$, P(PASS) = 22% and P(CAUTION) = 74%; the 4% false positive rate indicates that spurious FAILs are rare even without confounding. Exploring higher numbers of bootstrap or permutation replicates could further tighten these estimates.

Family-wise error of the aggregated verdict. To validate the Fisher-calibrated aggregation rule (Section 4.4) beyond the semi-synthetic model, we permute labels on MIMIC-IV ($\delta = 0$; $N = 100$ replicates) and recompute the full battery and overall verdict on each replicate. The empirical family-wise error rate of the overall FAIL verdict is 0.000. Varying the

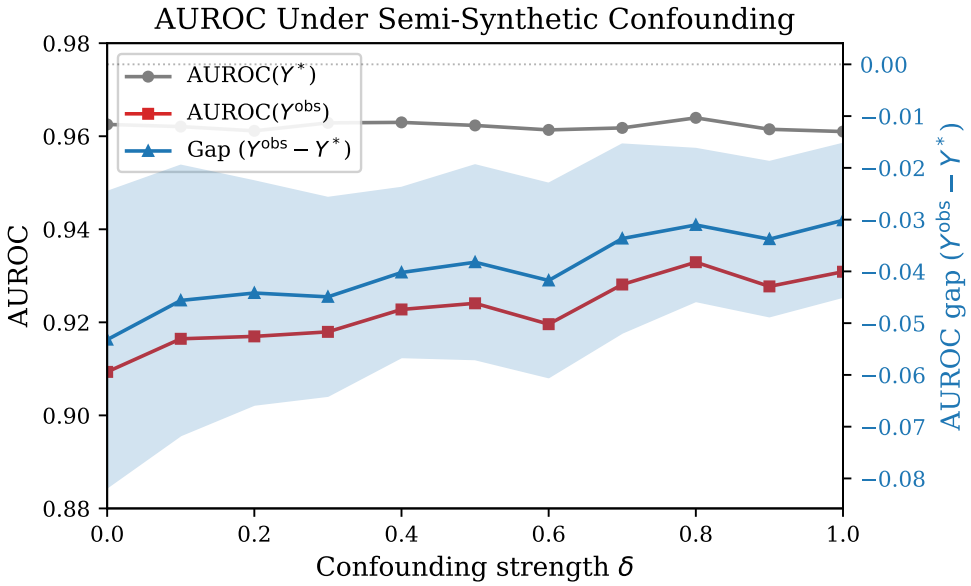

Figure 2: AUROC under semi-synthetic confounding. Gray: AUROC(Y^*) (true labels, ≈ 0.962). Red: AUROC(Y^{obs}) (operational labels). Blue ribbon: gap ± 1 SD across 50 replicates. The operational AUROC is *lower* than the oracle due to the stochastic $d = 1$ default adding label noise (see text).

Cramér's V cutoff from 0.05 to 0.20 leaves the FAIL verdicts on both real registries unchanged (observed $V = 0.390$ and 0.145 exceed all cutoffs), with empirical family-wise error ≤ 0.02 throughout.

Appendix J. Bootstrap Analysis Details

We use a cluster bootstrap with clinician-level clusters to respect the dependence structure of the data. The registry contains 10 clinician identifiers: the 8 with $n \geq 100$ patients define the primary environments used in the audit and the blocked evaluations, and the remaining 2 low-volume identifiers are retained as additional clusters here, yielding 10 bootstrap clusters in total. In each of 1,000 replicates, we sample clinicians with replacement, include all patients from each sampled clinician, and relabel duplicated clinician clusters to avoid identifier collisions. We then compute the full attribution hierarchy (IID, era-blocked, clinician-blocked, residualized) and run the audit battery on each bootstrap sample. Of 1,000 replicates, 989 produced valid results (11 excluded due to degenerate label distributions in resampled data).

The bootstrap serves two purposes: (i) providing 95% CIs for the attribution quantities in Table 3, and (ii) assessing audit stability—the fraction of bootstrap replicates in which the overall diagnostic agrees with the point estimate.

Audit stability. Across 1,000 bootstrap replicates, the audit produces a FAIL diagnostic in 81% of samples, CAUTION in 19%, and PASS in 0%, confirming that the FAIL conclusion

is robust to resampling variability. (PASS, CAUTION, and FAIL here denote the overall verdict of the decision rule of Section 4.4, recomputed in full on each bootstrap replicate.) The wide CI on the era-blocked AUROC ($[0.668, 0.925]$) reflects the small number of clinician clusters (10), underscoring that uncertainty in environment-blocked evaluations is inherently higher than in IID evaluations. Figure 3 visualizes the full bootstrap distributions for the four evaluation regimes and the three attribution quantities.

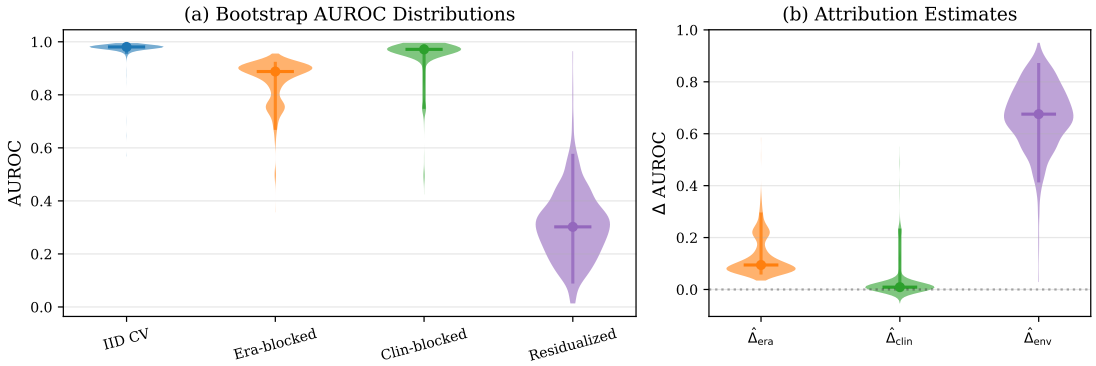

Figure 3: Bootstrap distributions for attribution quantities. (a) AUROC under four evaluation regimes; vertical bars show 95% CIs. (b) Attribution estimates $\hat{\Delta}_{\text{era}}$, $\hat{\Delta}_{\text{clin}}$, $\hat{\Delta}_{\text{env}}$.