Clinically-Ready Label-Flip Detection for Medical AI

Daniel Schönle [1], Christoph Reich [2]

[1] Affiliation
daniel.schoenle@hs-furtwangen.de
[2] Affiliation
christoph.reich@hs-furtwangen.de

Abstract. Medical AI pipelines face integrity risks from label flipping—mislabeling that harms thresholds, calibration, and parity. Because anomalies are rare, evolving, and often mislabeled, a purely supervised detector tends to miss new problems and flood reviewers with false alarms; a triage loop—rank strong model-vs-label disagreements, review a small top slice, fix, retrain—keeps effort low and results trustworthy. We present a lightweight procedure: basic plausibility/duplicate checks; leakage-safe K-fold cross-fitting; calibration; and Confident Learning to derive per-example flip scores (and the confident joint). High-scoring cases receive budgeted chart-review; we then selectively relabel or reweight, retrain, and recalibrate. We evaluate flip-ranking (PR-AUC, precision@k, TPR@low-FPR) and downstream AUROC/PR-AUC, ECE/Brier, and parity deltas. A HiRID ICU case demonstrates integrity and calibration gains with limited review effort.

Keywords: : label noise; confident learning; calibration; medical AI; fairness; governance; anomaly detection

1 Introduction

We target label flipping in EHR datasets used for model development & validation. Adversaries or process faults can inject or flip labels at data entry, curation, or during merges. Security goals are to maintain discrimination, calibration, & subgroup parity, & to provide auditable evidence trails (provenance, adjudication notes, versioned lineage). EHR-specific flip vectors include ICD miscodes & mapping drift, rule-based phenotype errors, timestamp/encounter misalignment, & contradictions between billing codes & clinical evidence (vitals/labs/notes).

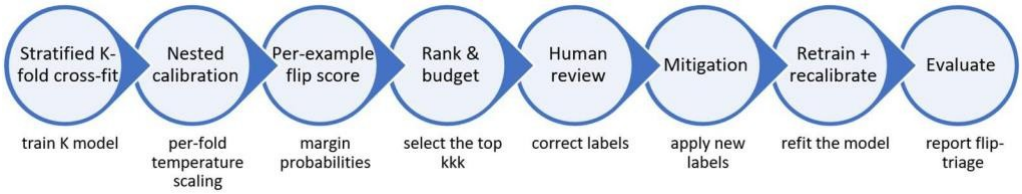

Figure 1. Eight-step ML-only triage loop for label-flip detection.

Blueprint: (1) Ingress rules check provenance, duplicates, mutual exclusivity, value ranges, unit harmonization (e.g., mg/dL↔mmol/L), physiologic plausibility, & encounter/windowing consistency; (2) Disagreement analytics from targeted chart-review & code–evidence conflict checks; (3) Leakage-safe K-fold cross-fit to obtain out-of-fold probabilities; (4) Calibration on a nested holdout (temperature by default; Dirichlet optional under imbalance); (5) Confident Learning to compute the confident joint (CJ) & estimate T; (6) Rank suspected flips where label & calibrated predictions disagree w/ high confidence; (7) Mitigate via budgeted relabeling,

reweighting, or relabel/weight w/ T; (8) Retrain, recalibrate, & log decisions. We use gradient-boosted trees (LightGBM/XGBoost) on benchmark features.

2 Evaluation Protocol

Flip ranking: PR-AUC, ROC-AUC, TPR@1% FPR, precision@k under review budgets, & lift over random; report bootstrap confidence intervals. Downstream (before/after mitigation): AUROC/PR-AUC, ECE & Brier, reliability diagrams, & subgroup parity deltas (e.g., ΔFNR). Subgroup analyses use ICU-relevant strata (sex, age bands, unit type, SOFA/APACHE bins, admission year) w/ temporal splits to approximate deployment drift; top-k suspects can be chart-reviewed against notes/labs when available.

3 Datasets & Experiment

Detectors only triage candidates; labels change only after expert adjudication; the system does not make clinical decisions. HiRID
HiRID access requires PhysioNet credentialing & a data-use agreement; data are de-identified. Limitations include site/process shifts across years/units & potential bias in Tunder covariate shift; subgroup-specific T can reduce bias at the cost of variance.

Primary dataset: HiRID (high-resolution ICU EHR; PhysioNet). We use the published HiRID-ICU-Benchmark pipeline to materialize features & tasks, minimizing engineering effort. Tasks: (i) Flip detection on a noisy, code-derived phenotype (e.g., pneumonia/CHF) constructed w/ the benchmark's labeling rules; (ii) Downstream evaluation on a stable outcome (in-hospital mortality or intervention onset). Pipeline: benchmark features → LightGBM/XGBoost w/ 5-fold cross-fit → temp-scaling (Dirichlet optional) → Confident Learning (CJ, T) → budgeted relabeling (top 1–2%) or relabel/weight → retrain + recalibrate.
Replication dataset: eICU Collaborative Research Database (eICU-CRD) can be processed w/ the same pipeline to validate cross-site robustness.

References

1. Northcutt, C. G., Jiang, L., & Chuang, I. (2021). Confident Learning: Estimating uncertainty in dataset labels. Journal of Artificial Intelligence Research.
2. Patrini, L., Rozza, A., Menon, A., Nock, R., & Qu, L. (2017). Making deep neural networks robust to label noise: A loss correction approach. CVPR.
3. Hyland, S. L., Faltys, M., Hüsser, M., et al. (2020). HiRID: A high-time-resolution ICU dataset for research. Scientific Data.
4. Goldberger, A. L., et al. (2000). PhysioBank, PhysioToolkit, and PhysioNet. Circulation.
5. HiRID-ICU-Benchmark (GitHub repository and accompanying paper): Reproducible tasks and pipelines for HiRID.