ML Lab Research Proposal

Team: [Team Letter & Names]
Date: [Submission Date]
Working title: Pick Two — A Fairness–Identifiability Trilemma for Self-Recalibrating Clinical Triage AI

---

1. Problem Statement

Clinical early-warning models (sepsis, acute kidney injury, deterioration) are trained at one hospital and deployed at others, where they wrap a "defer to clinician" abstention layer and recalibrate themselves online from the cases they process. We show this self-updating loop systematically erodes coverage for the vulnerable demographic group: uncertainty-greedy deferral removes the hardest minority cases, so the model's self-estimate of the deployment patient mix is biased, which worsens its per-group thresholds over time. This matters because it is a silent, self-reinforcing equity failure in deployed systems that current per-group / weighted / online conformal methods neither detect nor prevent.

2. Proposed Approach

We frame the failure as a **trilemma** — {defer optimally by uncertainty, recalibrate from retained cases only, keep per-group coverage} cannot all hold — extending the classic "pick two" fairness-impossibility tradition (Kleinberg–Mullainathan–Raghavan; Chouldechova) to the online selective-prediction setting. Our three theoretical contributions: (a) deferral-rate **parity ⟺ unbiased self-recalibration** — a fairness criterion is exactly an estimation condition, a new bridge between abstention-fairness and identifiability; (b) a **retained-only impossibility** — no recalibration from kept cases alone can bound the per-group gap, so the deferred cases' covariates are provably necessary; (c) a **Θ(Δ_r) frontier** quantifying coverage error as a function of the deferral-rate disparity Δ_r, with matching upper and lower bounds. The deployable system is a post-hoc wrapper on a frozen base model: RKHS mixture estimator + density-ratio weighted per-group conformal calibration + robust worst-case handling of missing demographics + a debiased (all-arrival) online update. We chose post-hoc / model-agnostic over retraining so it deploys on existing early-warning systems without vendor cooperation, and conformal over Bayesian UQ for distribution-free finite-sample guarantees.

3. Dataset

Primary (in hand now): a synthetic EHR simulator (`propc/gcs/simulator.py`) with controllable group mixture, group-correlated acuity, within-group covariate shift, and missing-label fraction — used to validate every theorem and gate. **Target (credential-gated):** MIMIC-IV v2.2 (~63k ICU stays, source) → eICU-CRD v2.0 (~200k stays, 208 hospitals, deployment), public via PhysioNet. Preprocessing: FIDDLE-style hourly bins, 34 vitals/labs + missingness indicators, patient-level temporal splits (train/cal/deploy, no patient overlap), sepsis (Sepsis-3) as the primary task. Natural cross-hospital demographic shift is the primary real-data track; synthetic composition resampling is the controlled stressor.

4. Evaluation Metrics

- **Worst-group coverage gap** (primary) = max_k |0.90 − coverage_k|, reported separately for *selective/issued* coverage and *full-group* coverage (never conflated).
- **Deferral-rate disparity** Δ_r = max_k |r_k − r̄| (the bias driver) and per-group deferral burden.
- **Mixture-estimate error** $\|\hat{q} - q\|_1$ (naive retained-only vs parity vs all-arrival) — the trilemma signal.
- **Frontier fit:** correlation of coverage/estimate error with Δ_r (target ≥ 0.95).
- **Set size / efficiency** and worst-group false-negative disparity.

Current synthetic results: marginal vulnerable-group coverage 0.69 → robust 0.90 (= oracle); trilemma mixture-error 0.105 (naive-global) → 0.005 (parity) → 0.004 (all-arrival); frontier corr(Δ_r, error) = 0.997.

5. Compute Requirements

The project runs on the single 192 GB-VRAM node and uses the GPU where it pays off. **GPU-accelerated:** (a) base-model training — GRU backbone(s) plus GPU-hist XGBoost across 3 tasks, ~24–40 GPU-hours; (b) random-Fourier-feature embeddings and frozen-model inference over the full eICU cohort (~200k stays × D≈1–4k features) precomputed once on GPU and reused across the sweep, ~8–12 GPU-hours; (c) the main parameter sweep (≈4,000 config × seed × task runs) parallelized across 2–3 GPUs of the node, with embeddings/inference batched on-device. **CPU-bound by nature (node's 64 cores):** the simplex QP mixture estimator, conformal quantile computation, and the inherently sequential online recalibration loop — these do not benefit from GPU and are not forced onto it. **Totals:** ~40–50 GPU-hours and ~700–800 CPU-hours; peak VRAM < 80 GB (fits one A100; multi-GPU used only for sweep parallelism, never exceeding the 192 GB ceiling); ~73 GB scratch for derived features. Cloud-equivalent ≈ $450 (~$400 GPU spot + ~$40 CPU); negligible on owned hardware. Dependencies: numpy / scipy / scikit-learn (built), plus OSQP / MAPIE, PyTorch (GRU + GPU embeddings), and FIDDLE / pyhealth for extraction.

6. Expected Deliverables (two weeks)

1. **Theory note + proofs** (Thms 1–7: bias, feedback, impossibility, the parity equivalence, the frontier) — done, with a hard-asserting numerical validation (8 checks, exit 0).
2. **Open-source `gcs-conformal` package** — Modules 0–3 + weighted CP + robust group handling + baselines; runs on synthetic now, plug-and-run for real data.
3. **The trilemma demonstration** (naive-global biased vs parity vs all-arrival, multi-seed).
4. **Must-beat baseline suite B8–B15** implemented and run on synthetic.
5. **Short paper draft** (FAccT-style: fairness-meets-identifiability headline) + the operational-audit tool.

Stretch (only if PhysioNet credentials land in time): first real MIMIC→eICU H1/H2 numbers.

7. Risk & Mitigation

- **Data-access timeline.** PhysioNet credentialing can take several weeks. Mitigation: the two-week deliverable (theory, synthetic validation, baselines, audit tool) is fully self-contained and stands on its own; the real-data run is scoped as an extension, and credentialing begins on day one so it proceeds in parallel.

- **Strong within-group shift.** Under the most extreme within-group covariate shifts, the weighted-conformal layer may not fully reach nominal per-group coverage. Mitigation: it already closes most of the gap toward the weighted-oracle ceiling, and we will add the exact conditional (Gibbs–Cherian–Candès) calibration to tighten coverage to nominal where needed.
- **Active research area.** Related work on fairness under distribution shift is moving quickly. Mitigation: our trilemma characterization and the coverage–disparity frontier are clearly differentiated contributions; we prioritize a fast first public release and target FAccT, where the fairness↔identifiability bridge is the strongest fit.
- **Missing-not-at-random demographics.** Real demographic missingness may violate the MAR assumption. Mitigation: the robust worst-case completion (already implemented) degrades gracefully and supports partial-identification bounds rather than relying on a single imputation.