# Integrating Machine Learning for Pharmacovigilance Adverse Event Signal Detection compared to Classical Disproportionality

Natanael Alpay

Department of Mathematics, University of California, Irvine

## Objectives

Quantify when machine-learning classifiers provide earlier and more precise adverse-event signal detection than classical disproportionality on public FAERS, using temporal holdout and reviewer-relevant metrics, and deliver practical guidance for safety triage.

## Introduction

**Signal detection** in pharmacovigilance aims to identify drug–adverse event pairs that occur more often than expected in spontaneous reports. Classical methods rely on disproportionality statistics computed from aggregated counts. However, modern reporting databases contain additional structured context (e.g., seriousness, reporter type, concomitant drug burden) that may help *prioritize* likely true risks.

## Outline/Method

1. **Construct 2 × 2 counts** For each drug $d$, adverse event $e$ (MedDRA PT), and time (quater) $t$, we form the $2 \times 2$ table of FAERS reports:

$$\begin{array}{c|cc} & e & \neg e \\ \hline d & a & b \\ \neg d & c & d \end{array} \qquad N = a + b + c + d.$$

2. **Classical scores:** compute PRR, ROR, IC, and EBGM-proxy from $(a, b, c, d)$.
3. **Feature sets:**
   - Classical: PRR, ROR, IC, EBGM-proxy
   - Report-quality: seriousness share, mean concomitant drugs, reporter type share
   - Temporal (tested): $a_{t-1}$ and burst indicator (removed in the final “SAFE” model)
4. **ML model:** Gradient Boosting trained to predict if a drug–event pair is a known positive.
5. **Calibration:** isotonic calibration on training data; use calibrated probabilities for ranking.
6. **Evaluation:** temporal holdout (future quarters) + rolling 4-quarter windows; report AUROC/AUPRC, and simple FDR-style cutoffs.

## Performance Stability Across Time

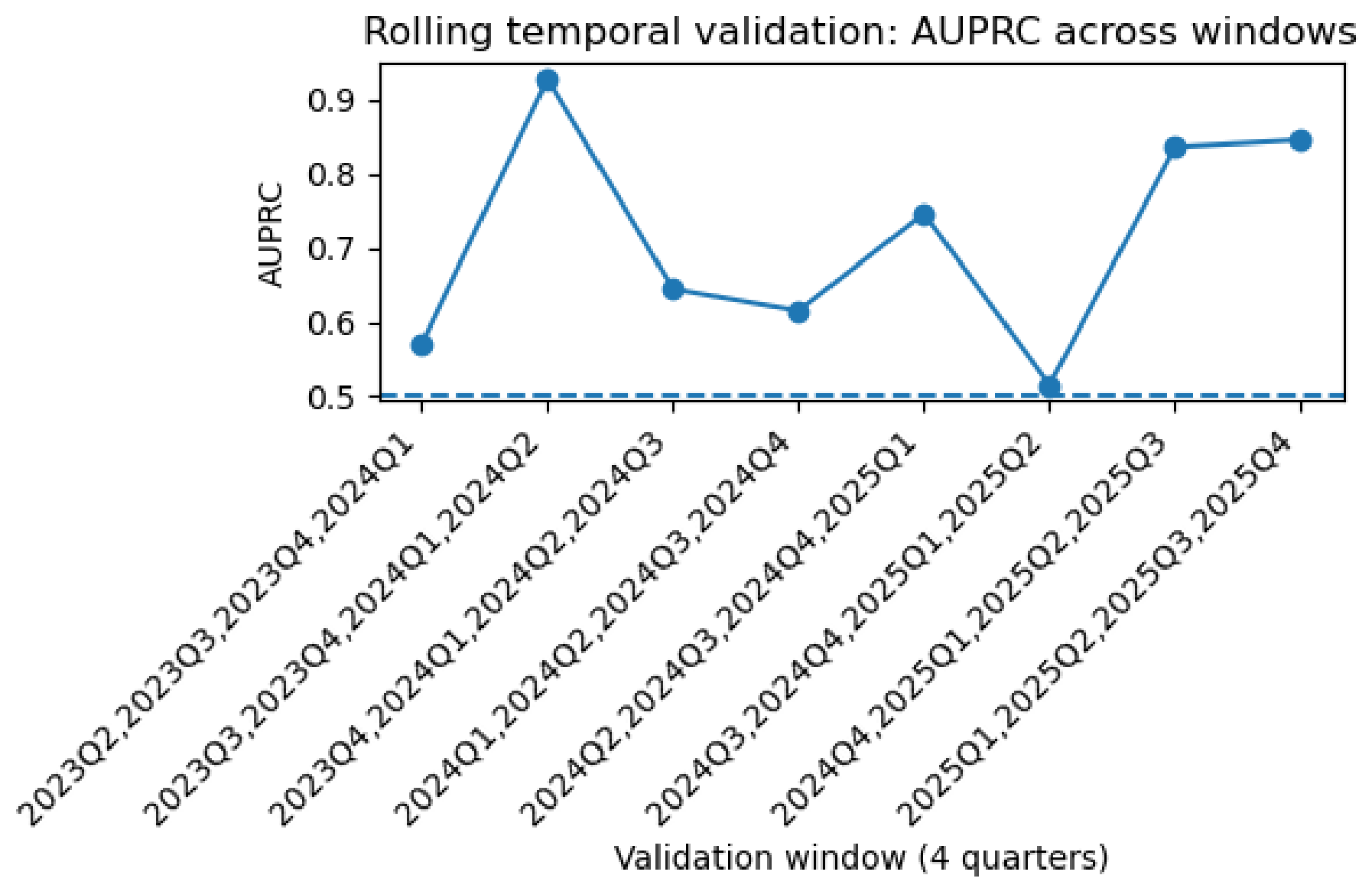

Figure 1: Rolling temporal validation (AUPRC). Each point shows AUPRC when training on all quarters prior to a 4-quarter validation window and evaluating on that future window. Performance varies across time but remains above baseline, indicating robustness under time-respecting evaluation.

## Feature Groups Contribution

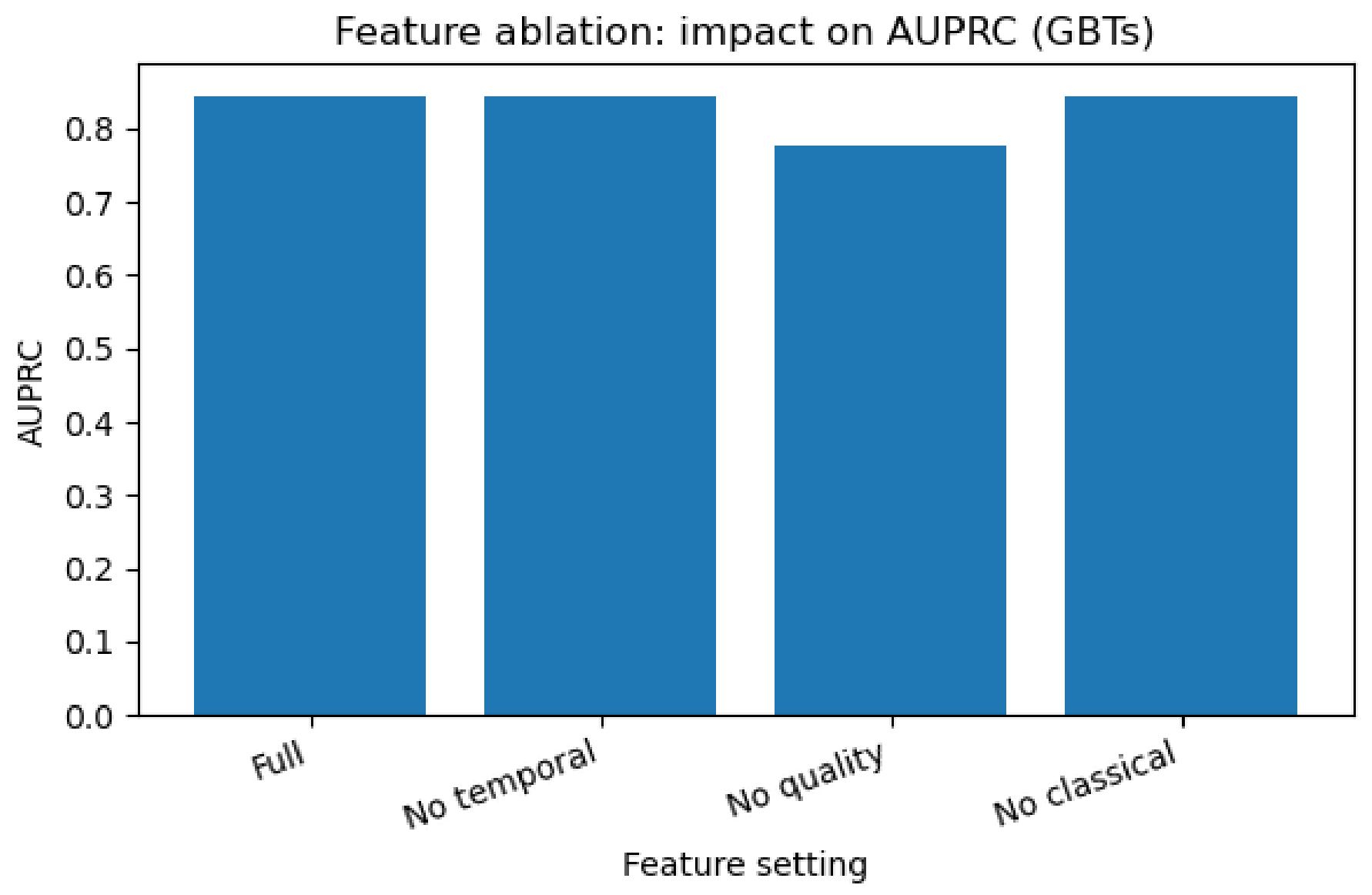

Figure 2: Feature ablation study (GBTs). AUPRC under different feature group removals: Full, No temporal (drop burst/$a_{\text{prev}}$), No quality (drop seriousness/reporting proxies), and No classical (drop PRR/ROR/IC/EBGM-proxy). This isolates which information contribute most to signal ranking.

## Important Result

**Main finding:** performance improves substantially after removing unstable temporal features.

- “Full” feature set: GBT AUPRC ≈ 0.68 (near baseline in this pilot).
- “SAFE” model (drop: `burst`, `a_prev`): AUPRC ≈ 0.85 on a temporal holdout.
- Rolling 4-quarter validation: mean AUPRC ≈ 0.71 (std ≈ 0.15), showing time-window variability.

**Takeaway:** ML can add value for prioritization, but temporal features require careful design to avoid instability/leakage.

## Mathematical Formulation

**Supervised ML:** For each $(d, e, t)$ we build a feature vector $x_{d,e,t} \in \mathbb{R}^p$ from counts $(a, b, c, d)$, classical scores (PRR, ROR, IC, EBGM-proxy), and report-quality features. We predict label $y \in \{0, 1\}$ indicating whether $(d, e)$ is a known positive control.

**Gradient boosting (nonlinear ML).**

$$F_M(x) = \sum_{m=1}^{M} \nu \, f_m(x), \qquad \hat{p}(y = 1 \mid x) = \sigma(F_M(x)),$$

where each $f_m$ is a shallow decision tree fit sequentially to reduce the loss, and $\nu$ is the learning rate.

**Training.** We train on earlier quarters and validate on later quarters, reporting AUROC/AUPRC, rolling-window stability, and operational metrics.

## Quality Under Class Imbalance

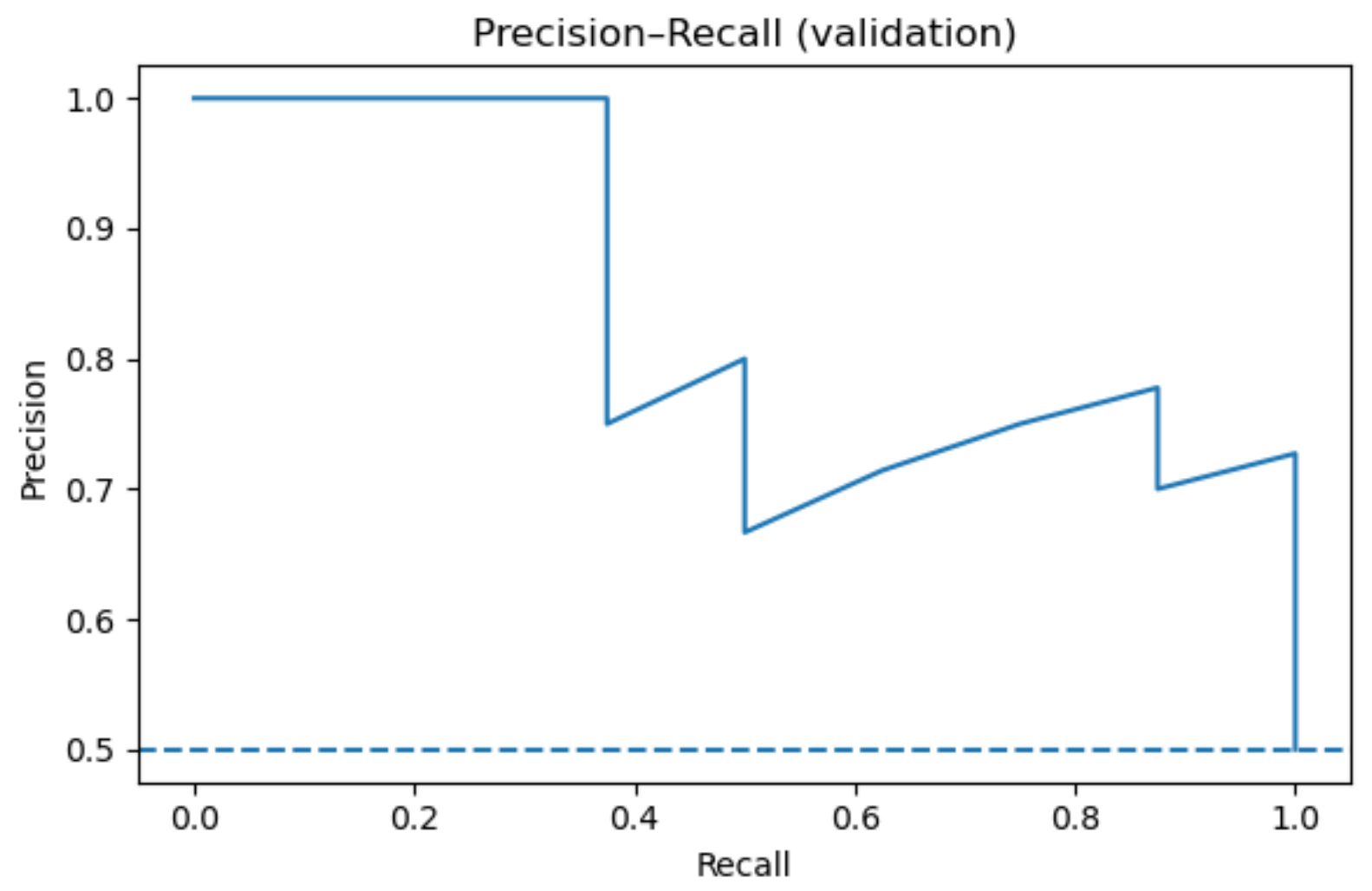

Figure 3: Precision–Recall curve on temporal holdout. PR curves summarize ranking performance under class imbalance. The dashed line indicates baseline precision. Higher area under the curve reflects improved prioritization of true drug–event signals.

## Reliability of Predicted Risk

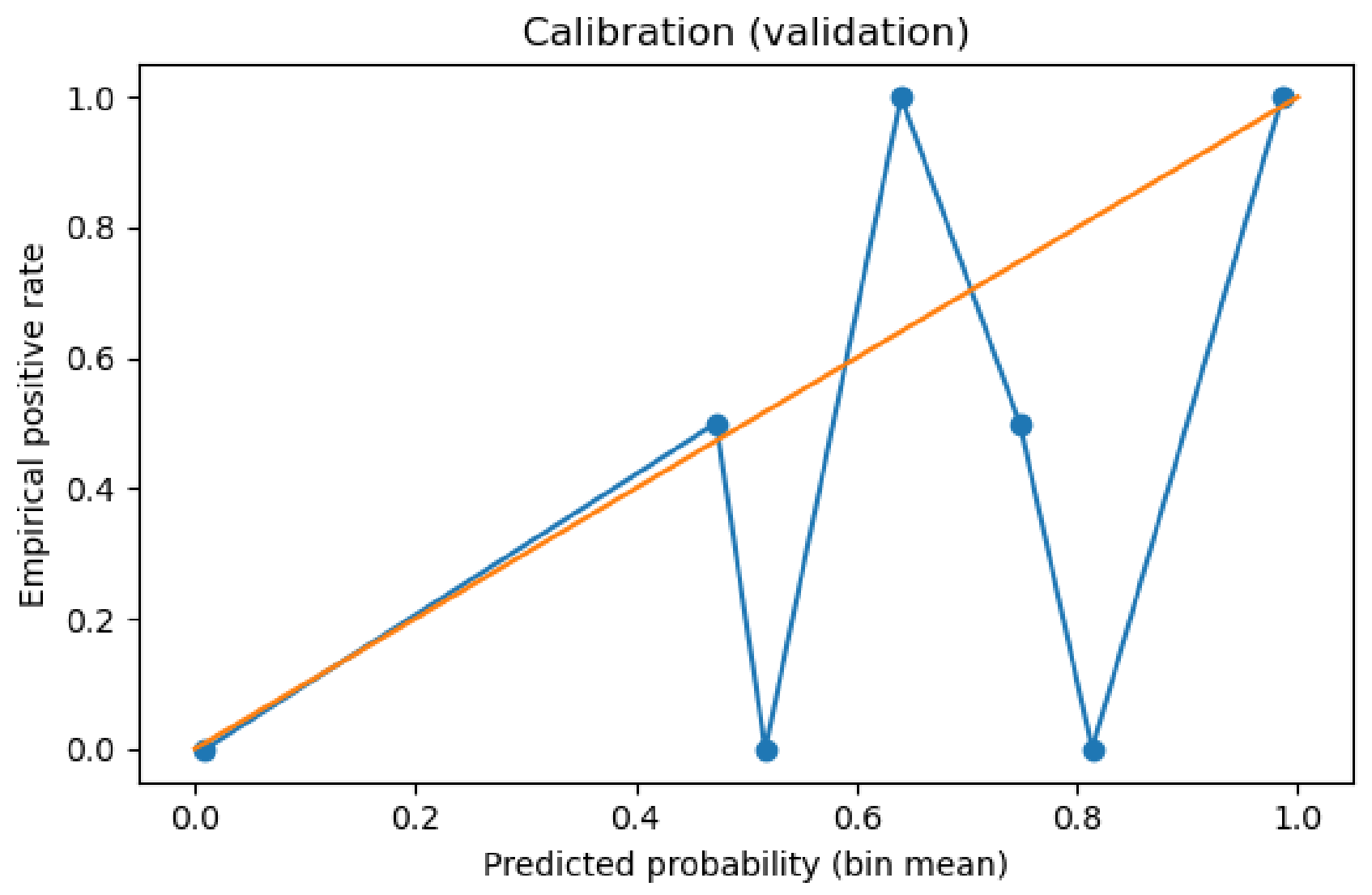

Figure 4: shows a calibration (reliability) diagram comparing predicted probabilities to observed event rates. The diagonal line represents perfect calibration, while deviations reflect over- or under-confidence; despite noise from limited sample sizes, probabilities are sufficiently calibrated to support threshold-based triage (e.g., FDR control).

## Conclusion

- Gradient Boosting models can improve **ranking/triage** of potential safety signals compared to using a single disproportionality.
- Temporal validation reveals **window-to-window variability**; stability improves with careful feature design.
- Next steps: expand labeled reference sets, evaluate pure-classical baselines (PRR/ROR/IC/EBGM as rankers), and redesign temporal features to avoid leakage.

## Contact Information

- Web: alpaynatanael.github.io/index.html
- Email: nalpay@uci.edu
- Phone: +1 (714) 710 098

UCI School of Physical Sciences