SUPPLEMENTARY MATERIAL

Ensemble-Based Label Adjudication (EBLA) to Resolve Discordant Observations in Clinical Monitoring

Guillermo Gutierrez, MD, PhD[1] and Fernando Papa, BS[2]

1 - The George Washington University, Washington, DC, USA.

2- Hitss Argentina, Buenos Aires, Argentina.

Corresponding author:

Guillermo Gutierrez, MD, PhD

700 New Hampshire Ave, NW

Suite1107

Washington, DC 20037

gutier@gwu.edu

Section 1: EBLA Method Description

Problem Definition

Label noise arising from imperfect, observer-dependent reference standards is common in clinical research but is seldom addressed explicitly. Conventional diagnostic performance metrics assume error-free labels; when the reference is noisy, this assumption introduces bias. Existing label-correction approaches typically rely on a single noisy label source and do not exploit the structure available when two imperfect instruments are compared.

Conceptual Framework

Ensemble-based label adjudication (EBLA) leverages agreement between two independently noisy instruments. Paired observations are partitioned into a confusion matrix using predefined thresholds for each method. Observations are classified as:

- **Concordant** (true positives and true negatives): both instruments agree
- **Discordant** (putative false positives and false negatives): instruments disagree

Concordant observations are treated as a **high-confidence subset**, as both instruments would have to err simultaneously and independently for misclassification. For example, if each method has an error rate of 15%, the expected error rate in the concordant subset is approximately 2.25%, compared with 27.75% in the full dataset.

Figure 1 (main manuscript). Distribution of paired observations into concordant and discordant regions defined by method-specific thresholds.

Adjudication Procedure

The EBLA workflow consists of:

1. **Identification of concordant observations**
2. **Training of a classifier committee** on the concordant subset
3. **Adjudication of discordant observations** using majority voting

A committee of five diverse classifiers: Random Forest (100 trees), XGBoost, Logistic Regression (max iterations 1,000), Support Vector Machine (RBF kernel), and k-Nearest Neighbors (k = 5) is trained on the concordant data using stratified 5-fold cross-validation. Discordant observations are then classified by committee vote using predefined thresholds (e.g., ≥3/5, ≥4/5, or 5/5 agreement). Observations not meeting the selected threshold are excluded as ambiguous.

This approach yields a corrected classification of discordant cases for the new method while leaving the original concordant observations unchanged.

Section 2: Validation Summary

Simulation Framework: EBLA was evaluated using the Wisconsin Breast Cancer Dataset (n = 569; Breast Cancer Wisconsin (Diagnostic) - UCI Machine Learning Repository), which provides pathology-confirmed ground truth. The dataset contains 30 continuous features derived from digitized fine-needle aspiration images (e.g., radius, texture, perimeter, area, smoothness). Python program for EBLA analysis of the Wisconsin Breast Cancer Dataset: **https://doi.org/10.6084/m9.figshare.31988589**

Two independent noisy label sets were generated by randomly introducing classification errors to simulate imperfect instruments and five error-rate scenarios were examined:
(10%, 10%), (15%, 15%), (20%, 15%), (20%, 20%), and (30%, 20%)

Key Results: Across all scenarios, EBLA consistently improved classification performance relative to the noisy input labels:

- **Accuracy improvement:** approximately 9% to 14%
- **Sensitivity and specificity:** both improved, with the largest gains observed in specificity
- **Data retention:** >98% of observations retained at the supermajority threshold (≥4/5)

The supermajority voting threshold provided the best balance between accuracy and retention.

Interpretation: Performance gains reflect the enrichment of label quality in the concordant subset and reduced exposure to label noise during model training. Improvements were robust across both symmetric and asymmetric error conditions.

Comparative analysis with probabilistic label-cleaning approaches (e.g., Cleanlab) is provided in the main manuscript. Cleanlab comparison Python program: **https://doi.org/10.6084/m9.figshare.31988610**

Section 3: Application of EBLA to a clinical dataset (Summary).

EBLA was applied to a clinical dataset comparing a continuous physiological measure of inspiratory effort ($P_{mus}PTP$) with an observer-based dyspnea scale (MV-RDOS) in mechanically ventilated patients.

- **Dataset: 281 paired observations**
- **Concordance: 227 (80.8%; 189TN, 38TP)**
- **Discordance: 54 (19.2%; 27FP, 27FN)**

The 227 concordant instances were used to train the five diverse, independent machine learning classifiers. Features were standardized using parameters (peak airway pressure and its covariance, compliance, driving pressure, respiratory rate, tidal volume respiratory variability index) derived from the concordant set to prevent information leakage.

Using a supermajority threshold (≥4/5):

- **87.2% of discordant cases were adjudicated**
- **97.9% of observations were retained**
- **Accuracy improved from 80.8% to 86.5%**
- **Improvement was driven primarily by increased specificity (87.5% TO 95.2%), with no change in sensitivity (58.5%; 200TN, 38TP, 10FP, 27FN and 6 excluded).**

Interpretive Considerations: Because no independent gold standard exists, EBLA-adjusted results should be interpreted as reflecting the agreement structure between instruments rather than absolute diagnostic performance. Discordant cases may arise from error in either instrument or from physiological mechanisms not captured by one of the measures.

Conclusion: EBLA provides a simple and reproducible framework for addressing discordant observations when validating clinical monitoring tools against imperfect reference standards. By leveraging concordant observations as a high-confidence training set, the method improves classification performance while maintaining transparency regarding its assumptions and limitations.

Python program for EBLA application: **https://doi.org/10.6084/m9.figshare.31988595**

Datasets used in the dyspnea EBLA study**:** **https://doi.org/10.6084/m9.figshare.31988655**