Robust classification of inflammatory subphenotypes using penalized logistic regression on highly available clinical variables

Clove S. Taylor, L. Nelson Sanchez-Pinto, Matt S. Zinter, Daniela Markovic, Daniel Balcarcel, Nadir Yehya, Anoopindar Bhalla, Robert Khemani, Michael Agus, Pratik Sinha, Anil Sapru

UCLA Sapru Lab at UCLA

Background

Inflammatory subphenotypes derived from circulating proteins identify elevated risk of mortality, longer PICU stays, and differential treatment responses across different conditions.

Subphenotypes originally derived in ARDS have replicated this enrichment at baseline in other conditions, but generalization is still limited.

Additionally, subphenotyping is limited by the need for biomarker assays that induce extra cost and can take up to 24-48 hours to collect, placing a significant barrier to rapid personalized clinical decision making.

Our work focuses on developing a classification model of inflammatory subphenotypes based entirely on standardized clinical data available in most patients, with the goal of real-time and zero-cost subphenotype estimation.

Aims:

1. *Accurate prediction of inflammatory subphenotypes across cohorts with different composition;*
2. *Feature selection limited to standardized and highly available clinical variables: zero-cost and real-time.*

Methods

1. Data is collected from each cohort using PRISM-III scores (n=17 features), PELOD-2 scores (n=10 features), and basic data (age, sex, weight);

2. Features are filtered for missingness < 30% and remaining missingness is imputed using median imputation to permit regression without adding bias to imputed features;

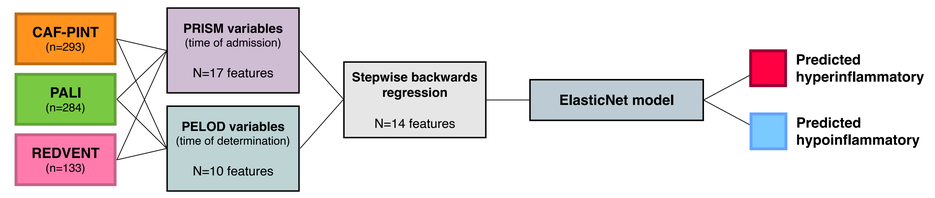
3. Preliminary feature selection is performed using stepwise backwards regression to optimize feature set by P-value (olsrr);

4. ElasticNet regression is performed using leave-one-out crossvalidation to optimize mixing hyperparameters;

5. Model is refitted using all data and including cohort as variable to assess generalizability and provide best estimate of performance;

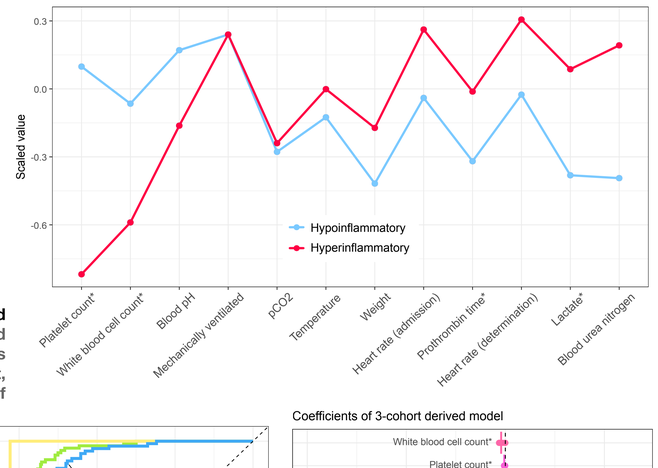
6. Performance metrics are computed using ROC curves, and coefficients are measured as odds ratios including best fit (vertical bars) and full distribution across gamma regularization (points)

A **ElasticNet model architecture.**



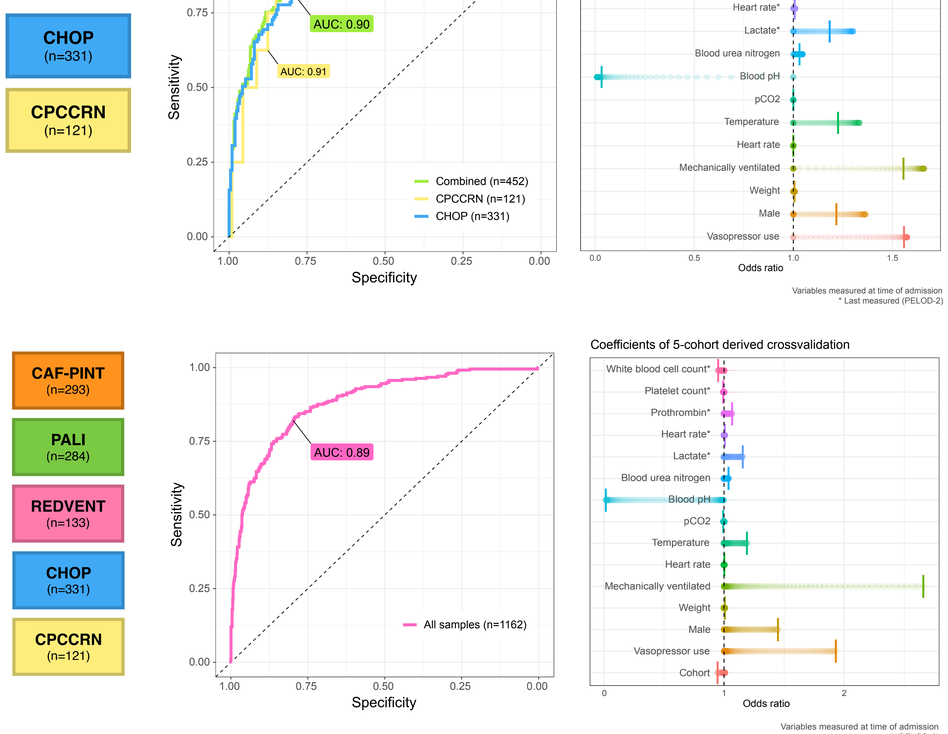
B

Continuous variables passing feature selection. Values are normalized per-cohort.



C

Results of validation on CPCCRN and CHOP cohorts. ROC curves (left) and odds ratios for model coefficients (right). Bars correspond to best fit, points depict different values of gamma.



D **Leave-one-out crossvalidation.** *Left:* ROC curve for all samples. *Right:* Odds ratios for model coefficients. Bars correspond to best fit, points depict different values of gamma. Cohort included to assess generalizability.

Results

Hyperinflammatory patients differ most in platelet & white blood cell counts, acidosis, lactate, BUN, and rates of ventilation and use of vasopressors (A).

2-cohort validation performance on 3-cohort derived model produces good performance with reasonable interpretation (B).

- CHOP and CPCCRN cohorts produce total AUC = 0.90 (CHOP = 0.89, CPCCRN = 0.91) on model derived from CAFPINT, PALI, and REDVENT cohorts
- Predicted hyperinflammatory patients generally display greater acidemia, rates of ventilation and vasopressor use, higher temperature, and higher lactate values

5-cohort crossvalidation performance remains similar, with little indication of batch effects.

- Model derived on all data produces total AUC = 0.89
- Predicted hyperinflammatory patients have similar characteristics to previous model
- Effect size of cohort as variable is nonzero but small relative to other predictors, suggesting good representation of true performance

Conclusion

1. Inflammatory subphenotypes can be classified from standardized clinical data with reasonably strong performance.

- Performance is similar across all cohorts used.

2. Use of standard and highly available clinical data supports value behind preliminary estimation of subphenotypes.

- Predictions are performed on data already collected in most patients.
- Estimated subphenotypes come in real time at no extra cost.

Current work is focused on:

- Expanding datasets to reduce reliance on ARDS patients and improve generalization;
- Additional feature selection of other standardized variables with strong predictive value.

Acknowledgements

We would like to thank the families of the patients enrolled in the CAFPINT, PALI, CPCCRN, REDVENT, and CHOP cohorts, as well as the site investigators of the PALISI network.