Characterizing Distinct Clinical Populations and Evaluating Feature Importance for Predicting Length of Stay in the NICU Using Interpretable Models

An AIM-AHEAD Bridge2AI for Clinical Care Project

Scott McGrath[1], Jennifer Woo[2], Salar Khaleghzadegan[3], Erika Jank[4], Allen J. Yiu[5], Zheng Zhang[6], Brynne Sullivan[7]

[1]University of California, Berkeley [2]University of Texas, Arlington [3]Dartmouth College [4]University of California, Los Angeles [5]Children's National Hospital [6]University of Michigan [7]University of Virginia

OBJECTIVE/PURPOSE

Length of stay (LOS) for preterm infants in the NICU is shaped by gestational maturity, acquired morbidities, and clinical trajectory, yet predictive models in this domain often trade interpretability for accuracy.

This project pursues two linked aims:

- **Characterize distinct clinical phenotypes** within a multi-institutional cohort of preterm NICU infants.
- **Develop an interpretable ML pipeline** for predicting LOS that produces transparent, clinically verifiable feature-importance rankings.

We prioritize a complete, glass-box pipeline over exhaustive optimization, so the team can surface data issues early and judge which refinements matter most.

METHODOLOGY

Data source. CHoRUS AIM-AHEAD-60 dataset in an OMOP CDM environment.

Cohort definition. We intended to define the cohort by gestational age, but no infants had populated gestational-age values, so we pivoted to a birth-weight–based definition (OHDSI ATLAS cohort 1161). Infants enter on their earliest qualifying birth-weight event, with no admission-timing restriction. Requiring both diagnosis and birth-weight data plus a computable LOS yields the analytic cohort.

Features. Demographics (gender, race, ethnicity), unique diagnosis count, unique medication count, and baseline weight (first recorded birth-weight value, kg, as a birth-weight proxy). LOS = admission → discharge. Medication exposure spans 205 formulations across 89 drug classes, used as-extracted this pass; clinically-guided class consolidation is reserved for a later iteration. Mortality is captured to treat in-hospital death as a competing outcome.

Models. We implemented a progressive modeling pipeline across a standardized feature set.

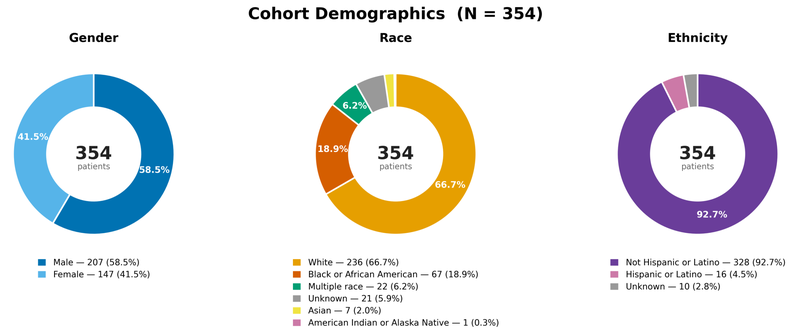
- **GLM**: Served as the interpretable, linear baseline.
- **XGBoost**: Tweedie objective for the strictly positive, right-skewed LOS; histogram trees; early stopping on an 80/10/10 split; feature attributions from **SHAP**.
- **EBM**: purely additive (no pairwise interactions; default settings), read through global and per-term shape graphs.

Models are compared on held-out performance and on the rank order and direction of their feature contributions.

Characteristic	Analytic cohort (N = 354)	Age ≤ 1 year (N = 1,250)
Gender — n (%)		
Male	207 (58.5)	708 (56.6)
Female	147 (41.5)	542 (43.4)
Race — n (%)		
White	236 (66.7)	786 (62.9)
Black or African American	67 (18.9)	178 (14.2)
Multiple race	22 (6.2)	27 (2.2)
Unknown	21 (5.9)	114 (9.1)
Asian	7 (2.0)	33 (2.6)
American Indian or Alaska Native	1 (0.3)	3 (0.2)
Other / not recorded	—	109 (8.7)
Ethnicity — n (%)		
Not Hispanic or Latino	328 (92.7)	1,097 (87.8)
Hispanic or Latino	16 (4.5)	112 (9.0)
Unknown	10 (2.8)	41 (3.3)

Table 1. Baseline characteristics of the analytic cohort (birth-weight–based ATLAS cohort 1161) vs. the broader age ≤ 1 year cohort. 71.2% of the analytic cohort were admitted on the day of birth; none later than 30 days. "Other / not recorded" aggregates small or missing race categories in the comparison cohort.

RESULTS

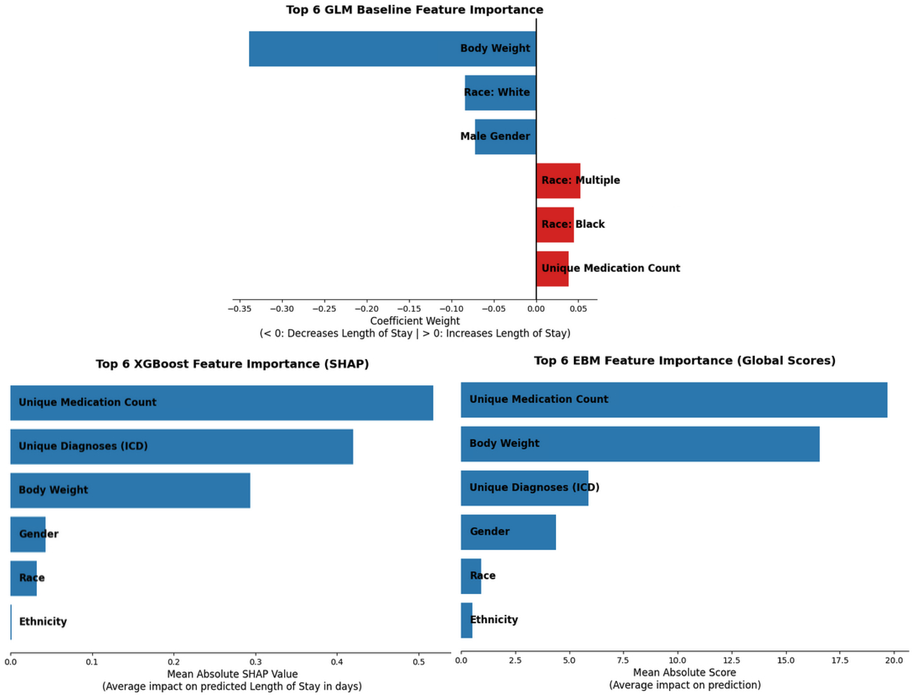


Most frequent diagnoses. Neonatal jaundice of prematurity, respiratory distress syndrome, patent ductus arteriosus, and apnea of the newborn; non-specific newborn-observation codes are the most common coded entries overall.

Comparing performance and feature importance across methods

- **GLM**: MAE 14.84 days | RMSE 26.89 days
- **XGBoost**: MAE 14.21 days | RMSE 23.68 days

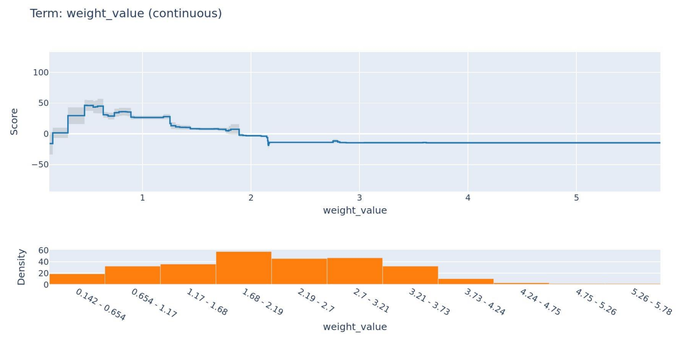
XGBoost reduced RMSE by 3.2 days, demonstrating superior handling of extreme, right-skewed LOS outliers.



Feature Importance compared between models.

Linear vs. Non-Linear. The linear baselines over-indexed on demographic proxies, while non-linear models successfully isolated actual clinical complexity to predict extended hospitalizations. Both EBM and XGBoost recovered the same top three clinical signals (weight, medication count, and diagnosis count) well ahead of demographics.

Clinical Takeaway. Given cohort noise, these three variables should be interpreted collectively as the dominant drivers of length of stay, rather than a strict ordinal ranking.



EBM weight-value shape function. The weight term contributes most at low birth weights and flattens as weight increases; its behavior at the lowest weights is under clinical review as a candidate data artifact.

Interpreting the signals. The medication-count term rises monotonically with LOS, but because a longer stay mechanically accrues more medications, this likely reflects length-of-stay accrual rather than illness severity, motivating a planned medication-frequency (count-per-day) analysis. Surfacing these candidate artifacts early is itself a benefit of the interpretable, glass-box approach.

These estimates are descriptive and not yet validated; finalized model performance and phenotyping results will be presented at the conference.

CONCLUSIONS

- **A clinically coherent NICU cohort and feature set can be assembled from multi-institutional OMOP data** using birth-weight criteria and structured preprocessing.
- **Across an additive glass-box model (EBM) and a black-box model with post-hoc explanations (XGBoost+SHAP), feature-importance rankings converge** on the same clinically plausible drivers.
- **Prioritizing a complete, interpretable pipeline** over exhaustive feature refinement lets the team surface issues early and identify which refinements matter most for the next iteration.

IMPLICATIONS

This work offers a transferable framework for NICU predictive modeling in OMOP environments and clarifies whether structured neonatal data alone can support LOS prediction.

Limitations.

- LOS reflects hospital (not NICU-specific) discharge.
- Gestational age is not reliably available in the dataset.
- Medication labeling varies across sites.

These constraints motivate interpretable models that can separate true clinical signal from documentation artifacts.

Future directions: waveform and clinical-note integration, refined drug-class grouping, and expansion to the forthcoming summer cohort.

ACKNOWLEDGEMENTS & DISCLAIMER

The research reported in this poster was supported by AIM-AHEAD Coordinating Center at the University of North Texas Health Science Center at Fort Worth. This research was, in part, funded by the National Institutes of Health (NIH) Agreement No. 1OT2OD032581. The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the NIH.

This project was supported in part by the NIH Bridge2AI CHoRUS program (NIH Office of Director OT2OD032701).