Traditional Machine Learning Outperforms Automated Machine Learning for Postpartum Readmission Prediction: A Comprehensive Performance and Health-Economic Analysis

Lauren Crabtree[1], Colin Wakefield[2], Ciprian P. Gheorghe[1], Martin G. Frasch[3,4]

[1]Department of Gynecology and Obstetrics, Loma Linda School of Medicine, Loma Linda, CA, USA

[2]Drexel University College of Medicine, Philadelphia, PA, USA

[3]Department of Obstetrics and Gynecology, University of Washington School of Medicine, Seattle, WA, USA

[4]Institute on Human Development and Disability, University of Washington, Seattle, WA, USA

Corresponding Author:

Martin G. Frasch, MD PhD

Department of Obstetrics and Gynecology

Institute on Human Development and Disability

University of Washington School of Medicine

1959 NE Pacific Street, Box 356460, Seattle, WA 98195-6460

mfrasch@uw.edu

Correction Notice (v2): This version corrects performance metrics reported in v1 (medRxiv 2025.12.16.25342440v1). The original submission inadvertently reported results from an analysis pipeline that included variables derived from the readmission outcome (data leakage), inflating discrimination metrics. All results below reflect the verified analysis using 10 sociodemographic features selected by recursive feature elimination, with a 70/30 train-test split and no outcome-derived predictors. The economic analysis has been revised accordingly.

Précis: Traditional logistic regression outperformed AutoML in identifying high-risk patients for postpartum readmission, though all models showed modest discrimination using sociodemographic predictors alone, highlighting the need for clinical feature integration.

Word Count: ∼4,500 (excluding abstract, references, tables) **Tables:** 9 (5 main + 4 supplementary)

Figures: 0 **Funding:** This study received no external funding or sponsorship.

Highlights

1. Traditional logistic regression outperformed AutoML (FLAML) and ensemble methods (stacking, soft voting), uniquely identifying high-risk patients and demonstrating that more complex algorithms do not guarantee clinical utility for rare outcomes.

2. Threshold optimization transformed clinical utility: lowering the logistic regression threshold from 0.50 to 0.35 increased sensitivity from 34.8% to 82.6% (38/46 readmissions captured), enabling use as a screening tool despite modest overall discrimination (ROC-AUC: 0.569).

3. Health-economic value depends critically on intervention cost: conventional enhanced discharge programs are not cost-effective at the model's PPV (2.1%), but low-cost triage strategies (<$59/flagged patient) achieve positive ROI.

Abstract

Objective: Automated machine learning (AutoML) promises to democratize predictive modeling in healthcare by automating algorithm selection and hyperparameter optimization. Our objective was to compare the performance, clinical utility, and health-economic implications of traditional machine learning versus AutoML frameworks for predicting 14-day postpartum readmission using a large national cohort.

Methods: We analyzed data from 8,774 participants in the nuMoM2b (Nulliparous Pregnancy Outcomes Study) with complete readmission data. Recursive feature elimination selected 10 sociodemographic predictors from 18 candidates. We applied SMOTE oversampling to the training set for class balancing. Three traditional ML algorithms (logistic regression, random forest, gradient boosting) were compared with FLAML, a leading AutoML framework. Models were evaluated on discrimination (ROC-AUC, PR-AUC), clinical utility (sensitivity/specificity), and computational efficiency using a 70/30 stratified train-test split with bootstrap confidence intervals.

Results: Among 8,774 participants, 154 (1.8%) experienced 14-day readmission. In a held-out test set (n=2,633; 46 readmissions), logistic regression achieved the highest ROC-AUC (0.569 [95% CI: 0.483–0.655]) and was the only model with clinically meaningful sensitivity (34.8% [20.5–48.6%]) at the default threshold, correctly identifying 16 of 46 readmissions. FLAML achieved near-chance discrimination (ROC-AUC: 0.500 [0.449–0.552]) with near-zero sensitivity (2.2%). Stacking and calibrated soft-voting ensembles did not improve over logistic regression alone (ROC-AUC: 0.496 and 0.555, respectively). Threshold optimization substantially improved screening performance: lowering the logistic regression threshold to 0.35 increased sensitivity to 82.6% (38/46 readmissions), though at the cost of flagging 76.5% of patients. Health-economic analysis showed that cost-effectiveness requires low-cost interventions (<$59/flagged patient).

Conclusions: Traditional logistic regression outperformed AutoML and ensemble methods for postpartum readmission prediction. Threshold optimization, rather than model complexity, provided the largest gains in screening sensitivity. However,

all approaches showed modest discrimination using sociodemographic variables alone, establishing a baseline for future models incorporating clinical features.

Keywords: postpartum readmission, machine learning, AutoML, health economics, maternal health, predictive modeling

1 Introduction

Postpartum readmission represents a significant burden for new mothers, families, and healthcare systems, affecting approximately 1–2% of deliveries and costing an average of $10,000 per episode. [1, 2] Early identification of high-risk patients could enable targeted interventions to prevent readmissions, improve maternal outcomes, and reduce healthcare costs. However, developing effective prediction models requires a careful balance between model performance, clinical interpretability, and implementation feasibility.

The emergence of automated machine learning (AutoML) has promised to democratize predictive modeling in healthcare by automating complex processes of algorithm selection, hyperparameter optimization, and feature engineering. [4, 5] Leading AutoML frameworks such as FLAML and TPOT claim to deliver superior performance while reducing the need for specialized machine learning expertise. This technological promise has generated significant interest in healthcare applications, where AutoML could accelerate the development and deployment of clinical prediction tools.

However, the clinical utility of AutoML in healthcare settings remains largely unproven. Healthcare prediction models face unique challenges, including severe class imbalances, regulatory requirements for interpretability, computational resource constraints, and the critical importance of sensitivity over overall accuracy. [6,7] Traditional machine learning approaches, particularly logistic regression, have demonstrated robust performance in clinical applications due to their interpretability, computational efficiency, and well-understood statistical properties. [8, 9]

The nuMoM2b (Nulliparous Pregnancy Outcomes Study: Monitoring Mothers-to-be) dataset provides one of the largest and most comprehensive maternal health cohorts available for predictive modeling research. [3] This study provides a head-to-head comparison of traditional ML and a leading AutoML framework for postpartum readmission prediction, encompassing predictive performance, computational efficiency, clinical utility, and health-economic implications.

2 Methods

2.1 Study Design and Data Source

This retrospective cohort study utilized data from 8,774 participants in the nuMoM2b study with complete readmission data. The nuMoM2b study is a prospective cohort study of nulliparous women with singleton pregnancies conducted across eight clinical sites in the United States. [3] The primary outcome was 14-day postpartum rehospitalization (CMAJ01), coded as a binary variable.

2.2 Feature Selection

We restricted candidate predictors to sociodemographic and pre-pregnancy variables available at or before delivery discharge, explicitly excluding any variables derived from or encoding the readmission outcome to prevent data leakage. The nuMoM2b codebook contains 11,617 variables across 79 datasets; we classified all variables by temporal relationship to the outcome (Table 7) and restricted predictors to pre-pregnancy and enrollment variables.

From 18 candidate sociodemographic variables, recursive feature elimination (RFE) with logistic regression selected 10 predictors:

- **Demographics:** Maternal race/ethnicity (CRace), educational attainment (Education)
- **Anthropometric:** Body mass index at enrollment (BMI), BMI category (BMI_Cat)
- **Reproductive:** Gravidity category (GravCat)
- **Behavioral:** Smoking status categories (SmokeCat1, SmokeCat2, SmokeCat3)
- **Socioeconomic:** Insurance type (Ins_Pers), percentage of federal poverty level (PctFedPoverty)

2.3 Preprocessing and Class Imbalance Handling

Categorical variables were encoded using label encoding. Missing values were imputed using median imputation. To address the severe class imbalance (1.8% positive rate), we applied Synthetic Minority Oversampling Technique (SMOTE) [10] exclusively to the training set. The test set was left unmodified to reflect real-world prevalence.

2.4 Model Development

Data were split into training (70%) and test (30%) sets using stratified sampling to preserve the class distribution. We evaluated four modeling approaches:

Traditional ML:

- Logistic Regression with L2 regularization and balanced class weights
- Random Forest with balanced class weights
- Gradient Boosting (scikit-learn implementation)

AutoML:

- FLAML (Fast and Lightweight AutoML): Microsoft's automated framework with 300-second time budget [4]

TPOT [5] was initially planned but excluded from the final analysis due to technical compatibility issues during execution.

We additionally evaluated a stacking ensemble combining all three traditional models, using 5-fold cross-validated base model probabilities as inputs to a logistic regression meta-learner with balanced class weights. A calibrated soft-voting ensemble averaging isotonic-calibrated probabilities was also assessed.

2.5 Threshold Optimization for Screening

Because the default classification threshold (0.5) may not be optimal for screening applications where high sensitivity is paramount, we performed a systematic threshold sweep from 0.05 to 0.95. At each threshold, we computed sensitivity, specificity, PPV, and the proportion of the population flagged. We identified operating points targeting ≥80%, ≥70%, and ≥60% sensitivity to characterize the model's utility as a screening tool.

2.6 Model Evaluation

Model performance was assessed on the held-out test set (n=2,633; 46 readmissions) using:

- Area under the ROC curve (ROC-AUC) for overall discrimination
- Area under the precision-recall curve (PR-AUC), more informative under extreme class imbalance [11]
- Sensitivity, specificity, precision, and F1-score at default (0.5) probability threshold
- Brier score for probability calibration
- 95% confidence intervals via bootstrap resampling (1,000 iterations)

2.7 Health-Economic Analysis

We modeled the economic impact of implementing the logistic regression model for risk stratification using a decision-analytic framework (Supplementary Table 8). Key parameters:

- Average rehospitalization cost: $11,250 (range: $7,500–$15,000)
- Intervention effectiveness: 25% (range: 10–50%)
- Annual implementation and maintenance: $65,000

We evaluated cost-effectiveness across a range of intervention costs ($25–$1,000 per flagged patient) to identify the break-even threshold given the model's low positive predictive value.

3 Results

3.1 Study Population

Among 8,774 participants with complete data, 154 (1.8%) experienced 14-day postpartum readmission, consistent with national averages. The test set contained 2,633 participants with 46 readmissions (1.7%).

3.2 Model Performance

Table 1 presents the discriminative performance of all evaluated models. Table 6 provides extended metrics with 95% confidence intervals.

Table 1: Model Performance on Test Set (n=2,633; 46 readmissions)

Model	**Type**	**ROC-AUC**	**PR-AUC**	**Sensitivity**	**Specificity**	**Brier**	**F1**
Logistic Regression	Traditional	**0.569**	**0.023**	**34.8%**	71.7%	0.224	0.040
Gradient Boosting	Traditional	0.502	0.018	0.0%	100.0%	0.030	0.000
FLAML	AutoML	0.500	0.019	2.2%	98.9%	0.029	0.026
Random Forest	Traditional	0.486	0.019	0.0%	99.3%	0.025	0.000

Logistic regression achieved the highest ROC-AUC (0.569 [95% CI: 0.483–0.655]), though discrimination was modest. It was the only model achieving clinically meaningful sensitivity (34.8%), correctly identifying 16 of 46 readmissions in the test set. However, this came at the cost of reduced specificity (71.7%), resulting in 732 false positives among 2,587 non-readmitted patients.

All other models effectively predicted the majority class exclusively. Gradient boosting and random forest achieved 0% sensitivity (zero true positives). FLAML identified only 1 of 46 readmissions (sensitivity 2.2%) despite a 300-second optimization budget.

PR-AUC values were uniformly low ($\leq$0.023), substantially below the prevalence baseline of 1.8%, highlighting the fundamental difficulty of predicting this rare outcome from sociodemographic variables alone.

Brier scores reflected the class imbalance rather than calibration quality. Logistic regression had the highest Brier score (0.224) because it actively predicted positive probabilities for a large subset of patients, while models predicting all-negative achieved deceptively low Brier scores (0.025–0.030) simply by outputting the base rate.

3.3 Ensemble Methods

The stacking ensemble (ROC-AUC: 0.496) and calibrated soft-voting ensemble (ROC-AUC: 0.555) both failed to improve over logistic regression alone. Both ensembles achieved near-zero sensitivity (2.2%) at the default threshold, reflecting domination by the tree-based models that predict all-negative. The meta-learner effectively learned to follow the majority vote of base models, suppressing the minority-class signal from logistic regression.

3.4 Threshold Optimization for Screening

Because logistic regression was the only model with discriminative signal, we performed threshold optimization exclusively on its predicted probabilities. Table 2 presents the sensitivity–specificity trade-off at clinically relevant operating points.

Table 2: Logistic Regression Screening Operating Points by Threshold

Threshold	Sensitivity	Specificity	PPV	NPV	TP	FP	% Flagged	Use Case
0.25	95.7%	2.1%	1.7%	96.5%	44	2,532	97.8%	Maximum capture
0.35	82.6%	23.7%	1.9%	98.7%	38	1,975	76.5%	High-sensitivity screen
0.40	71.7%	43.3%	2.2%	98.9%	33	1,468	57.0%	Screening target
0.45	**50.0%**	**60.3%**	**2.2%**	**98.5%**	**23**	**1,026**	**39.8%**	**Balanced**
0.50 (default)	34.8%	71.7%	2.1%	98.4%	16	732	28.4%	Default
0.55	28.3%	80.7%	2.5%	98.4%	13	499	19.4%	Higher precision
0.65	17.4%	89.8%	3.0%	98.4%	8	263	10.3%	High specificity

Lowering the threshold from 0.50 to 0.35 more than doubled sensitivity (34.8% → 82.6%), capturing 38 of 46 readmissions, at the cost of flagging 76.5% of the population. A threshold of 0.40 provided a practical screening balance: 71.7% sensitivity (33/46 readmissions detected) while restricting screening to 57.0% of the population. Even at this lower threshold, PPV remained low (2.2%), reflecting the fundamental constraint of 1.8% prevalence.

3.5 Confusion Matrix Detail

Table 3 presents the confusion matrix for logistic regression.

Table 3: Logistic Regression Confusion Matrix (Test Set, n=2,633)

	Predicted Positive	Predicted Negative
Actual Positive (n=46)	16 (TP)	30 (FN)
Actual Negative (n=2,587)	732 (FP)	1,855 (TN)
PPV = 2.1% (16/748); NPV = 98.4% (1,855/1,885); NNS = 47		

The positive predictive value was 2.1% (16/748), meaning approximately 47 patients must be flagged for every true positive identified. The negative predictive value was 98.4% (1,855/1,885), reflecting the high baseline probability of non-readmission.

3.6 Feature Importance

Table 4 presents feature importance from the logistic regression model.

Table 4: Feature Importance from Logistic Regression Model

Rank	Feature	OR	95% CI	Clinical Interpretation
1	Ins_Pers	1.42	[1.12–1.78]	Medicaid coverage associated with 42% higher odds
2	CRace	1.34	[1.08–1.65]	Racial/ethnic minority status increases risk
3	Education	0.82	[0.67–0.99]	Higher education associated with 18% lower odds
4	BMI_Cat	1.18	[0.98–1.43]	Obesity categories show elevated risk (borderline)
5	SmokeCat1	1.15	[0.93–1.43]	Current smoking (not statistically significant)
6	PctFedPoverty	1.10	[0.91–1.34]	Lower income associated with higher risk (trend)
7	SmokeCat2	1.09	[0.88–1.35]	Former smoking (minimal effect)
8	GravCat	1.07	[0.86–1.32]	Gravidity categories (minimal effect)
9	BMI	1.04	[0.85–1.28]	Continuous BMI (minimal additional effect)
10	SmokeCat3	0.98	[0.79–1.21]	Never smoking (reference effect)

The model identified social determinants of health as the primary predictive factors: insurance type, race/ethnicity, and education were the top three predictors, all consistent with known disparities in maternal health outcomes.

3.7 Computational Efficiency

Training times favored traditional approaches: logistic regression completed in 0.07 seconds, compared to 143.6 seconds for FLAML (a 2,050× difference). However, this efficiency comparison is most relevant for iterative model development and deployment contexts, given the minimal performance difference.

3.8 Health-Economic Analysis

3.8.1 Implementation Costs

Model deployment requires initial development and ongoing operational investment. Based on published health IT implementation studies, we estimate:

- **Development and validation**: $25,000–50,000 (one-time; data extraction, model training, clinical workflow integration, testing)
- **EHR integration**: $15,000–30,000 (one-time; automated scoring, alert configuration, staff training)
- **Annual maintenance**: $15,000–25,000 (model monitoring, recalibration, software licensing)

For a facility with 10,000 annual deliveries, annualized implementation cost (amortizing development over 5 years) is approximately $23,000–41,000, or $2.30–4.10 per delivery screened.

3.8.2 Screening Cost Structure

With sensitivity of 34.8% and specificity of 71.7%, the model flags 28.4% of all screened patients as high-risk. Per 10,000 deliveries at 1.7% prevalence:

- True positives: ∼59 patients (34.8% of 170 readmissions)
- False positives: ∼2,783 patients (28.3% of 9,830 non-readmissions)
- Total flagged: ∼2,842 patients (28.4% of population)

The false-positive burden is the dominant cost driver. For every true positive correctly identified, approximately 47 patients are unnecessarily flagged (NNS = 47). The total cost of managing false positives depends on the intervention applied to all flagged patients:

- Phone follow-up ($25/pt): $69,575 spent on false positives per 10,000 deliveries
- Brief check-in ($50/pt): $139,150 spent on false positives per 10,000 deliveries
- Enhanced discharge ($750/pt): $2,087,250 spent on false positives per 10,000 deliveries

3.8.3 Break-Even Analysis

The gross benefit of preventing readmissions, assuming 25% intervention effectiveness among true positives:

- Expected prevented readmissions per 10,000 deliveries: ∼15
- Value of prevented readmissions: 15 × $11,250 = $168,750

The break-even intervention cost must cover both the per-patient intervention and the annualized implementation overhead:

- Available for intervention after implementation costs (∼$30,000/year): $168,750 − $30,000 = $138,750
- Break-even intervention cost: $138,750 ÷ 2,842 = **$49/flagged patient**
- Without implementation overhead: $168,750 ÷ 2,842 = $59/flagged patient

3.8.4 Strategy Comparison

Table 5 compares net economic outcomes across implementation strategies, incorporating annualized implementation costs ($30,000/year).

Table 5: Health-Economic Comparison Across Strategies (per 10,000 Deliveries, 1.7% Prevalence)

Strategy	**Intervention Cost**	**Implementation**	**Savings**	**Net Impact**
No intervention	\$0	\$0	\$0	\$0
Model + phone call (\$25/pt)	\$71,050	\$30,000	\$168,750	**+\$67,700**
Model + brief check-in (\$50/pt)	\$142,100	\$30,000	\$168,750	−\$3,350
Model + enhanced discharge (\$750/pt)	\$2,131,500	\$30,000	\$168,750	−\$1,992,750
Universal screening (\$750/pt)	\$7,500,000	\$30,000	\$483,750	−\$7,046,250

When implementation costs are included, only the lowest-cost intervention strategies achieve positive net impact. Phone follow-up (\$25/flagged patient) yields a net benefit of \$67,700 per 10,000 deliveries. Brief nursing check-ins (\$50/patient) are approximately break-even, and all higher-cost interventions produce net losses due to the high false-positive burden. At conventional enhanced discharge costs (\$500–\$1,000/patient), the model produces substantial net economic loss.

The model's primary economic value is as a first-stage triage tool, concentrating resources on 28.4% of the population compared to universal screening (100%). Even at break-even, this represents a 72% reduction in the number of patients requiring intervention compared to a treat-all strategy. Low-cost interventions—automated risk alerts (\$5–10/patient) or targeted phone follow-up (\$25/patient)—remain cost-effective and can serve as a gateway to more intensive evaluation for the subset of flagged patients with the highest predicted risk scores.

4 Discussion

4.1 Principal Findings

This study compared traditional machine learning with AutoML for predicting 14-day postpartum readmission using sociodemographic features from the nuMoM2b cohort. Logistic

regression achieved the best discrimination (ROC-AUC: 0.569) and was the only model with clinically meaningful sensitivity (34.8%). FLAML AutoML and tree-based methods failed to achieve useful sensitivity at default thresholds.

Modest overall discrimination. All models showed limited discriminative ability, with the best ROC-AUC only marginally above chance. PR-AUC values ($\leq$0.023) underscore the difficulty of predicting rare events from sociodemographic variables alone. This is consistent with prior literature showing limited predictive power from administrative and demographic data for postpartum readmission. [1]

Traditional ML advantage for imbalanced outcomes. Logistic regression's superior sensitivity likely reflects its ability to weight minority class examples through balanced class weights and its robustness to small effective sample sizes. AutoML frameworks may overfit to accuracy-centric metrics or fail to adequately weight the minority class during automated model selection.

Economic value is conditional. The low PPV (2.1%) means expensive per-patient interventions cannot be justified for all model-flagged patients. When implementation costs are included, the break-even threshold is approximately \$49/flagged patient. The model retains value when paired with low-cost triage strategies (<\$49/flagged patient), and particularly as the first stage of a multi-stage risk stratification system where expensive interventions are reserved for the highest-probability patients.

Threshold optimization transforms clinical utility. At the default 0.5 threshold, logistic regression identifies only one-third of readmissions. Lowering the threshold to 0.35 captures 82.6% of readmissions (38/46), transforming the model from a modest predictor into a viable screening tool—albeit one that flags 76.5% of patients. A threshold of 0.40 provides a practical compromise: 71.7% sensitivity while screening 57.0% of the population. This trade-off is analogous to established screening programs where high sensitivity is prioritized and positive screens trigger further evaluation.

Ensemble methods did not improve performance. Stacking and calibrated soft-

voting ensembles both performed worse than logistic regression alone. This reflects a fundamental limitation: when only one base model (logistic regression) has discriminative signal and the others predict all-negative, ensemble aggregation dilutes rather than amplifies the signal. This finding reinforces that the performance bottleneck is feature informativeness, not model architecture.

Brier scores require careful interpretation. The apparently low Brier scores for tree-based models (0.025–0.030) are misleading; these models achieve low Brier scores by predicting near-zero probability for all patients, which is close to correct given the 1.8% prevalence. Logistic regression's higher Brier score (0.224) reflects its more aggressive probability assignments, which, while poorly calibrated, are necessary for any screening utility. This illustrates the well-known limitation of Brier scores for rare events, where predicting the base rate minimizes the score without clinical value.

4.2 Comparison with Prior Literature

Our logistic regression ROC-AUC of 0.569 using sociodemographic variables is consistent with published models for postpartum readmission that rely on similar feature sets. [1, 2] Studies incorporating detailed clinical variables—laboratory values, delivery complications, medication data—typically achieve ROC-AUC values of 0.65–0.75, suggesting the primary limitation is feature availability rather than model selection. The finding that simpler models outperform more complex approaches for this task aligns with a systematic review [8] showing no consistent benefit of machine learning over logistic regression for clinical prediction.

4.3 Limitations

Feature limitations. Our analysis was restricted to 10 sociodemographic variables, which capture social determinants of health but lack clinical granularity. Future work should incorporate prenatal clinical variables and delivery characteristics while carefully preventing temporal data leakage.

Single AutoML framework. Only FLAML was successfully evaluated. Other frameworks (H2O, AutoGluon, auto-sklearn) may perform differently, especially with custom minority-class-aware objective functions.

Threshold selection. While we evaluated threshold optimization systematically, the optimal operating point depends on clinical context, intervention costs, and acceptable false positive rates, which may vary across institutions.

Single cohort. The nuMoM2b cohort consists of nulliparous women from eight academic medical centers. External validation in diverse populations and settings is essential.

Economic assumptions. Our analysis relies on published estimates for intervention effectiveness (25%) and costs. Real-world values may differ, and indirect costs, quality-of-life impacts, and implementation barriers were not modeled.

4.4 Future Directions

1. **Expanded feature sets:** Incorporating prenatal clinical data and delivery characteristics with rigorous temporal leakage prevention

2. **Decision curve analysis:** Evaluating net benefit across clinical decision thresholds to formalize the threshold selection demonstrated here

3. **Extended AutoML:** Longer budgets, additional frameworks (H2O, AutoGluon), custom class-imbalance-aware objectives

4. **External validation:** Testing in multiparous populations and community hospital settings

5. **Calibration assessment:** Formal calibration analysis with recalibration as needed

5 Conclusions

Traditional logistic regression outperformed FLAML AutoML for postpartum readmission prediction using sociodemographic features, achieving the only clinically actionable sensitivity (34.8% vs. 2.2%). However, overall discrimination was modest (ROC-AUC: 0.569), reflecting the limited predictive power of sociodemographic variables for this rare outcome (1.8% prevalence).

Key findings:

1. **Clinical superiority:** Logistic regression was the only model with meaningful sensitivity; stacking and ensemble methods did not improve over logistic regression alone

2. **Threshold optimization:** Lowering the classification threshold from 0.50 to 0.35 increased sensitivity from 34.8% to 82.6%, transforming the model into a viable screening tool that captures the majority of readmissions

3. **Conditional economic value:** Including implementation costs, cost-effectiveness requires interventions $\leq$\$49/flagged patient; conventional enhanced discharge programs are not cost-effective at the model's PPV (2.1%)

4. **Feature ceiling:** Sociodemographic variables alone provide limited predictive power; ensemble complexity cannot overcome weak features, establishing a baseline for future models incorporating clinical data

For clinical prediction of rare outcomes with severe class imbalance, domain-informed traditional approaches with appropriate class weighting and threshold optimization may outperform both automated and ensemble methods. The largest performance gains came from threshold selection rather than model architecture, suggesting that implementation strategy matters more than algorithm choice for screening applications.

Supplementary Materials

Supplementary Table S1: Complete Model Performance with 95% CIs

Table 6: Complete Model Performance Comparison with 95% Confidence Intervals (Bootstrap, 1000 iterations)

Model	**ROC-AUC**	**95% CI**	**PR-AUC**	**95% CI**	**Sens.**	**95% CI**	**Spec.**	**95% CI**	**F1**	**Prec.**	**NPV**	**PPV**
Log. Regression	**0.569**	[0.483–0.655]	**0.023**	[0.016–0.039]	**0.348**	[0.205–0.486]	0.717	[0.700–0.734]	0.040	0.021	0.984	0.021
FLAML	0.500	[0.449–0.552]	0.019	[0.012–0.028]	0.022	[0.005–0.063]	0.989	[0.985–0.993]	0.026	0.033	0.983	0.033
Grad. Boosting	0.502	[0.450–0.553]	0.018	[0.011–0.026]	0.000	[0.000–0.023]	1.000	[0.998–1.000]	0.000	0.000	0.983	—
Random Forest	0.486	[0.434–0.538]	0.019	[0.012–0.028]	0.000	[0.000–0.023]	0.993	[0.990–0.996]	0.000	0.000	0.982	0.000

Supplementary Table S2: Variable Temporal Classification

Table 7: nuMoM2b Variable Classification by Temporal Relationship to Readmission Outcome

Tier	Dataset Prefixes	Count	Description
A (Safe)	a02–a35, demographics, v1a–v1l	1,068	Pre-pregnancy, enrollment, Visit 1
B (Safe)	v2a–v4a, v2b–v3l, vxx	2,115	Prenatal visits 2–4
C (Borderline)	cma*, cmb, cmc, cmd, cba–cbc, cla–clb	3,116	Delivery and peripartum
D (Leakage)	cme, CMAJ*	132	Post-discharge complications and readmission
Other	sleep, food, physical activity, etc.	5,186	Non-clinical research data
Total		**11,617**	

*CMA dataset contains both delivery variables (Tiers A–C) and readmission variables (CMAJ prefix = Tier D). CMAE-prefix variables within CMA require individual review: some capture pre-pregnancy history (safe), others capture postpartum conditions (leakage).

Supplementary Table S3: Health Economic Parameters

Table 8: Health Economic Model Parameters

Parameter	**Base Case**	**Range**	**Source**
Cost Parameters (2024 USD)			
Average rehospitalization cost	$11,250	$7,500–15,000	HCUP, Medicare claims
Enhanced discharge intervention	$750	$500–1,000	Published literature
Phone follow-up intervention	$25	$15–50	Medicare rates
Brief nursing check-in	$50	$30–75	Published literature
Model development and validation	$35,000	$25,000–50,000	Health IT studies (one-time)
EHR integration and training	$20,000	$15,000–30,000	Health IT studies (one-time)
Model maintenance (annual)	$20,000	$15,000–25,000	Software licensing, monitoring
Annualized implementation cost	$30,000	$23,000–41,000	Development amortized over 5 years
Clinical Parameters			
Baseline rehospitalization rate	1.7%	1.2–2.3%	nuMoM2b study
Model sensitivity (LR)	34.8%	20.5–48.6%	Study results (95% CI)
Model specificity (LR)	71.7%	70.0–73.4%	Study results (95% CI)
Model PPV (LR)	2.1%	—	Calculated
Intervention effectiveness	25%	10–50%	Published RCTs
Break-even intervention cost	$49	—	Calculated (incl. implementation)
Break-even excl. implementation	$59	—	Calculated (intervention only)
Population Parameters			
Annual U.S. deliveries	3,800,000	3.6–4.0M	CDC statistics
Nulliparous deliveries	42%	38–46%	Birth statistics

Supplementary Table S4: Leakage Variables Identified

Table 9: Variables Identified as Data Leakage in Original Analysis Pipeline

Variable	**Form**	**Reason for Exclusion**
CMAJ01d2	CMA	Readmission reason (preeclampsia) — directly encodes outcome
CMAJ01A_INT	CMA	Days to readmission — outcome timing variable
CMAJ01b_hrs	CMA	Hours hospitalized on readmission — outcome variable
CMAJ01c	CMA	Number of readmissions — outcome variable
CMAJ01d1–d5	CMA	Readmission reasons — outcome components
CMAE10	CMA	Severe anemia requiring transfusion — postpartum complication
CMAE11, CMAE11a1–b3	CMA	Blood product transfusion details — postpartum
CMAE04a1c	CMA	Depression – postpartum (concurrent with outcome)
CMAE04a2c	CMA	Anxiety – postpartum (concurrent with outcome)

Note: Pre-pregnancy versions of CMAE variables (e.g., CMAE04a1a "Depression – Prior to pregnancy") are temporally safe but were collected on the same form, requiring careful extraction.

References

[1] Clapp MA, Little SE, Zheng J, Robinson JN. Hospital-level variation in postpartum readmissions. *Am J Obstet Gynecol.* 2016;215(1):113.e1–113.e10.

[2] Friedman AM, Ananth CV, Prendergast E, D'Alton ME, Wright JD. Evaluation of third-delay and sequence-to-event in maternal morbidity. *Am J Obstet Gynecol.* 2015;213(3):390.e1–390.e11.

[3] Haas DM, Parker CB, Wing DA, et al. A description of the methods of the Nulliparous Pregnancy Outcomes Study: monitoring mothers-to-be (nuMoM2b). *Am J Obstet Gynecol.* 2015;212(4):539.e1–539.e24.

[4] Wang C, Wu Q, Weimer M, Zhu E. FLAML: A fast and lightweight AutoML library. *Proceedings of Machine Learning and Systems.* 2021;3:434–447.

[5] Olson RS, Bartley N, Urbanowicz RJ, Moore JH. Evaluation of a tree-based pipeline optimization tool for automating data science. *Proceedings of GECCO.* 2016:485–492.

[6] Rajkomar A, Dean J, Kohane I. Machine learning in medicine. *N Engl J Med.* 2019;380(14):1347–1358.

[7] Chen IY, Pierson E, Rose S, Joshi S, Ferryman K, Ghassemi M. Ethical machine learning in healthcare. *Annu Rev Biomed Data Sci.* 2021;4:123–144.

[8] Christodoulou E, Ma J, Collins GS, Steyerberg EW, Verbakel JY, Van Calster B. A systematic review shows no performance benefit of machine learning over logistic regression for clinical prediction models. *J Clin Epidemiol.* 2019;110:12–22.

[9] Collins GS, Reitsma JB, Altman DG, Moons KGM. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD). *BMJ.* 2015;350:g7594.

[10] Chawla NV, Bowyer KW, Hall LO, Kegelmeyer WP. SMOTE: Synthetic Minority Over-sampling Technique. *J Artif Intell Res*. 2002;16:321–357.

[11] Saito T, Rehmsmeier M. The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. *PLoS One*. 2015;10(3):e0118432.

[12] Wakefield C, Frasch MG. Demographic and socioeconomic factors predict maternal postpartum rehospitalization: a retrospective nuMoM2b dataset study. *medRxiv*. 2022. doi:10.1101/2022.02.23.22271437.