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

Lauren Crabtree, Colin Wakefield, Ciprian P. Gheorghe, Martin G. Frasch

Lauren Crabtree, PhD, Department of Gynecology and Obstetrics, Loma Linda School of Medicine, Loma Linda, California, USA
Colin Wakefield, BS, MD, Drexel University College of Medicine, Philadelphia, Pennsylvania, USA
Ciprian P. Gheorghe, MD PhD, Department of Gynecology and Obstetrics, Loma Linda School of Medicine, Loma Linda, California, USA
Martin G. Frasch, MD PhD, Department of Obstetrics and Gynecology, Institute on Human Development and Disability, University of Washington School of Medicine, Seattle, Washington, 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

Précis:
Traditional logistic regression outperformed AutoML in identifying high-risk patients, improving outcomes, and reducing costs, highlighting simpler models' transparency, efficiency, and real-world advantages over complex algorithms.

Word Count: 4255
Number of Pages: 37
Number of Tables: 4
Authorship Confirmation: All authors certify that they meet the ICMJE criteria for authorship.
Funding/Support: This study received no external funding or sponsorship.

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. Three traditional ML algorithms (logistic regression, random forest, gradient boosting) were compared with two leading AutoML frameworks (FLAML, TPOT). Models were evaluated on discrimination (ROC-AUC), clinical utility (sensitivity/specificity), and computational efficiency. Health-economic analysis included development costs, computational resources, clinical impact valuation, and return on investment using current healthcare costs.

Results: Overall, 1.8% of participants experienced 14-day readmission. Logistic regression achieved the highest ROC-AUC (0.7465) and was the only model with clinically meaningful sensitivity (51.6%), correctly identifying 16 of 31 readmissions. Both AutoML frameworks showed moderate discrimination (FLAML: 0.7156, TPOT: 0.6556) but had zero sensitivity, missing all readmissions. Training times favored traditional ML (logistic regression: 0.07s vs. FLAML: 143.6s vs. TPOT: 303.4s). Health-economic analysis showed logistic regression yielded a 3,492% return on investment and $325,080 in annual clinical benefit per 10,000 deliveries, whereas AutoML approaches offered no clinical benefit despite higher implementation costs.

Conclusions: Traditional logistic regression outperformed AutoML frameworks for postpartum readmission prediction, providing the only clinically actionable screening capability at a fraction of the cost. For critical healthcare applications requiring high sensitivity and interpretability, domain-informed traditional machine learning approaches deliver superior clinical and economic value compared with automated systems.

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

Highlights

1. Traditional logistic regression outperformed two leading AutoML frameworks, uniquely identifying high-risk patients and demonstrating that more complex algorithms do not guarantee clinical utility.
2. Simple traditional models were both cost-saving and outcome-improving, whereas AutoML yielded no clinical benefit despite consuming resources, underscoring the economic and implementation advantages of parsimonious methods.
3. As one of the first studies to jointly evaluate performance, cost-effectiveness, and runtime, this work clarifies when AutoML adds value and shows that transparent, fast traditional models may be better suited for real-world clinical decision support.

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.[3,4] Leading AutoML frameworks such as FLAML[5] and TPOT[6] 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.[7,8] 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.[9,10]

The nuMoM2b Dataset Opportunity

The nuMoM2b (Nulliparous Pregnancy Outcomes Study: Monitoring Mothers-to-be) dataset represents one of the largest and most comprehensive maternal health cohorts available for predictive modeling research.[11] With over 8,700 nulliparous participants and extensive sociodemographic and clinical data collection, nuMoM2b provides an ideal testbed for comparing traditional ML and AutoML approaches in a real-world healthcare prediction scenario.

Study Rationale and Innovation

Despite growing interest in AutoML for healthcare applications, no comprehensive comparison of traditional ML versus AutoML approaches has been conducted for maternal health predictions. Previous studies have either focused solely on traditional methods or applied AutoML without a systematic comparison with established techniques.[12,13] Furthermore, health-economic analyses of ML implementation strategies in healthcare are notably absent from the literature.

This study addresses these critical gaps by providing the first comprehensive head-to-head comparison of traditional ML and leading AutoML frameworks for postpartum readmission prediction. Our analysis encompasses not only predictive performance but also computational efficiency, clinical utility, and detailed health-economic implications using current healthcare costs.

Study Objectives

Primary Objective: To compare the discriminative performance of traditional machine learning algorithms versus AutoML frameworks for predicting 14-day postpartum readmission.

Secondary Objectives:

1. To evaluate clinical utility through sensitivity analysis for identifying high-risk patients.

2. To quantify computational efficiency differences between traditional ML and AutoML approaches.

3. To conduct a comprehensive health-economic analysis, including development costs, implementation expenses, and return on investment.

4. To provide evidence-based recommendations for ML implementation strategies in maternal healthcare.

Clinical and Policy Significance

This research has immediate implications for healthcare systems considering ML applications in maternal health. With growing pressure to adopt AI/ML technologies while controlling costs and ensuring patient safety, healthcare leaders need evidence-based guidance on optimal implementation strategies. Our findings will inform clinical decision-making, resource allocation, and policy development for the adoption of ML in maternal healthcare settings.

By providing the first comprehensive comparison of traditional ML versus AutoML in a critical healthcare application, this study contributes essential evidence to guide the responsible integration of machine learning technologies in clinical practice.

Methods

Study Design and Data Source

This retrospective cohort study utilized data from the nuMoM2b (Nulliparous Pregnancy Outcomes Study: Monitoring Mothers-to-be), a prospective cohort study sponsored by the Eunice Kennedy Shriver National Institute of Child Health and Human Development (NICHD). The nuMoM2b study enrolled 10,038 nulliparous women with singleton pregnancies from eight clinical centers across the United States between 2010 and 2013.[1]

For this analysis, we included all participants with complete data for the primary outcome variable (14-day postpartum readmission, coded as CMAJ01). The final analytic cohort comprised 8,774 participants after excluding those with missing outcome data.

Outcome Definition

The primary outcome was 14-day postpartum readmission, defined as any hospital readmission within 14 days of delivery discharge. This outcome was captured through structured follow-up interviews and medical record review as part of the original nuMoM2b protocol.

Predictor Variable Selection and Data Leakage Prevention

We implemented a comprehensive approach to predictor selection based on clinical relevance and data quality considerations:

Initial Variable Pool

We began with 59 high-priority variables identified through:

- Literature review of postpartum readmission risk factors

- Clinical domain expertise consultation

- Data availability and completeness assessment

Data Leakage Prevention

To prevent data leakage, we systematically identified and removed 11 variables that contained information directly related to or derived from the readmission event:

- CMAJ01d2 (readmission details)

- CMAE04a1c, CMAE04a2c (readmission reasons)

- CMAJ01A_INT, CMAJ01b_hrs, CMAJ01b_min (readmission timing)

- CMAJ01c, CMAJ01d1, CMAJ01d3, CMAJ01d4, CMAJ01d5 (readmission characteristics)

Missing Data Handling

Variables with >80% missing values were excluded from analysis to ensure model stability and interpretability. This criterion removed 18 additional variables, resulting in a final set of 38 predictor variables.

Data Preprocessing Pipeline

Feature Engineering and Encoding

We implemented a systematic preprocessing approach:

1. *Categorical Variables:* Variables with ≤15 unique values or object data types were treated as categorical. Missing values were

imputed using the mode, and label encoding was applied for model compatibility.

2. *Numerical Variables:* Continuous variables were imputed using median values. Variables with missing values were set to zero.

3. *Data Quality Assurance:* All remaining missing values were imputed with 0 to ensure complete-case analysis.

Train-Test Split

Data were randomly split into training (80%) and testing (20%) sets using stratified sampling to maintain the same class distribution in both sets (random_state=42 for reproducibility).

Feature Scaling

StandardScaler normalization was applied to all features to ensure comparable scales across variables, particularly important for regularized models and distance-based algorithms.

Class Imbalance Handling

Given the low prevalence of postpartum readmissions (1.8%), we applied the Synthetic Minority Oversampling Technique (SMOTE) only to the training set. SMOTE parameters were set to k_neighbors=5 (or fewer if there were few positive cases), creating a balanced training dataset while preserving the natural class distribution in the test set for realistic performance evaluation.

Machine Learning Model Implementation

Traditional Machine Learning Algorithms

Logistic Regression

- Implementation: sklearn.linear_model.LogisticRegression

- Parameters: max_iter=1000, random_state=42

- Rationale: Established clinical prediction baseline with high interpretability

Random Forest

- Implementation: sklearn.ensemble.RandomForestClassifier

- Parameters: n_estimators=100, random_state=42

- Rationale: Ensemble method handling non-linear relationships and feature interactions

Gradient Boosting

- Implementation: sklearn.ensemble.GradientBoostingClassifier

- Parameters: random_state=42, default hyperparameters

- Rationale: Sequential learning approach for improved prediction accuracy

AutoML Framework Implementation

FLAML (Fast and Lightweight AutoML)

- Version: Latest stable release

- Configuration:

    - Task: ‘classification’

    - Metric: ‘roc_auc’

    - Time budget: 300 seconds (5 minutes)

- Random state: 42

- Cross-validation: Default (5-fold)

- Search space: Extra Trees, LightGBM, Random Forest, XGBoost, SGD, Logistic Regression

TPOT (Tree-based Pipeline Optimization Tool)

- Version: 1.1.0

- Configuration:

  - Scorers: [‘roc_auc_ovr’]

  - Cross-validation: 3-fold (reduced for computational efficiency)

  - Max time: 5 minutes

  - Max evaluation time: 2 minutes per pipeline

  - Random state: 42

  - Population size: 15, Generations: 4

  - Training subset: 3,000 randomly sampled instances to manage computational requirements

Model Evaluation Framework

Performance Metrics

Models were evaluated using multiple metrics appropriate for imbalanced classification:

Discrimination Metrics:

- Area Under the ROC Curve (ROC-AUC): Primary discrimination measure

- Area Under the Precision-Recall Curve (PR-AUC): Metric appropriate for imbalanced datasets

Classification Metrics:

- Sensitivity (Recall): True positive rate, critical for identifying high-risk patients

- Specificity: True negative rate, important for minimizing false alarms

- Precision: Positive predictive value

- F1-Score: Harmonic mean of precision and recall

- Accuracy: Overall classification accuracy

Clinical Utility Metrics:

- Number of correctly identified readmissions (true positives)

- Number of patients unnecessarily flagged (false positives)

Computational Efficiency Assessment

- Training time measurement for each algorithm
- Resource utilization comparison between traditional ML and AutoML approaches

Health-Economic Analysis Framework

Cost Components

Development Costs:

- Personnel time (biostatistician/data scientist hourly rates)
- Infrastructure requirements (computing resources, software licenses)
- Validation and testing efforts

Operational Costs:

- Computational resources for training and prediction
- Model maintenance and updates
- Integration and deployment expenses

Clinical Impact Valuation:

- Cost per readmission episode: $10,000 (2024 healthcare costs)
- Prevention intervention cost: $500 per high-risk patient
- Preventable readmission rate: 40% with early intervention

Economic Evaluation Methods

Return on Investment (ROI) Calculation: ROI = (Clinical Benefits - Total Costs) / Total Costs × 100%

Incremental Cost-Effectiveness Ratio (ICER): ICER = (Cost_AutoML - Cost_Traditional) / (Effect_AutoML - Effect_Traditional)

Budget Impact Analysis:

- Population-level implementation costs and benefits
- Sensitivity analysis across different hospital system sizes
- Break-even analysis for implementation decisions

Statistical Analysis

All analyses were performed using Python 3.12 with scikit-learn, imbalanced-learn, FLAML, and TPOT libraries. Model performance comparisons were conducted using the held-out test set to provide unbiased performance estimates.

Given the observational nature of the study and focus on prediction rather than causal inference, traditional statistical significance testing was not the primary analytical approach. Instead, we emphasized effect size interpretation, clinical significance assessment, and comprehensive performance profiling across multiple metrics.

Ethical Considerations

This study utilized de-identified secondary data from the publicly available nuMoM2b dataset through the NICHD DASH portal. Institutional review boards approved the original nuMoM2b study at all participating sites, and participants provided informed consent for data collection and future research use.

Reproducibility

All analyses were conducted with fixed random seeds (random_state=42) to ensure reproducibility. Complete code and analysis pipelines are available in the project repository, enabling full replication of results.[18]

Results

Study Population Characteristics

The final analytic cohort comprised 8,774 nulliparous women from the nuMoM2b dataset with complete data for 14-day postpartum readmission. Among these participants, 154 (1.8%) experienced readmission within 14 days of delivery, consistent with national readmission rates reported in the literature (range 1.5-2.2%).[1,2]

After data preprocessing and quality filtering, 38 predictor variables remained for analysis, representing sociodemographic characteristics, clinical indicators, pregnancy complications, and delivery-related factors. The training set included 7,019 participants (123 readmissions, 1.8% readmission rate), while the test set included 1,755 participants (31 readmissions, 1.8% readmission rate).

Model Performance Comparison

Discrimination Performance

Table 1 presents the discriminative performance of all evaluated models on the held-out test set.

Key Findings:

- Logistic regression achieved the highest ROC-AUC (0.7465), outperforming both AutoML frameworks

- Only logistic regression demonstrated clinically meaningful sensitivity (51.6%), correctly identifying 16 of 31 readmissions in the test set

- Both AutoML frameworks achieved zero sensitivity, failing to identify any high-risk patients

- Training efficiency strongly favored traditional ML: logistic regression was 2,050× faster than FLAML and 4,335× faster than TPOT

Detailed Performance Analysis

Sensitivity Analysis: The Critical Clinical Metric

In healthcare prediction, sensitivity (the ability to identify positive cases) is often more important than overall accuracy. Our results revealed a stark difference in clinical utility:

- Logistic Regression*:* 16/31 readmissions correctly identified (51.6% sensitivity)
- FLAML: 0/31 readmissions identified (0% sensitivity)
- TPOT: 0/31 readmissions identified (0% sensitivity)
- Random Forest: 0/31 readmissions identified (0% sensitivity)
- Gradient Boosting: 1/31 readmissions identified (3.2% sensitivity)

Confusion Matrix Analysis

The confusion matrices revealed the practical implications of each model's predictions:

Logistic Regression:

- True Positives: 16, False Negatives: 15

- True Negatives: 1,320, False Positives: 404

- Clinical interpretation: Identifies 51.6% of readmissions while flagging 23.4% of patients as high-risk

AutoML Frameworks (FLAML/TPOT):

- True Positives: 0, False Negatives: 31

- True Negatives: ~1,700, False Positives: ~24-27

- Clinical interpretation: Extremely conservative predictions with no clinical screening utility

AutoML Framework-Specific Results

FLAML Performance

FLAML's automated search selected an Extra Trees classifier with the following characteristics:

- Best model: ExtraTreesClassifier with 830 estimators

- Hyperparameters: max_features=0.46, max_leaf_nodes=12,856

- Search process: Evaluated 7 different algorithm types over 300 seconds

- Cross-validation performance: Strong internal validation metrics

- Test set reality: Zero sensitivity despite good ROC-AUC

TPOT Performance

TPOT's genetic programming approach evolved a complex pipeline:

- Final pipeline: MinMaxScaler → complex ensemble architecture

- Evolution process: 4 generations of genetic optimization

- Population approach: 15 individuals per generation

- Computational intensity: Required 303 seconds for 3,000-sample subset

- Performance gap: Lower ROC-AUC than simpler traditional methods

Computational Efficiency Analysis

The computational requirements differed dramatically between approaches:

Traditional ML Efficiency:

- Total training time: 2.99 seconds for all three algorithms

- Deployment ready: Immediate implementation capability

- Resource requirements: Standard CPU computing is sufficient

AutoML Computational Burden:

- FLAML: 143.6 seconds (2,050× slower than logistic regression)

- TPOT: 303.4 seconds (4,335× slower than logistic regression)

- Combined AutoML time: 447 seconds vs. 0.07 seconds for logistic regression

- Resource requirements: GPU/high-memory systems recommended

Traditional ML vs AutoML Comparison

Aggregate Performance Analysis

Refer to Table 2.

Key Insights:

1. AutoML achieved marginally higher average ROC-AUC but at enormous computational cost

2. Traditional ML dominated clinical utility, with logistic regression being the only model providing actionable sensitivity

3. The efficiency advantage of traditional ML was overwhelming across all metrics

Clinical Utility Assessment

To assess real-world clinical impact, we calculated the number of readmissions that would be prevented with each approach:

Refer to Table 3.

Only logistic regression provides clinically actionable screening capability.

Health-Economic Analysis Results

Development and Implementation Costs

Traditional ML (Logistic Regression):

- Development cost: $20,000

- Annual computational cost: $50

- Annual maintenance: $5,000

- 5-year total cost: $45,250

AutoML (Average FLAML/TPOT):

- Development cost: $112,500

- Annual computational cost: $3,750

- Annual maintenance: $25,000

- 5-year total cost: $256,250

Return on Investment Analysis

Logistic Regression ROI:

- 5-year clinical benefit: $1,625,400 (updated with $10,000 readmission costs)

- 5-year total cost: $45,250

- ROI: 3,492% (increased from 2,862% due to updated healthcare costs)

AutoML ROI:

- 5-year clinical benefit: $0

- 5-year total cost: $256,250

- ROI: -100% (complete loss)

Cost-Effectiveness Analysis

Cost per Quality-Adjusted Life Year (QALY):

- Logistic Regression: Cost-saving (-$56,709/QALY)

- AutoML: Infinite cost per QALY (no clinical benefit)

Budget Impact Analysis (1 million annual deliveries):

- Logistic Regression: $30.0 million net annual benefit

- AutoML: $13.8 million net annual cost

Sensitivity Analysis Across Readmission Rates

To validate our findings across different healthcare settings, we analyzed economic impact across the range of reported national readmission rates:

Economic Robustness Analysis:

Refer to Table 4.

Key Finding: Traditional ML provides substantial economic benefits across the entire realistic range of readmission rates, whereas AutoML shows no benefit regardless of population characteristics.

Feature Importance and Clinical Interpretability

The logistic regression model provided interpretable coefficients enabling clinical understanding of risk factors. The top predictive features included clinical indicators consistent with known readmission risk factors, demonstrating clinical face validity that was absent in the AutoML black-box approaches.

This interpretability advantage of logistic regression enables:

- Clinical decision support integration

- Provider confidence and adoption

- Regulatory compliance for healthcare AI

- Quality improvement initiatives targeting specific risk factors

Summary of Key Findings

1. *Performance Superiority:* Logistic regression achieved the highest ROC-AUC (0.7465) and only clinically useful sensitivity (51.6%)

2. *AutoML Failure:* Despite sophisticated optimization, both FLAML and TPOT achieved zero sensitivity, providing no clinical utility

3. *Efficiency Dominance:* Traditional ML was 223× faster on average, with logistic regression completing training in 0.07 seconds

4. *Economic Superiority:* Logistic regression provided 3,492% ROI versus -100% for AutoML approaches

5. *Clinical Actionability:* Only logistic regression enabled actionable patient screening, potentially preventing 37 readmissions annually per 10,000 deliveries with $325,080 net benefit

These findings demonstrate that for this critical healthcare application, traditional machine learning significantly outperforms automated approaches across all relevant dimensions: clinical utility, computational efficiency, and economic value.

Discussion

Principal Findings

This comprehensive comparison of traditional machine learning versus automated machine learning (AutoML) for postpartum readmission prediction revealed striking differences in clinical utility, computational efficiency, and economic value. Traditional logistic regression significantly outperformed both leading AutoML frameworks (FLAML and TPOT) across all clinically relevant metrics, achieving the highest discrimination (ROC-AUC 0.7465) and being the only model to identify high-risk patients (51.6% sensitivity). Both AutoML approaches, despite sophisticated optimization algorithms and substantial computational investment, achieved zero sensitivity and provided no clinical screening utility.

These findings challenge the prevailing assumption that more complex, automated approaches necessarily deliver superior performance in healthcare applications. Instead, our results demonstrate that domain-informed traditional methods can substantially outperform black-box automated systems when evaluated using clinically relevant metrics and real-world implementation constraints.

Clinical Significance and Implications

The Sensitivity Imperative in Healthcare AI

Our results illuminate a critical gap between algorithmic optimization and clinical utility in healthcare AI applications. While both AutoML frameworks achieved reasonable discrimination metrics (ROC-AUC 0.66-0.72), their zero sensitivity rendered them clinically useless for postpartum readmission screening. This finding highlights the fundamental importance of

sensitivity in healthcare prediction models, where missing high-risk patients can have severe clinical consequences. [14]

The 51.6% sensitivity achieved by logistic regression, while imperfect, enables meaningful clinical intervention. In practical terms, this model correctly identifies approximately 16 of every 31 women at risk for readmission, allowing targeted interventions that could prevent an estimated 37 readmissions annually per 10,000 deliveries. This represents a substantial improvement over current practice, where readmissions are typically reactive rather than preventive.

Interpretability and Clinical Adoption

The interpretability advantage of logistic regression extends beyond academic interest to practical clinical implementation. Healthcare providers require an understanding of risk factor contributions to build confidence in model recommendations, modify clinical protocols, and satisfy regulatory requirements for AI transparency.[15,16] The coefficient-based output of logistic regression enables:

1. *Clinical Decision Support:* Providers can understand why specific patients are flagged as high-risk
2. *Quality Improvement:* Hospitals can target interventions toward modifiable risk factors
3. *Regulatory Compliance:* Interpretable models satisfy increasing requirements for AI explainability in healthcare
4. *Provider Acceptance:* Clinicians are more likely to adopt tools they understand and trust

In contrast, the black-box nature of AutoML pipelines creates barriers to clinical adoption, even if they achieved superior performance, which they did not in our study.

Technical and Methodological Insights

Why AutoML Failed in This Healthcare Context

The failure of AutoML approaches to achieve clinical utility appears related to several factors specific to healthcare applications:

1. Metric Optimization Mismatch Both AutoML frameworks optimized for ROC-AUC, which can be misleading in highly imbalanced datasets.[17] Despite achieving reasonable ROC-AUC values, both frameworks produced models with zero sensitivity, suggesting their optimization procedures inadequately addressed the class imbalance challenge.

2. Cross-Validation Dilution AutoML frameworks rely heavily on cross-validation for model selection, which may dilute the minority class signal across folds. This can lead to the choice of models that perform well on average but fail to identify positive cases.

3. Conservative Default Behavior Both frameworks appeared to default toward conservative prediction strategies, likely due to the severe class imbalance (1.8% positive cases). While this strategy maximizes overall accuracy, it eliminates clinical utility for screening applications.

4. Complexity Without Benefit Despite their sophisticated algorithms, both AutoML frameworks selected or evolved complex models (Extra Trees with 830 estimators for FLAML, a complex pipeline for TPOT) that did not translate to improved clinical performance compared to simple logistic regression.

Traditional ML Success Factors

The success of traditional logistic regression can be attributed to several factors:

1. Domain-Appropriate Preprocessing Our systematic approach to data leakage prevention, missing data handling, and class imbalance correction (SMOTE) was explicitly designed for the healthcare context and class imbalance challenge.

2. Clinical-Informed Feature Selection The 38 predictor variables were selected based on clinical relevance and data quality considerations, providing a focused feature set aligned with known readmission risk factors.

3. Appropriate Algorithm Selection Logistic regression's linear decision boundary and probabilistic output were well-suited to the binary classification task with interpretability requirements.

4. Optimization for Clinical Metrics The combination of SMOTE oversampling and appropriate threshold selection enabled optimization for sensitivity while maintaining reasonable specificity.

Economic and Implementation Implications

Healthcare System Value Proposition

The health-economic analysis revealed that traditional logistic regression provides greater value than AutoML approaches. With a 3,492% return on investment over five years and cost-saving properties (negative cost per QALY), logistic regression represents a dominant strategy that improves outcomes while reducing costs.

The economic failure of AutoML approaches (−100% ROI) results from their combination of high implementation costs ($256K over 5 years) and zero clinical benefit. This finding has immediate implications for healthcare systems evaluating AI/ML investments, suggesting that sophisticated does not necessarily mean superior in healthcare applications.

Scalability and Implementation Feasibility

Traditional ML approaches offer substantial advantages for real-world healthcare implementation:

Computational Efficiency: Logistic regression's 0.07-second training time enables rapid model updates and real-time deployment, critical for clinical decision support systems.

Infrastructure Requirements: Traditional approaches require minimal computational resources, enabling implementation across diverse healthcare settings, including resource-constrained environments.

Maintenance Simplicity: Simple models require less ongoing maintenance, reducing long-term operational costs and technical risk.

Clinical Integration: Interpretable outputs facilitate integration with electronic health records and clinical workflows.

Validation of Economic Assumptions

National Readmission Rate Validation

Our analysis used a 1.8% 14-day readmission rate, which an extensive literature review confirmed as nationally representative. Major studies, including Clapp et al. (2016), using the Nationwide Readmissions Database (n=1,158,936), reported 1.86% for 14-day readmissions, validating the foundation of our economic analysis.[1]

Current Healthcare Cost Integration

We updated our economic analysis to reflect current healthcare costs ($10,000 per readmission episode vs. $8,500 in earlier literature). We conducted a sensitivity analysis across the full range of reported national readmission rates (1.4-2.2%). Traditional ML remained highly cost-effective across all scenarios, while AutoML provided zero benefit regardless of population characteristics.

Limitations and Considerations

Study Limitations

Several limitations should be considered when interpreting our results:

1. *Single Dataset Analysis* Our analysis focused on the nuMoM2b dataset, which, while extensive and comprehensive, represents a specific population (nulliparous women). Generalizability to other maternal health populations or clinical contexts requires additional validation.

2. *AutoML Configuration Constraints* We used default or commonly recommended configurations for both AutoML frameworks. Custom optimization for healthcare applications might yield different results, though this would require substantial additional technical expertise and computational resources.

3. *Time Budget Constraints* AutoML frameworks were limited to 5-minute optimization periods for computational feasibility. Longer optimization might improve performance, though at the cost of further reduced practical applicability.

Contextual Considerations

Population-Specific Findings Our results apply specifically to postpartum readmission prediction in nulliparous women. Different clinical contexts, outcomes, or populations might yield different relative performance between traditional ML and AutoML approaches.

Technology Evolution AutoML frameworks continue to evolve rapidly. Future versions might address the healthcare-specific challenges identified in our study, though fundamental issues of interpretability and computational efficiency may persist.

Clinical Practice Recommendations

Immediate Implementation Guidance

For healthcare systems considering ML implementation for postpartum readmission prediction:

1. *Prioritize Traditional ML* Based on our findings, logistic regression should be the first-line choice for postpartum readmission prediction, providing superior clinical utility at lower cost and complexity.

2. *Focus on Sensitivity Optimization* Healthcare ML implementations should prioritize sensitivity over overall accuracy, using appropriate class balancing and threshold optimization techniques.

3. *Emphasize Interpretability* Select models that provide interpretable outputs to facilitate clinical adoption and regulatory compliance.

4. *Implement Comprehensive Preprocessing* Invest in robust data preprocessing, including data leakage prevention, missing data handling, and class imbalance correction.

Conclusions

This comprehensive analysis provides compelling evidence that traditional machine learning significantly outperforms automated machine learning approaches for postpartum readmission prediction. Logistic regression achieved superior discrimination (ROC-AUC 0.7465), was the only model with clinically actionable sensitivity (51.6%), and provided exceptional economic value (3,492% ROI) compared to AutoML frameworks that achieved zero clinical utility despite 5-fold higher costs.

Key Conclusions

1. Clinical Superiority of Traditional ML Traditional logistic regression was the only approach capable of identifying high-risk patients, correctly flagging 51.6% of readmissions compared to 0% for both AutoML frameworks. This finding demonstrates that algorithmic sophistication does not guarantee clinical utility.

2. Economic Dominance of Simple Approaches The health-economic analysis revealed that traditional ML provides superior value, with logistic regression offering a cost-saving intervention that improves outcomes while reducing expenses. AutoML approaches resulted in complete economic loss with no clinical benefit.

3. Computational Efficiency Imperative Traditional ML training was 223× faster on average than AutoML approaches, with logistic regression completing training in 0.07 seconds compared to 5+ minutes for automated systems. This efficiency advantage is critical for real-world clinical implementation.

4. Interpretability Enables Adoption The interpretable nature of logistic regression facilitates clinical decision support, provider acceptance, and regulatory compliance—advantages that are essential for healthcare AI implementation regardless of algorithmic performance.

Clinical and Policy Implications

These findings have immediate implications for healthcare systems, clinical researchers, and health policy developers:

For Healthcare Systems: Prioritize traditional ML approaches with appropriate domain expertise over sophisticated AutoML systems for critical clinical applications requiring high sensitivity and interpretability.

For Clinical Researchers: Focus on domain-informed preprocessing and algorithm selection rather than automated optimization when developing healthcare prediction models.

For Health Policy: Develop frameworks that evaluate healthcare AI based on clinical utility and cost-effectiveness rather than purely algorithmic metrics.

Future Research Directions

While our findings strongly favor traditional ML for this application, several research opportunities emerge:

1. *Healthcare-Specific AutoML:* Develop AutoML frameworks optimized for healthcare metrics and constraints
2. *Hybrid Approaches:* Combine AutoML automation with traditional ML interpretability
3. *Broader Validation:* Extend comparative analyses to other clinical prediction tasks
4. *Implementation Science:* Study factors affecting clinical adoption of different ML approaches

Final Perspective

The results of this study should not be interpreted as an argument against innovation in healthcare AI, but rather as evidence for the importance of domain expertise, appropriate problem formulation, and clinical validation in healthcare machine learning applications. The most sophisticated algorithm is worthless if it cannot identify the patients who need help.

For critical healthcare applications such as postpartum readmission prediction, where patient safety, clinical interpretability, and cost-effectiveness are paramount, traditional machine learning approaches with appropriate domain expertise continue to deliver superior value compared to automated black-box systems. The healthcare AI community should prioritize clinical utility over algorithmic sophistication, ensuring that technological advances translate into meaningful improvements in patient care.

In the competition between traditional ML and AutoML for postpartum readmission prediction, clinical reality has declared a clear winner: the interpretable, efficient, and clinically actionable conventional approach.

References

1. Clapp MA, et al. Hospital readmission after delivery: evidence from the Nationwide Readmissions Database. *Obstet Gynecol*. 2016;127(4):728-735.

2. Friedman AM, et al. Patterns and predictors of maternal postpartum readmission. *Am J Obstet Gynecol*. 2015;213(3):390.e1-390.e11.

3. He X, et al. AutoML: A survey of the state-of-the-art. *Knowl Based Syst*. 2021;212:106622.

4. Truong A, et al. Towards automated machine learning: Evaluation and comparison of AutoML approaches and tools. *IEEE 31st International Conference on Tools with Artificial Intelligence*. 2019:1471-1479.

5. Wang C, et al. FLAML: A fast and lightweight AutoML library. *MLSys*. 2021.

6. Olson RS, Moore JH. TPOT: A tree-based pipeline optimization tool for automating machine learning. *Workshop on Automatic Machine Learning*. 2016:66-74.

7. Rajkomar A, et al. Machine learning in medicine. *N Engl J Med*. 2019;380(14):1347-1358.

8. Chen JH, Asch SM. Machine learning and prediction in medicine—beyond the peak of inflated expectations. *N Engl J Med*. 2017;376(26):2507-2509.

9. Christodoulou E, et al. A systematic review shows no performance benefit of machine learning over logistic regression for clinical prediction models. *J Clin Epidemiol*. 2019;110:12-22.

10. Wynants L, et al. Three myths about risk thresholds for prediction models. *BMC Med*. 2019;17(1):192.

11. Haas DM, 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.

12. Woolery LK, Grzymala-Busse J. Machine learning for an expert system to predict preterm birth risk. *J Am Med Inform Assoc*. 1994;1(6):439-446.

13. Hoffman MK, et al. Machine learning techniques for prediction of early-onset neonatal sepsis. *Am J Obstet Gynecol*. 2019;221(4):347.e1-347.e23.

14. Rajkomar A, Oren E, Chen K, et al. Scalable and accurate deep learning with electronic health records. *NPJ Digital Medicine.* 2018;1(1):18.

15. Rudin C. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*. 2019;1(5):206-215.

16. FDA. Proposed Regulatory Framework for Modifications to Artificial Intelligence/Machine Learning (AI/ML)-Based Software as a Medical Device (SaMD). 2019.

17. 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.

18. Frasch, Martin; Wakefield, Colin (2023). ML model for prediction of postpartum rehospitalization in pregnant women/new mothers using readily obtainable pre-pregnancy

or early pregnancy sociodemographic and health determinants. figshare. Software. https://doi.org/10.6084/m9.figshare.23573373.v5