phuse

connect·share·advance

Single Day Events

Chennai, India

phuse.global

phuse

connect·share·advance

Practical Implementation of Machine Learning in Biostatistics

Dr. Vaishali Thakkar

19-Sep-2026

connect·share·advance

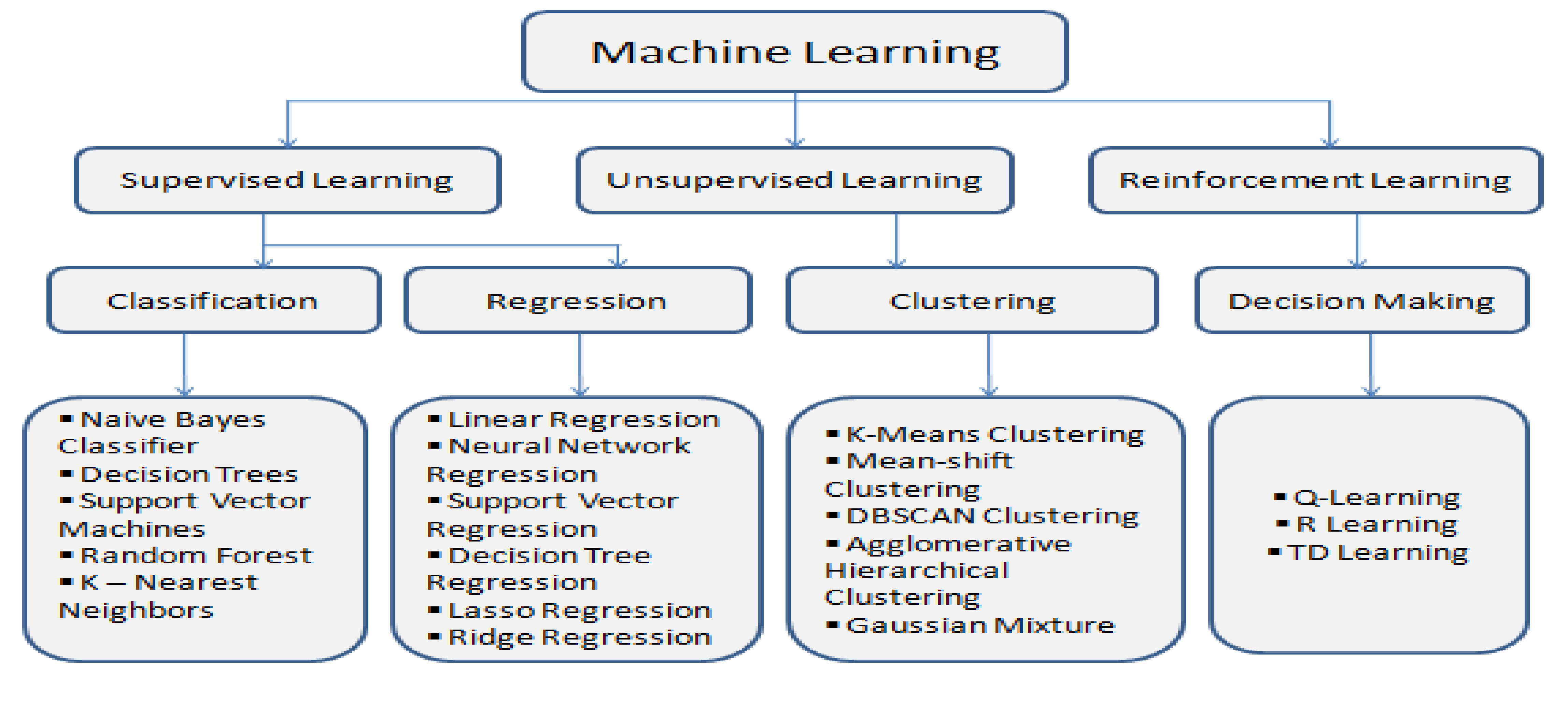
Machine Learning

Machine Learning (ML) comprises of algorithms and statistical models used by machines to perform a specific task and improve with experience by learning on it own.

Supervised Learning	Unsupervised Learning	Reinforcement Learning
Task Driven	Data Driven	Data Driven
Input Data	Input Data	Learn from Mistakes
Output Data – Training	Cluster Identification	
Can be used for predicting	Can not be used for predicting	

Machine Learning Models

connect·share·advance



Why Machine Learning in Biostatistics?

- Traditional biostatistics is primarily focused on inference
  - Estimate treatment effects
  - Test hypotheses
  - Quantify uncertainty
  - Control Type I error
  - Support regulatory decisions

- Machine learning adds a complementary capability
  - Prediction
  - Pattern recognition
  - Nonlinear relationships
  - Complex interactions
  - Risk stratification

Key Message: ML should complement-not replace-sound biostatistical principles.

What Makes Clinical Data Different?

Clinical/healthcare data are often:

- High-dimensional
- Heterogeneous
- Longitudinal
- Missing
- Correlated
- Subject to measurement error
- Imbalanced
- Vulnerable to bias

Example:
A clinical trial may contain:
1,500 participants × 100+ baseline/longitudinal variables

This creates opportunities—and challenges—for ML.

Where Can ML Add Value?

Area	Potential ML Utility
Patient recruitment	Identify likely eligible participants
Patient retention	Predict dropout
Safety	Identify high-risk participants
Efficacy	Predict treatment response
Biomarkers	Identify predictive features
Precision medicine	Individual risk/benefit prediction
RWE	Patient phenotyping
Clinical trials	Site/recruitment optimization

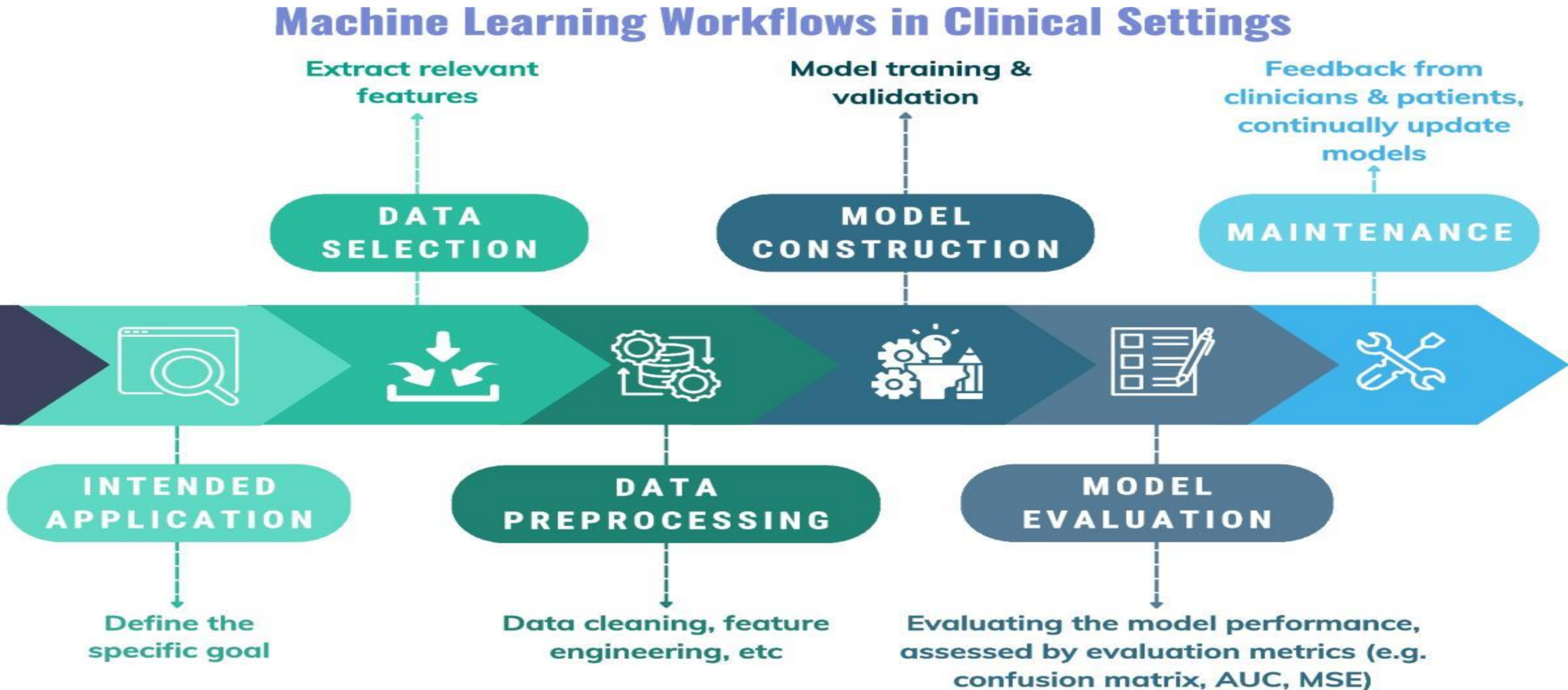
phuse

connect·share·advance

Practical ML Workflow

connect·share·advance

Where Does ML Fit in Clinical Research?



Step 1: Define the Clinical Question

- **Before choosing an algorithm:**
  - What is the clinical objective?
  - Who is the target population?
  - What is the outcome?
  - When is prediction required?
  - What information is available at prediction time?
  - What decision will the prediction support?

Example
Predict probability of clinical-trial dropout between Week 5 and Week 12 using information available through Week 4.

Step 2: Data Quality & Exploration

- **Evaluate:**
  - Missing values
  - Outliers
  - Distribution
  - Correlation
  - Class imbalance
  - Duplicate records
  - Measurement inconsistencies
  - Site effects

Key principle:

Garbage in → garbage out.

A sophisticated algorithm cannot compensate for poor-quality clinical data

Step 3: Missing Data & Feature Engineering

- **Missing Data**
  - Complete case
  - Single imputation
  - Multiple imputation
  - Model-based methods

Feature engineering
Examples:
- Change from baseline
- Number of missed visits
- Compliance percentage
- AE count
- Laboratory trends

Critical requirement:
Avoid using information that was unavailable at the prediction time.

Step 4: Feature Selection

- **Potential approaches:**
  - Clinical knowledge
  - Univariate screening
  - LASSO
  - Elastic Net
  - Recursive Feature Elimination
  - Tree-based importance

Critical issue:
Feature selection performed before validation can cause **optimistic model performance**.

Therefore:
Feature selection should be incorporated within the resampling/validation procedure.

Step 5: Train / Validation / Test

Example: **1,500 participants**

- Training — 60%
- Validation — 20%
- Test — 20%

Step 6: Model Performance

For binary prediction:

- ROC-AUC
- Sensitivity
- Specificity
- PPV
- NPV
- F1 score

For continuous prediction:

- RMSE
- MAE
- R^2

For survival prediction:

- C-index
- Time-dependent AUC
- Brier score

But predictive discrimination is not enough.

connect·share·advance

Case Study

Predicting Clinical Trial Dropout Risk

Clinical Problem

Participant dropout can result in:

- Missing outcome data
- Reduced statistical power
- Potential bias
- Increased sample size requirements
- Longer recruitment
- Increased cost

Research question

Can ML identify participants at high risk of dropout early enough to enable preventive intervention?

Hypothetical Study Dataset

N = 1,500 participants

Potential predictors:

- Age
- Sex
- Disease severity
- Disease duration
- Prior treatment
- Baseline laboratory values
- Missed visits
- Treatment compliance
- Early adverse events
- Concomitant medications
- Study site

Outcome: Dropout between Week 5 and Week 12.

Prediction Time Point

Prediction made at Week 4 -Only data available up to Week 4 are permitted.

Baseline —— Week 2 —— Week 4 | Week 5 ———————— Week 12

↑

Prediction point

Predict: Probability of dropout from Week 5–12.

Model Development

Compare:
Model 1: Logistic Regression
Model 2: Random Forest
Model 3: XGBoost

Workflow:

Clinical data
↓
Quality assessment
↓
Pre-processing
↓
Feature selection
↓
Cross-validation
↓
Model tuning
↓
Independent testing

connect·share·advance

Hypothetical/Illustrative Model Results

Metric	Logistic Regression	Random Forest	XGBoost
AUC	0.71	0.77	0.80
Sensitivity	68%	74%	78%
Specificity	67%	71%	72%
Brier score	0.19	0.16	0.15

Interpretation: XGBoost demonstrates better predictive performance in this hypothetical example but, AUC alone does not establish clinical utility.

Calibration & Clinical Utility

Discrimination
Can the model distinguish high-risk from low-risk participants?

Calibration
Do predicted probabilities correspond to observed probabilities?

Example:
Predicted dropout risk = 30%
Observed dropout ≈ 30%

Clinical utility
Could the prediction change management?

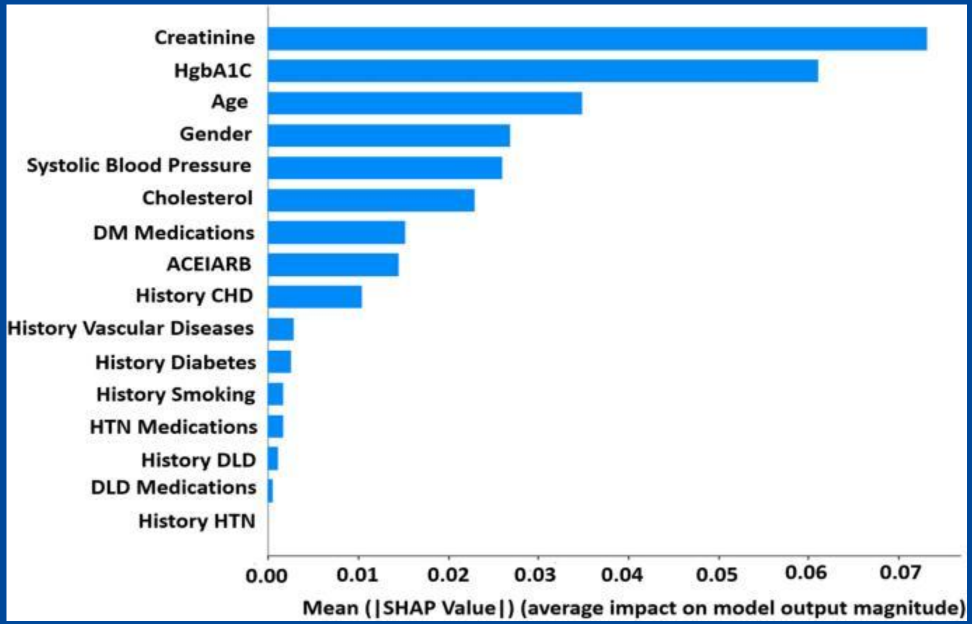
Potential interventions:
- Patient counseling
- Reminder calls
- Transportation support
- AE management
- Adherence support
- Additional site follow-up

Explainability: SHAP

Question
Why did the model classify this participant as high risk?

Potential influential variables:
- Frequent missed visits
- Low treatment compliance
- Early adverse events
- High disease severity
- Multiple concomitant medications

SHAP provides
- Global feature importance
- Direction of influence
- Individual-level explanation



From Prediction to Intervention

Without ML
Dropout occurs
→ Data become missing
→ Statistical consequences

With ML
Early prediction
→ Identify high-risk participants
→ Targeted intervention
→ Potentially reduce avoidable dropout
→ Improve trial efficiency

Key message
The value of ML is not merely predicting dropout—it is enabling timely action.

Criticality & Challenges

Critical Challenges

Overfitting: Model learns noise.

Data leakage: Future information enters model development.

Small sample size: Too many predictors relative to events.

Class imbalance: Few dropout events.

Site effects: Model learns site characteristics rather than participant risk.

Missing data: Incorrect handling may introduce bias.

Lack of external validation: Good internal performance may not generalize.

Biostatistical & Regulatory Considerations

Key considerations:

- Prespecification
- Statistical analysis plan
- Data provenance
- Reproducibility
- Version control
- Model validation
- External validation
- Explainability
- Performance monitoring
- Model updating/recalibration

Important distinction

An exploratory ML model for hypothesis generation is different from an ML model used to support a clinical or regulatory decision.

Key Takeaways

1. Start with the clinical question, not the algorithm.
2. Data quality is fundamental.
3. Establish a traditional statistical benchmark.
4. Prevent data leakage.
5. Control overfitting.
6. Use appropriate validation.
7. Assess discrimination and calibration.
8. Evaluate clinical utility—not AUC alone.
9. Use explainability where appropriate.
10. Integrate ML with biostatistical principles rather than treating ML as a replacement for biostatistics.

connect·share·advance

phuse

THANK YOU!

connect·share·advance

phuse

The Global Healthcare Data Science Community

Contact Channels

- office@phuse.global
- phuse.global

Social Media

- /company/phuse
- /phusetube
- /phuse_global
- /phusebook