Penn Medicine

An Introduction to Artificial Intelligence Clinical Prediction Models in Surgery

Mert Marcel Dagli, MD; Ali K Ozturk, MD; Jang W Yoon, MD; William C Welch, MD

June 10, 2024

Disclosures

None

CLINICAL PREDICTION MODELS

AI CLINICAL PREDICTION MODELS

OUTLINE DEVELOPMENT AND INTERNAL VALIDATION

AI MODELS FOR BINARY PROGNOSTIC CLINICAL PREDICTION MODELS

HYPERPARAMETER TUNING

PERFORMANCE EVALUATION

What are Clinical Prediction Models?

GENERAL

Definition: quantitative tools designed to aid healthcare professionals indiagnostic or prognostic decision-making based on individual patient data

Function: data synthesis from multiple predictors to estimate likelihood of having a specific disease or outcome, thereby informing clinical decision-making

Components: Typically, these models integrate patient demographics, clinical parameters, and laboratory results using statistical methods to predict health outcomes

How to Report Clinical Prediction Models?

GUIDELINES

Transparent Reporting of a multivariable prediction model for Individual Prognosis or Diagnosis (TRIPOD) Statement 2015: Clinical prediction models are poorly reported, indexed on EQUATOR Network with other guidelines, such as PRISMA

TRIPOD Guidelines: easily accessible 22-item checklist

Most studies utilize retrospective data: handling of missingness, reproducible and transparent development with internal validation, external validation, net benefit analysis are important to reach clinical application

Transition to AI Clinical Prediction Models?

Leap to AI: AI models can surpass traditional statistics

- complex data types (e.g., imaging, sequential/genetic data, etc.)
- recognizing patterns beyond human capacity
- large datasets with intricate variables
- potentially superior performance (benchmarking and comparisons)
- opaqueness (explainability vs black box)

Outline Development and Internal Validation of a Prognostic AI Clinical Prediction Model

Data Source: retrospective database

Step 1: Data cleaning, evaluation and handling of missingness (MCAR/MAR/MNAR?, multiple imputation – MICE (R) or Iterative Imputer with random states in Python)

Step 2: Correlation matrix

Step 3: Panel of professionals (clinician, statistician) & literature review for candidate predictor selection

- Advanced: can be done at later steps: non-linearity, cubic, polynomial feature engineering

Step 4: Building the models/pipeline

- Language of choice: Python

Step 5: Hyperparameter Tuning

Step 6: Internal Validation and Evaluation of Performance

Different Models for Prognostic Clinical Prediction Model (Binary Outcome)

Logistic Regression (LR)

- **Opaqueness**: Low. Direct and transparent

Support Vector Machine (SVM)

- **Function**: Separates groups with a clear boundary
- **Opaqueness**: Moderate. The use of kernels can obscure understanding

Random Forest (RF)

- **Function**: Combines multiple trees for robust predictions
- **Opaqueness**: Moderate. Individual trees are simple, but ensemble complexity adds some opacity

Gradient Boosting (XGBoost)

- **Function**: Sequentially corrects errors, improving accuracy
- **Opaqueness**: Moderate to high. Sequential complexity and interactions can be hard to trace

Explainable Boosting Machine (EBM)

- **Function**: Balances accuracy with interpretability
- **Opaqueness**: Low. Designed for transparency in predictions and feature importance. Excellent for re-evaluating continuous predictor cut-offs

Artificial Neural Network (ANN)

- **Function**: Mimics the brain, learning from data
- **Opaqueness**: High. "Black box" nature due to complex internal workings

Ensemble Boosted Artificial Neural Network (ebANN)

- **Function**: Improves via learning from sequential errors
- **Opaqueness**: High. Combining multiple ANNs increases complexity and reduces transparency

Hyperparameter Tuning – Part 1

STEP 5 – EXAMPLE ANN FOR PROGNOSTIC BINARY CLASSIFICATION

Hyperparameter Tuning Essentials

- Objective: Adjust key settings to enhance model's ability to learn and predict accurately.
- Data split: A 70-15-15 split for training, validation, and testing ensures we have enough data to learn from, validate adjustments, and independently test performance. Val enables early stopping to prevent overfitting and increase generalizability.

Core Parameters

- Learning Rate: Speed at which a model learns. Think of it as the pace of studying for an exam—too slow might not cover all topics, too fast might miss important details.
- Neurons per Layer: Controls the brainpower of the network. More neurons can grasp complex patterns, like a larger team solving a problem together.
- Number of Layers: Layers are like steps in problem-solving. More layers allow for more complex reasoning, from simple to sophisticated tasks. Black box in-between input and output (hidden).
- Dropout Rate: Prevents the model from relying too much on any one feature, like studying a broad range of subjects to do well on a comprehensive exam.

Hyperparameter Tuning – Part 2

STEP 5 – EXAMPLE ANN FOR PROGNOSTIC BINARY CLASSIFICATION

Adjusting for Performance

- Optimizers: Tools that help adjust the model's knowledge based on learning, akin to choosing the best study technique. F.e. AdamW incorporates weight decay to prevent overfitting.
- Criterion (Loss Function): Measures how far off predictions are from actual outcomes, guiding improvement like feedback on practice tests. Binary Cross Entropy (BCE) loss for binary tasks f.e.

Balance is Key: Too many neurons or layers might make the model memorize rather than learn. Aim for just enough complexity to accurately predict without memorizing.

Advanced: grid search, prespecified range that the model gets iterated on to find the best settings, resource-intensive (time, computing power)

Guiding Principle

- Iterative Testing: Start with baseline settings, then test and adjust. Learning from data should gradually improve, much like honing a skill over time.

Internal Validation and Performance Evaluation

STEP 6

Evaluation of confusion matrix to evaluate predictions (TN, TP, FN, FP) is a form of internal validation.

Internal validation with testing set (random split, temporal can be considered intermediate validation in-between internal and external). External validations are unfortunately very rare, but important for clinical application.

Selection of standard performance metrices derived from testing set:

- Sensitivity: True Positive Rate
- Specificity: True Negative Rate
- Discrimination AUC-ROC: model's ability to discriminate between classes
- F1-Score: Harmonic mean between precision and sensitivity (minimizing FP, max. TP)
- Precision (PPV): ratio of TP predictions : all positive predictions
- NPV: ratio of TN predictions : all negative predictions
- Yousef's optimal threshold: selecting the optimal cutoff for best performance

Internal Validation and Performance Evaluation

STEP 6

Evaluation of confusion matrix to evaluate predictions (TN, TP, FN, FP) is a form of internal validation.

Internal validation with testing set (random split, temporal can be considered intermediate validation in-between internal and external). External validations are unfortunately very rare, but important for clinical application.

Selection of standard performance metrices derived from testing set:

- Sensitivity: True Positive Rate
- Specificity: True Negative Rate
- Discrimination AUC-ROC: model's ability to discriminate between classes
- F1-Score: Harmonic mean between precision and sensitivity (minimizing FP, max. TP)
- Precision (PPV): ratio of TP predictions : all positive predictions
- NPV: ratio of TN predictions : all negative predictions
- Yousef's optimal threshold: selecting the optimal cutoff for best performance

Conclusions

Proper development, validation, and transparent reporting of AI Clinical Prediction Models is very important and complex

A spectrum of explainable and opaque models to develop the best predictive model, while also including information on reasoning is important and guides future research

External validation and net benefit analysis is recommended, but rarely done

The future of AI in healthcare will most likely be driven by complexity due to predictive power, but explainable AI (XAI) should not be disregarded

Penn Medicine