FluidAI

A Clinician-Integrated Framework for Developing Postoperative Complication Prediction Models with Machine Learning and EHR Data

Finn O'Hara[1], Trevor Yu[1], Nour Helwa[1]

1. 1. FluidAI Medical (Formerly NERv Technology Inc.), Kitchener, Ontario, Canada

Abstract

Machine learning (ML) models have the potential to enable earlier prediction of postoperative complications in gastrointestinal surgeries, such as anastomotic leak (AL), bleeding, and sepsis. Developing ML models using retrospective electronic health record (EHR) data can be disconnected from clinical workflows and risks producing algorithms that are not actionable in practice. We present a structured, clinician-guided methodology for building binary classification risk prediction models that integrates clinical input at each stage. By emphasizing actionability, patient safety, and clinician usability alongside technical performance, it addresses key barriers to adoption of ML in clinical practice. Future work will apply this framework to multi-institutional EHR datasets for GI surgery.

1 Cohort Definition

Clinicians provide initial definitions and help adjudicate edge cases

Establishing a target cohort defines exactly who the model applies to, such as patients undergoing a specific surgery or individuals newly diagnosed with a disease. The index date is set to the date of the surgery of interest.

Methods

1. **Rule-Based Design:** Cohorts are assembled using concept sets built from standardized codes such as ICD-10 or CPT.
2. **NLP and Text Mining:** Unstructured clinical text, including progress notes, surgical notes, etc. is scanned to extract specific medical concepts. These concepts are then mapped into a standardized ontology so they can be consumed by the model
3. **Probabilistic Design:** A machine-driven approach in which supervised ML learns the defining attributes of a cohort from a set of labeled examples.

2 Model Framing

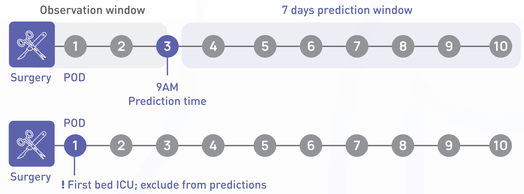
Clinicians define the scope and timing of the prediction problem

Model framing translates the clinical question into a structured prediction problem for a machine learning model. Clinicians help establish the observation window (the period of historical data used as model input) and the prediction window (the future time period in which the outcome is expected to occur).

Methods

1. **Defining the Prediction Structure:** Frame the question as: Among [target cohort], who will develop [outcome] within [prediction window]?
2. **Selecting the Clinical Use Case:** Determine whether the model predicts disease onset or progression, need for treatment/intervention, need for further workup, etc.
3. **Setting Temporal Constraints:** Define when the prediction window begins relative to the index date. Require a minimum observation window to ensure patients have sufficient follow-up time.
4. **Setting Exit Events:** Define events after which predictions are no longer valid (e.g. discharge, ICU escalation) to avoid spurious predictions during non-actionable clinical contexts.

Example AL model framing



3 Label Definition

Clinicians define relevant events indicative of complications

Label definition determines the outcome status for each patient within the prediction window. Patients who experience the outcome event (e.g. diagnosis, intervention) are assigned a positive label while those who do not are assigned a negative label. Multiple methods can be used in combination to cross-validate labels and strengthen confidence in their accuracy.

Methods

1. **Code-Based Phenotyping:** Administrative and billing codes (e.g., ICD, CPT) extracted from the EHR are used to flag the presence of an outcome.
2. **NLP-based Labelling:** NLP algorithms extract outcome-relevant concepts from unstructured clinical notes.
3. **Chart Review:** Clinical experts manually review patient charts to establish a gold-standard label.

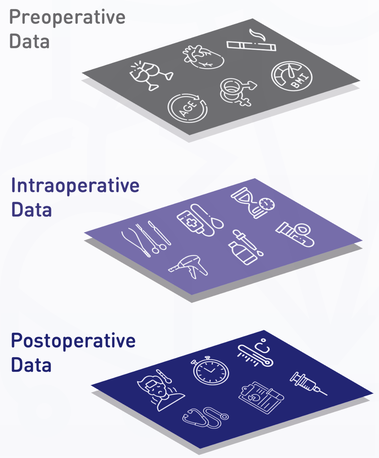
4 Feature Selection

Clinicians suggest features relating to the pathophysiology of the complication

This step identifies and selects the patient-level variables that will serve as inputs to the model. The features available for selection are determined by the observation window and the clinical context defined in previous steps. Depending on the prediction tasks features may be strictly preoperative, intraoperative, postoperative, or a combination.

Methods

1. **Filter and Wrapper Methods:** Filter methods use statistical tests (e.g., hypothesis testing, correlation) to pre-select features based on a significance cutoff. Wrapper methods evaluate different feature subsets iteratively through forward selection or backward elimination.
2. **Implicit Feature Selection:** Certain algorithms perform feature selection during training. Regularized regression (e.g., LASSO) shrinks unhelpful coefficients to zero, while tree-based methods select features based on their importance in splitting the data.
3. **Missingness Indicator:** Missingness itself can be a clinically meaningful feature, the absence of a measurement may reflect a provider's clinical judgement. Absence of a feature can be encoded as a predictor. When missingness is not informative, imputation methods such as mean/median replacement, k-nearest neighbors, or multiple imputation via Bayesian regression can be used to fill in gaps.



5 ML Experimentation

Clinicians verify interpretability of algorithms

This step involves selecting and fitting the algorithms that will learn the relationship between the features and the label. Because no single algorithm works best for all data multiple methods are typically evaluated. Each algorithm is further optimized through hyperparameter tuning, where grid searches identify the optimal settings (e.g., number of trees, learning rate) for each method.

Methods

1. **Regularized Logistic Regression:** Penalizes model complexity to prevent overfitting and handles large feature sets well.
2. **Tree-Based Ensembles:** Random Forests and Gradient Boosting Machines (e.g., XGBoost, AdaBoost) combine multiple decision trees to improve accuracy and handle non-linear relationships natively.
3. **Deep Learning / Neural Networks:** Architectures such as Multilayer Perceptrons or augmented Gated Recurrent Units are particularly adept at handling sequential or time-series data with missingness.
4. **Interpretable Algorithms:** Methods such as SHAP are used to interpret feature importance to predictions and validated with clinicians.

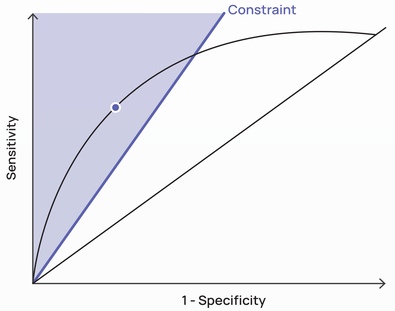
6 Threshold Selection

Clinicians inform tolerance for error, sensitivity, and specificity

Machine learning models typically output a predicted probability: a risk score from 0 to 1. Threshold selection is the process of choosing the cutoff value at which a patient is classified as high-risk (predicted positive) versus low-risk (predicted negative). Clinician tolerance for errors may introduce performance constraints on feasible thresholds.

Methods

1. **Youden's J Index:** A statistical method that identifies the threshold yielding the highest sum of sensitivity and specificity, representing the optimal balance point on the ROC curve.
2. **Clinical Utility Trade-offs:** The threshold is selected based on the real-world costs of a false positive versus a false negative.
3. **Exploring Multiple Cutoffs:** Performance characteristics are calculated across a wide range of thresholds (e.g., 0.4, 0.5, 0.6) to provide a comprehensive view of how the sensitive the model is to threshold selection.



7 Validation

Clinicians conduct usability testing

This step measures how accurately the finalized model predicts outcomes on new, unseen data to ensure it generalizes well and is not overfit to the training data. In addition to model validation, user workflow validation is conducted to ensure model outputs are clinically actionable.

Methods

1. **Internal Validation (Data Splitting):** The dataset is divided into a training set and a holdout test set (e.g., 75/25). This is often done with a temporal split, where the model trains on older data and is tested on newer data. Cross-validation is also used to repeatedly leave out subsets during training.
2. **External Validation:** The finalized model is tested on an entirely different database, healthcare system, or geographic location to demonstrate its generalizability to other sites.
3. **Performance Metrics:** Discrimination is evaluated using the Area Under the ROC Curve (AUROC), Area Under the Precision-Recall Curve (AUPRC), Sensitivity, Specificity, Positive Predictive Value (PPV), and Negative Predictive Value (NPV).
4. **Calibration Evaluation:** Patients are grouped by their predicted risk (e.g., into deciles) and plotted against the actual observed incidence of the outcome to ensure the model assigns numerically accurate risk scores.