

Explainable Estimation of Clinical Pathway Branching Factors from Multi-Institutional Electronic Medical Records

Miwa Sugitani¹⁾, Ryosuke Matsuo¹⁾, Tomoyoshi Yamazaki¹⁾, Kenji Araki¹⁾,
Masato Oguchi¹⁾, Haruo Yokota²⁾, Hieu Hanh Le¹⁾

1) Ochanomizu University 2) Josai University

ISMICT 2026

Contents

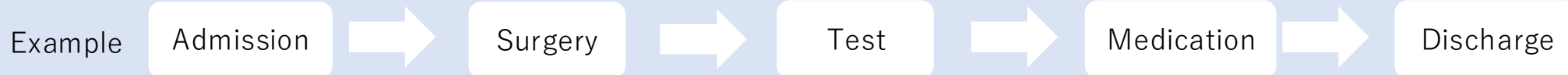
- Research Background
- Related Work: Issues and Our Goal
- Proposed Method
 - Step 1: Evaluating and Selecting Patterns
 - Step 2: Estimating Branching Factors (Statistical / Learning-based)
- Evaluation Experiments and Results
 - Pattern Accuracy / Model Validity / Clinical Evaluation of Factors
- Conclusion and Future Work

Research Background

Why Medical Data Analysis Matters

- Reusing electronic medical record (EMR) data speeds up analysis of care processes, helping make care more consistent
- To standardize care, hospitals widely use clinical pathways

Clinical pathway : a care plan that lists the standard tests, surgeries, and medications for patients with a given disease, in time order



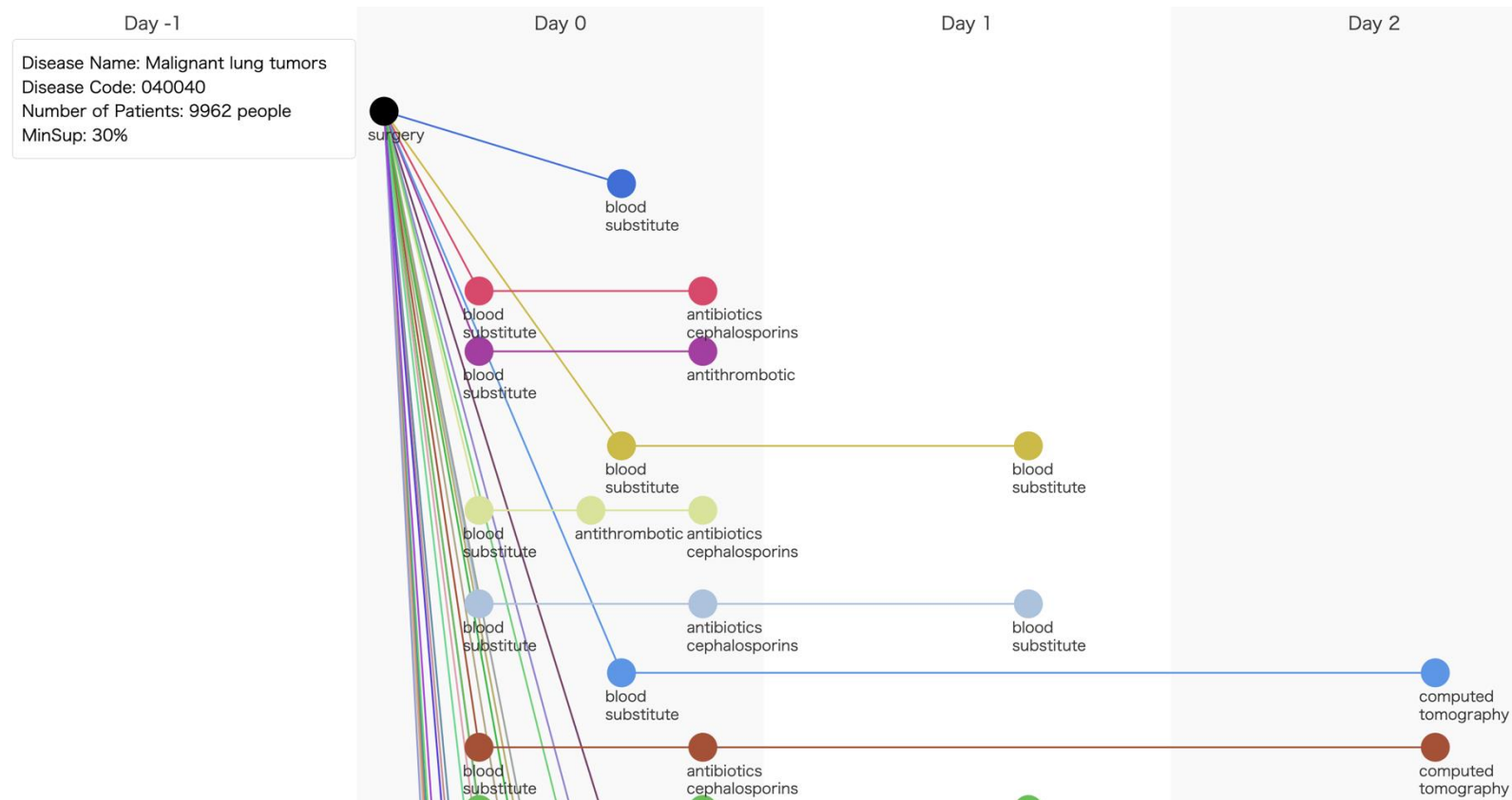
But Real Care Often Branches from the Standard

- Care processes differ across patients and hospitals [1]
- So beyond standardization, we need to identify exactly where and why care branches from the standard. Quantifying these **branching factors** is the goal of this study

Related Work (1/2)

Multi-Institutional Analysis and Variance Analysis of Clinical Pathways

- M. Sugitani, et al. (2025): used sequential pattern mining (SPM) to extract and visualize frequent medical instruction patterns from multi-institutional EMR data
- The number of extracted results is huge, with no objective standard for selecting which branches to analyze



Related Work (2/2)

Factor Estimation Using Logistic Regression

- Y. Hua, et al. (2025): reviewed how logistic regression, with odds ratios, gives interpretable clinical predictions
- Y. Guan, et al. (2022): added regularization (ridge regression) to stabilize risk prediction when variables are strongly correlated
 - But these methods assume a linear relationship, so they may miss non-linear thresholds and interactions that can arise in clinical decision-making

➔ This shows the need for a more explainable approach to factor estimation

Violin Plot of Predicted Probabilities by True Class

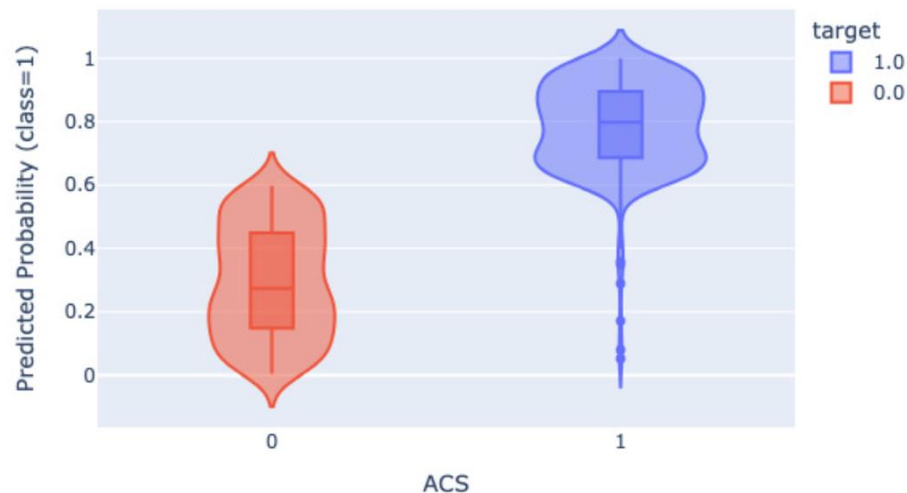
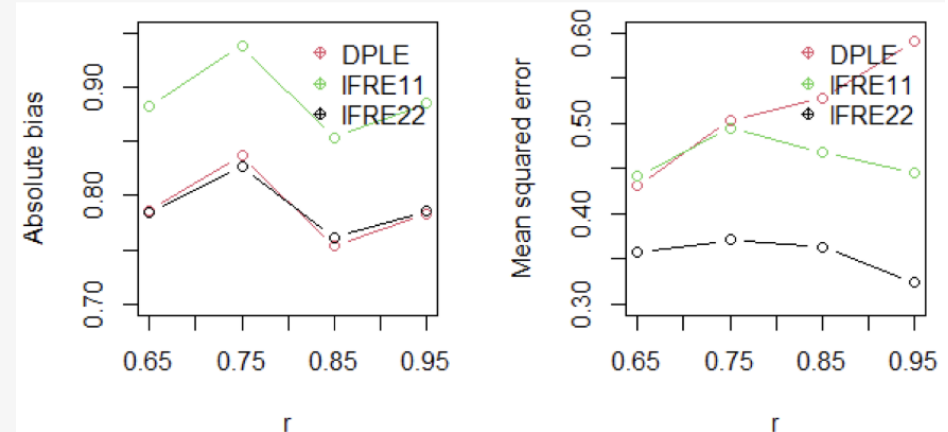


Figure 2. Plots of the bias and MSE for four different scenarios which combine a fixed sample size $n = 40$ with a variational correlation coefficient obtained by DPLE, IFRE11 and IFRE22.



Objective of This Study

Estimate the branching factors of frequent medical instruction patterns in an explainable manner, based on their consistency with the **standard clinical pathway**

Step 1: Objective Target Selection

F-measure quartiles are used to objectively and comprehensively identify branches that are, or are not, consistent with the standard clinical pathway

Step 2: Comparing Factor-Estimation Approaches

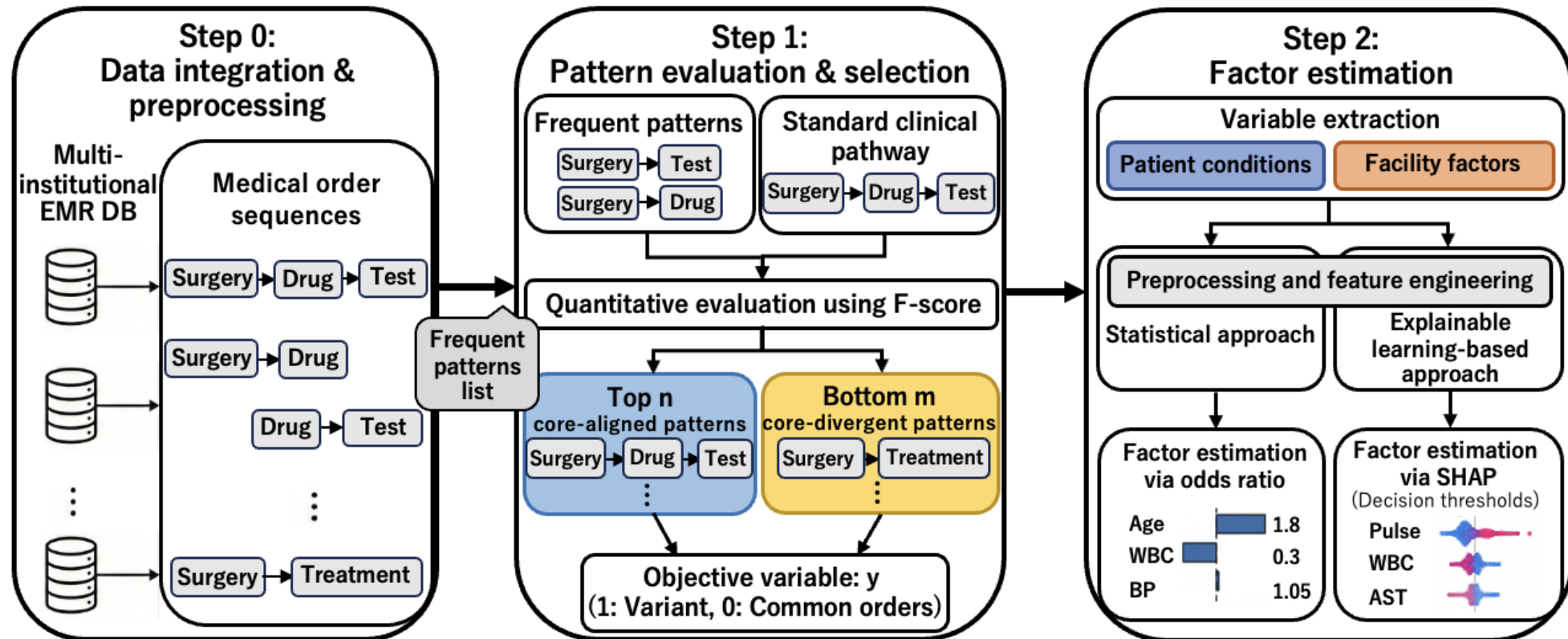
Statistical (L2-regularized logistic regression) vs. explainable learning-based (predictive model + SHAP) approaches are compared on real data, to see which extracts branching factors more effectively

Proposed Method: Overview

Step 0: Use sequential pattern mining (SPM) to extract frequent order patterns from multi-institutional EMR data

Step 1: Compare patterns with the standard clinical pathway to select which branches to study

Step 2: Estimate the factors behind each selected branch



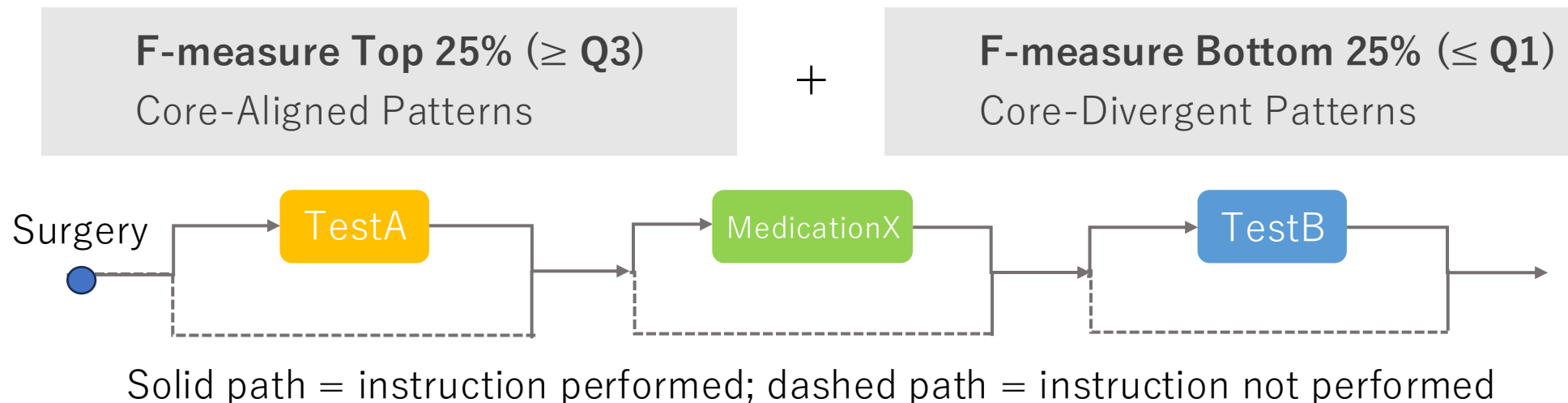
Step 1: Evaluating and Selecting Patterns

① Quantitative Evaluation Against the Standard Clinical Pathway

Measure how well each pattern matches the standard clinical pathway (the minimum necessary medical instructions defined by clinicians), using the F-measure (the harmonic mean of precision and recall)

② Selecting Analysis Targets (Objective Variable y)

Using F-measure quartiles, select the top 25% ($\geq Q3$) and bottom 25% ($\leq Q1$) patterns per disease as candidates, requiring at least 100 negative cases and a positive rate of 95% or less. Then split patterns by instruction, excluding common items (e.g., surgery), and set the rest as the objective variable $y \in \{0, 1\}$.



Step 2: Estimating Branching Factors

① Extracting Variables and Preprocessing

- Build explanatory variables from four types: static, dynamic, delta features, and facility information

② Comparing Factor Estimation Using Two Approaches

- To see which method extracts factors closer to real clinical judgment for complex medical data (strong correlation, non-linearity), we run the following two approaches

Statistical Approach

Feature: captures linear relationships between variables, a common method in medical research

Expected: extraction of global, universal risk factors via odds ratios

Explainable Learning-Based Approach

Feature: captures non-linear relationships and complex interactions between variables

Expected: extraction of real **non-linear decision thresholds**, such as "add a test once a value exceeds X"

① Extracting Variables and Preprocessing

Type of Explanatory Variable	Description
Static information (patient-specific)	Age, sex, BMI class (sex and BMI class are one-hot encoded)
Dynamic information (patient status)	Vital signs (temperature, blood pressure, pulse) and lab values (white blood cell count, red blood cell count, platelet count, hemoglobin, CRP, creatinine, etc.)
Delta features	Automatically generate day-to-day and whole-period differences to capture change over time
Facility information	Total patient count and disease-specific patient count at each facility, entered as continuous values

Preprocessing Policy

- Lab values and vital signs are not discretized. They are **entered as continuous values** to reduce information loss
- To preserve causal time order, only data **up to the day before the branch date** is used

②-1 Statistical Approach: L2-Regularized Logistic Regression

Statistical Model: L2 Regularization (Ridge)

- Lab values are often strongly correlated, which destabilizes coefficients. L2 regularization stabilizes them, separating patient effects from facility effects

Evaluating Factors

- Factors are judged mainly by their odds ratio
- Group sizes are also reported, to catch odds ratios inflated by small groups

②-2 Explainable Learning-Based Approach: Predictive Model + SHAP

Interpretable Learning-Based Model: Predictive Model + SHAP

SHAP [2] : explains how much each feature pushes a prediction up or down

- The predictive model captures linear and non-linear effects and interactions between features; SHAP then quantifies each feature's contribution

Two Key Ideas

- Missing values are not imputed; **not being measured is itself treated as information**
- The point where a SHAP value's sign reverses is treated as a candidate **decision threshold**
- For the predictive model, we compared logistic regression (LR), random forest (RF), and XGBoost

Evaluation Experiments: Purpose and Design

Purpose

- Test whether we can **explainably estimate the factors** behind branches selected for consistency with the standard clinical pathway

Experiment 1: Selecting Patterns for Studying

- Evaluate pattern accuracy and pattern extraction's stability across minSup settings

Experiment 2: Validating the Predictive Models

- Compare LR, RF, and XGBoost by AUC (Area Under the Curve) with 5-fold stratified cross-validation

Experiment 3: Estimating Branching Factors and Clinical Evaluation

- Case study on gastric malignancy, comparing factors extracted by the statistical and learning-based approaches
- Clinical specialists judge the extracted factors on a 3-point scale (○ Valid / △ Indirect / × Invalid)

Evaluation Experiments: Real Dataset

Real Data Overview

- EMR data from 27 institutions, collected 2015–2024 under the Next-Generation Medical Infrastructure Act
 - Anonymized, with consistent terminology
- 9 diseases, 58,598 patients total
- Case study for factor estimation: **5,961 patients with gastric malignancy**
 - Cases from 24 institutions (median 221 patients, range 4–983)
- Reviewed and approved by the Purpose-of-Use Review Committee of the Life Data Initiative (No. 2024_MIL_0004_A001)

Patients per Disease

Disease	Patients (n)
Lung malignancy	9,962
Acute myocardial infarction	4,874
Angina pectoris, chronic ischemic heart disease	11,148
Gastric malignancy	5,961
Liver and intrahepatic bile duct malignancy	2,030
Intervertebral disc degeneration, hernia	1,194
Breast malignancy	12,818
Bladder tumor	9,748
Cervical and uterine body malignancy	897
Total	58,598

Evaluation Experiment 1: Pattern Accuracy

Evaluated across all 9 diseases (minSup = minimum support [%]; values are the 9-disease average)

High Precision (0.74–0.97)

- Extracted patterns rarely include instructions outside the standard clinical pathway, showing few irrelevant results
- At minSup = 40%, core processes like surgery and IV drips are extracted stably

Recall Challenge (0.52–0.54)

- Date shifts from anonymization cause standard Day 1 instructions to appear as Day 2–3 in the real data
- Correcting this date shift fully remains future work



minSup	Patterns	Recall	Precision	F-measure
70	48	0.52	0.97	0.67
60	107	0.52	0.93	0.65
50	286	0.53	0.86	0.64
40	853	0.53	0.79	0.62
30	1,547	0.54	0.74	0.61

- We adopt minSup = 40%; these patterns become the input for Experiment 2's predictive models and Experiment 3's factor estimation

Evaluation Experiment 2: Validating the Predictive Models

Comparing 3 Models Across 9 Diseases and 81 Targets (5-fold Stratified Cross-Validation, AUC)

Day Group	Targets (N)	Logistic Regression	Random Forest	XGBoost
Day -1	6	0.812	0.952	0.958
Day 0	17	0.985	0.999	0.998
Day 1	38	0.958	0.995	0.996
Day 2	9	0.859	0.986	0.988
Day 3	6	0.942	1.000	1.000
Day 4	1	0.828	0.999	0.999
Day 7	4	1.000	1.000	0.999
Average	81	0.941	0.992	0.993

Results and Discussion

- LR averages 0.941; RF and XGBoost both exceed 0.99
- The gap is largest in Day groups with strong non-linearity
 - LR drops as low as 0.81–0.86
 - XGBoost stays at 0.95–1.00
- **Interactions and non-linear effects matter**

XGBoost adopted for factor estimation

RF and XGBoost have similar accuracy. XGBoost was chosen for its missing-value handling and its better fit with SHAP-based threshold extraction.

Evaluation Experiment 3: Validity Assessment by Clinical Specialists

Clinical specialists judge the extracted factors on a 3-point scale (○ Valid / △ Indirect / × Invalid)

Factor-Estimation Method	Factors Extracted (Targets)	○ Valid	△ Indirect	× Invalid	Validity Rate
Statistical: L2-Regularized Logistic Regression	6 (4)	0	3	3	0 %
Learning-Based: XGBoost + SHAP	35 (8)	16	6	13	45.7 %

Limits of the Statistical Approach

- All extracted factors are "downward trends"
- Low-abnormal groups are small, causing quasi-complete separation

Benefits of the Learning-Based Approach

- Far more medically supported findings
- Invalid cases are mostly due to confounding from routine tests

→ Validity rate: learning-based **45.7%**, statistical **0%**, confirming our method's effectiveness

Evaluation Experiment 3: Factor Estimation with the Statistical Approach

Gastric malignancy: extracted **6 factors** (from 4 of 9 targets) satisfying $|\beta^*| \geq 0.9$ and $p < 0.05$

Target	Related Factor	Normal-trend group n (rate)	Low-abnormal-trend group n (rate)	Odds Ratio	p-value	Verdict
Day -1 / Renal Function Test	Creatinine, downward trend	125 (2%)	2 (50%)	4.89	0.047	×
Day -1 / Renal Function Test	TP, downward trend	75 (3%)	1 (100%)	2.50	0.039	△
Day -1 / Blood Count Test	Creatinine, downward trend	125 (2%)	2 (50%)	5.08	0.047	×
Day -1 / Blood Count Test	TP, downward trend	75 (3%)	1 (100%)	2.54	0.039	△
Day 1 / Infection Test	CRP, downward trend	1,737 (99%)	81 (94%)	23,940.54	0.009	×
Day 1 / Blood Count Test	CRP, downward trend	1,737 (99%)	81 (95%)	26,279.61	0.014	△

(CRP: C-reactive protein; TP: total protein)

All extracted factors are "downward trends"

- The low-abnormal-trend groups have only 1–81 patients, causing **quasi-complete separation**
- CRP's odds ratio inflates past 20,000 (implementation rates are similar: 99% vs. 94%)

Judgment by Clinical Specialists

- ○ Valid: 0 / △ Indirect: 3 / × Invalid: 3 → **Validity rate: 0%**
- "Low creatinine is never a problem"; "An infection test the day after surgery is not normally expected"

Evaluation Experiment 3: Factor Estimation with the Learning-Based Approach

Gastric malignancy: extracted **35 factors** from 8 targets (Day 1 / Renal Function Test: 6 shown)

Related Factor	Decision Threshold	Below-threshold group n (rate)	At-or-above-threshold group n (rate)	Odds Ratio	p-value	Verdict
Pulse Δ (Day -1 $\rightarrow$ Day 0)	26.37 bpm	3,329 (62%)	178 (38%)	2.61	< 0.001	○
Diastolic BP (Day 0)	68.57 mmHg	553 (49%)	3,277 (61%)	0.59	< 0.001	○
Diastolic BP Δ (Day -1 $\rightarrow$ Day 0)	1.00 mmHg	1,755 (55%)	1,869 (62%)	0.76	< 0.001	○
Systolic BP (Day 0)	123.53 mmHg	685 (52%)	3,142 (61%)	0.68	< 0.001	○
Systolic BP Δ (Day -1 $\rightarrow$ Day 0)	2.74 mmHg	1,699 (53%)	1,913 (64%)	0.63	< 0.001	△
Alb (Day 0)	3.00 g/dL *	22 (86%)	1,490 (96%)	0.24	0.014	×

(BP: blood pressure; Alb: albumin)

* No SHAP sign-reversal point was found, so the lower limit of the normal range was used as the threshold

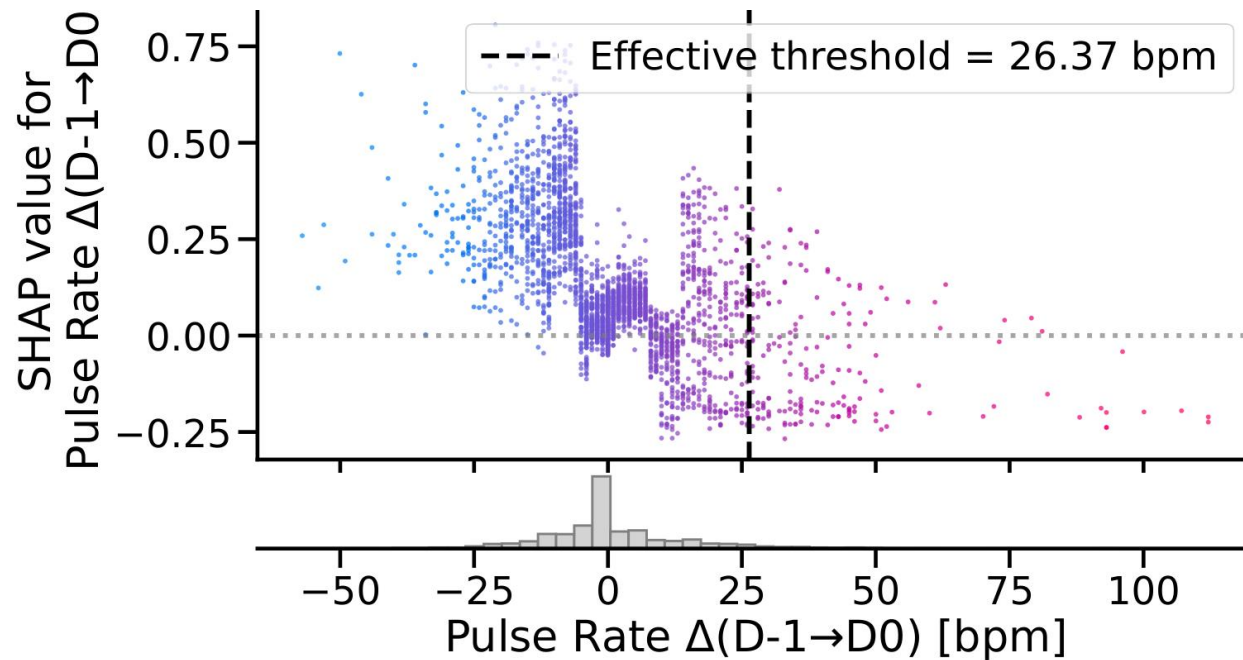
Blood Pressure and Pulse Drive the Decision

- Of the 6 factors, **5** relate to blood pressure or pulse, and **4** were judged valid

Causes of Invalid Cases

- For Alb, no SHAP threshold was found, and the class imbalance is severe (22 vs. 1,490 patients)
- Confounding from routine tests is the main cause of invalid cases

Discussion ①: Extracting Decision Thresholds with SHAP



Dependence plot for the change in pulse Δ (Day -1 $\rightarrow$ Day 0),
for the Day 1 / renal function test

Model-Derived Sign-Reversal Point

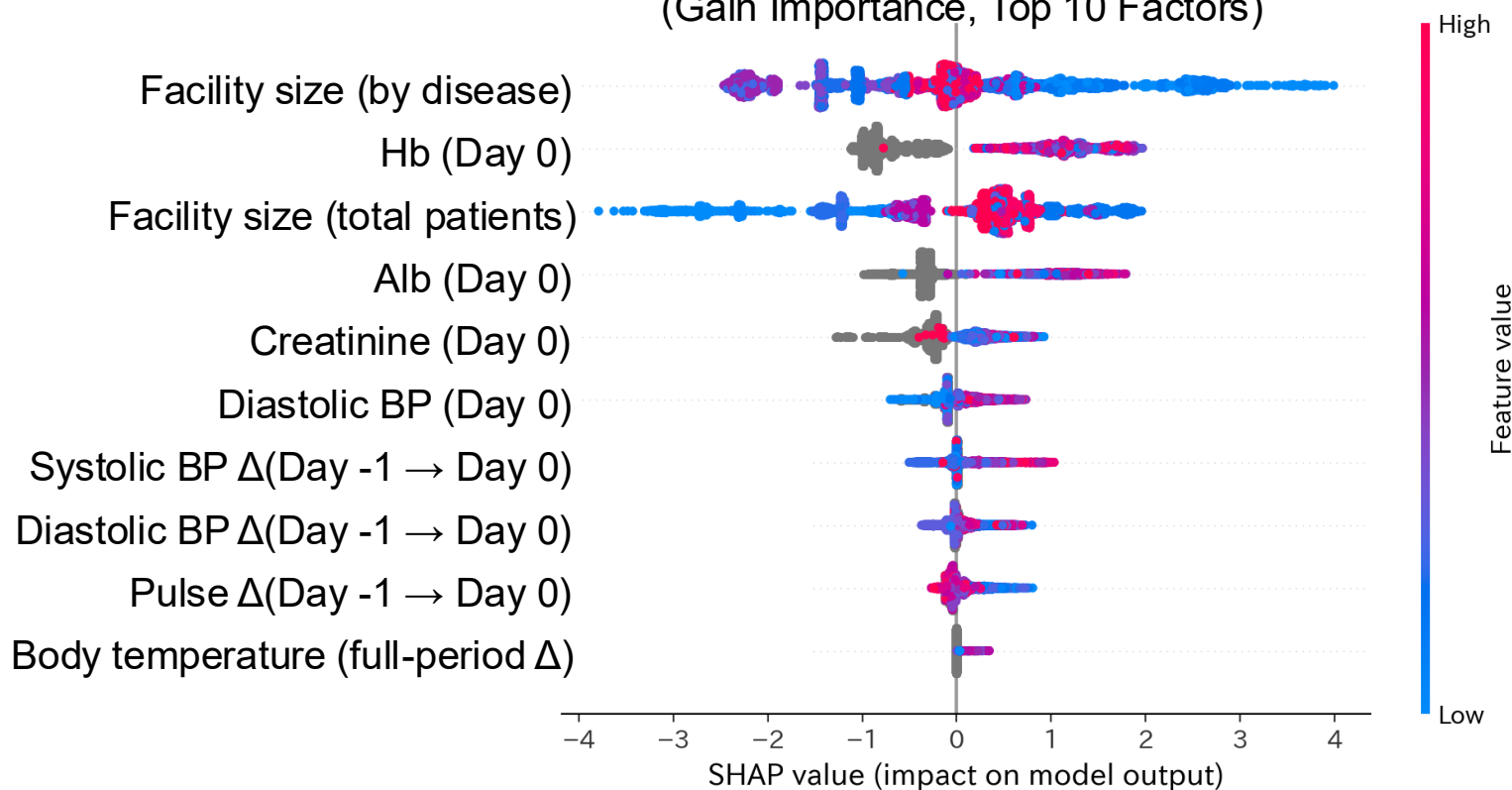
- SHAP contribution reverses at **26.37 bpm**
- Implementation rate: **62% below** (n=3,329) vs. **38% at or above** (n=178)

Clinical Interpretation by Specialists

- A pulse change of 26.37 bpm is clinically substantial. Patients with hypertension or arteriosclerosis may exhibit large perioperative pulse fluctuations, supporting the clinical plausibility of pulse change as a factor associated with this decision.
- **The non-linear threshold** was presented in a clinically interpretable way

Discussion ②: The Dominant Effect of Facility Factors and a Two-Layer Structure

Overall Trend Analysis: Post-Op Day 1 Renal Function Test
(Gain Importance, Top 10 Factors)



Facility Factors Rank Highest

- Among the top 10 factors by Gain Importance, **facility size** (disease-specific and total patient counts) ranks highest
- Within the top 10 for all 9 targets; total patient count ranks first for 6 targets

Discovering a Two-Layer Structure

- Facility factors: **wide-ranging**, acting globally
- Blood pressure and pulse: sign varies by patient, **acting locally** to adjust the decision
- This insight was possible only with multi-institutional data from 27 hospitals, though facility size may partly reflect unmeasured case-mix differences, which is future work to separate out

Conclusion and Future Work

Conclusion

- Proposed a method that objectively selects branch points using F-measure quartiles, and estimates their factors with two approaches: a statistical model and an explainable learning-based model (XGBoost + SHAP, average AUC 0.993)
 - In a case study of 5,961 gastric malignancy patients, the validity rate was **learning-based 45.7%, statistical 0%**, with interpretable thresholds such as a 26.37 bpm pulse change
 - Found that facility factors rank among the top predictors across all 9 targets, a finding unique to multi-institutional data

Future Work

- Model the admission year to separate pathway revisions from facility factors
- Estimate surgical procedures from order sequences to recover detail lost to anonymization
- Extend detailed factor estimation to the remaining 8 diseases, with continued clinician evaluation