

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

Miwa Sugitani*, Ryosuke Matsuo*, Tomoyoshi Yamazaki*, Kenji Araki*,
Masato Oguchi*, Haruo Yokota[†], and Hieu Hanh Le*

*Ochanomizu University, Tokyo, Japan

{g2120520, oguchi, le}@is.ocha.ac.jp, matsuo@ldi.or.jp, yamazaki.cp@gmail.com, araki6925@gmail.com

[†]Josai University, Tokyo, Japan - yokota.h.aa@gmail.com

Abstract—Analyzing clinical processes from electronic medical record (EMR) data is essential for improving and standardizing the quality of medical care. As these processes are not uniform across patients and institutions, identifying deviations from the standard process and their underlying factors is important. However, conventional factor estimation has been performed mainly within a single institution, with no framework for objectively selecting analysis targets across institutions. Moreover, owing to its linearity assumption, the widely used logistic regression struggles to capture decision thresholds and nonlinear interactions. This paper proposes a method for explainable estimation of the branching factors of frequent medical order patterns. It objectively selects branching points by their F-measure consistency with a standard clinical pathway and estimates factors through two approaches: a statistical one based on L2-regularized logistic regression, and an explainable learning-based one that combines a predictive model with SHapley Additive exPlanations (SHAP). In a case study on 5,961 patients with gastric malignancy from a 27-institution dataset, the combination of a predictive model with SHAP identified factors contributing to the prediction outcomes in an explainable manner, achieving a validity rate of 45.7% in clinical specialists’ judgments versus 0% for logistic regression, thereby confirming the effectiveness of the proposed method.

Index Terms—Clinical Pathway, Electronic Medical Records, Explainable AI, Health Care, Interpretability, Logistic Regression, Medicine, SHAP, XGBoost.

I. INTRODUCTION

In recent years, the secondary use of electronic medical record (EMR) data for analyzing medical processes has advanced, and is expected to improve the uniformity of healthcare quality and the effectiveness of treatment. Achieving this requires the standardization of clinical care, for which the clinical pathway is widely used. It is a care plan that chronologically organizes the standard examinations, surgeries, medications, and other interventions for inpatients with a given disease. However, as prior work [1] has pointed out, clinical processes are not uniform across patients and institutions, and support tailored to each patient’s condition is important in addition to standardization. Therefore, quantitatively identifying the

factors that cause branching from the standard clinical pathway is an essential task for both standardization and individualized support. Studies on identifying clinical factors using logistic regression have been conducted, such as estimating deviation factors from a pathway using multi-institutional data [2]. However, no process has been established for objectively identifying clinically meaningful branches among the vast number of frequent patterns spanning multiple institutions.

In addition, logistic regression, widely used for factor estimation, has two problems. First, owing to multicollinearity specific to medical data [3], such as severe cases concentrated at particular institutions (e.g., tertiary referral or university hospitals), it is difficult to separate whether a branch stems from patient condition or institutional characteristics. Second, because it assumes linearity, it struggles to capture the nonlinear decision thresholds and feature interactions physicians rely on. Interpretable AI combining predictive models with SHapley Additive exPlanations (SHAP) [4, 5] can overcome this limitation.

To address these problems, this study proposes a method for estimating the branching factors of frequent medical order patterns [6, 7] extracted from the EMR data of multiple medical institutions. The proposed method is novel in two respects: (1) it objectively selects the target branches based on the quartiles [8] of the F-measure consistency with the standard clinical pathway, addressing the lack of an objective branch selection criterion in prior work; and (2) it examines both a statistical approach (L2-regularized logistic regression [9]) and an explainable learning-based approach (predictive model with SHAP [4, 10]) for estimating branching factors, and the approach that extracts branching factors more effectively is evaluated on real data. The explainable learning-based approach, in particular, can capture both linear and nonlinear factor learning and thus overcome the linearity limitation of conventional logistic regression.

In a case study on 5,961 patients with gastric malignancy, drawn from 58,598 patients across 27 institutions, the predictive model with SHAP attained a validity rate of 45.7% in clinical specialists’ judgments versus 0% for

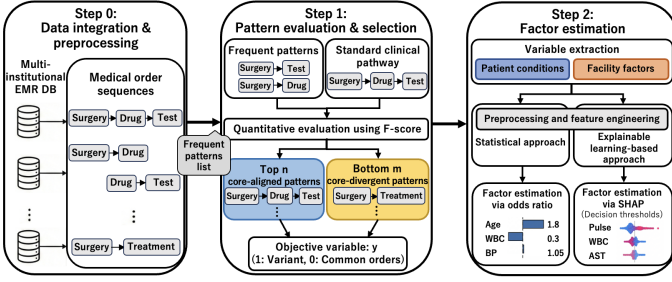


Fig. 1. Overview of the proposed method.

logistic regression.

II. RELATED WORK

We review prior work from three perspectives: (1) factor estimation using logistic regression, (2) multi-institutional and clinical pathway variance analysis, and (3) clinical applications of explainable machine learning.

Hua et al. [11] review the usefulness of logistic regression in clinical prediction models and its interpretability through odds ratios. In addition, Guan et al. [12] propose penalized estimation including ridge regression that simultaneously mitigates the separation and multicollinearity of logistic regression, effective for stabilizing risk prediction with strongly correlated explanatory variables. However, because these logistic-regression-based methods assume linearity, they have difficulty capturing the nonlinear feature interactions and decision thresholds physicians use.

Sugitani et al. [6, 7] proposed a method for extracting frequent medical order patterns across institutions, but the extraction results are enormous, and no objective criterion has been established for selecting the branches to analyze. Tou et al. [13] predicted pathway variances after lung cancer surgery using machine learning, but their study was limited to single-institution data and did not address institutional factors.

Maruyama et al. [5] predicted gastrectomy length of stay with XGBoost and SHAP [4, 10, 14] but did not address per-branch decisions. These studies use SHAP only to explain prediction contributions, without extracting sign-reversal points as decision thresholds or integrating with statistical factor estimation. In particular, Hur et al. [15] noted SHAP’s interpretive limitations.

Despite these advances, no prior study has objectively selected clinically meaningful branches from a standard clinical pathway while comparing statistical and explainable learning-based factor estimation. This study addresses this gap (Section III).

III. PROPOSED METHOD

The proposed method comprises two main steps (Fig. 1), preceded by Step 0, which integrates and preprocesses the multi-institutional EMR data into medical order sequences. Branching points are first selected by comparison with a standard clinical pathway (Section III-A), and their factors are then estimated by two approaches (Section III-B).

A. Pattern Evaluation and Selection (Step 1)

1) *Definition of the Standard Clinical Pathway*: The standard clinical pathway CP_{std} is the reference for comparison with frequent medical order patterns. Several clinical specialists defined the minimum necessary perioperative medical orders for each elapsed day, aligned with the medical order categories used in this study. CP_{std} is neither an institution-specific pathway nor a society-issued standard—both of which are revised as new drugs and techniques appear—but a minimum, version-independent reference; deviation from it does not imply inappropriate care. To compensate for anonymization-induced date shifts, we also use a shortened clinical pathway CP_{short} , removing the earlier day of consecutive-day duplicate orders.

Definition 1 (Standard and shortened clinical pathways). *The standard clinical pathway CP_{std} is a sequence of pairs of an execution day d and a medical order a :*

$$CP_{std} = \langle (d_1, a_1), \dots, (d_m, a_m) \rangle$$

When the same order appears on two consecutive days, the shortened version CP_{short} , which excludes the pair on day d , is defined as:

$$CP_{short} = \langle (d, a) \in CP_{std} \mid \neg IsDuplicateEarly(d, a) \rangle$$

The consistency of an extracted frequent pattern P_{freq} is measured by the F-measure $F = 2PR/(P + R)$. Here, the precision P is the proportion of medical orders in the pattern that match CP in both occurrence day and type, and the recall R is the proportion of the medical orders in CP that are included in the pattern. Each pattern is compared with both CP_{std} and CP_{short} , and the higher F-measure is adopted.

2) *Criteria for Selecting Branching Points*: Based on the quartiles [8] of the F-measure, the top 25% ($\geq Q_3$) and bottom 25% ($\leq Q_1$) patterns for each disease are taken as candidate targets, covering both within-standard and out-of-standard branches. To ensure learnability, we restrict the analysis to patterns with at least 100 negative cases and a positive rate of at most 95%, avoiding extreme imbalance. From the selected patterns, items such as surgeries common to all patients are excluded, and the remaining item is used as the objective variable $y \in \{0, 1\}$.

B. Factor Estimation (Step 2)

1) *Variable Extraction*: The explanatory variables x are shared by both methods and comprise the following four types, input as continuous values except one-hot encoded sex and BMI class. In this paper, the day of surgery is denoted Day 0, and the preceding and following days are expressed in relative days (Day $-d$, Day d).

- **Static information**: age, sex, and body mass index.
- **Dynamic information**: vital signs (temperature, blood pressure, pulse) and laboratory test values such as white blood cell count, red blood cell count, platelet

count, hemoglobin, C-reactive protein (CRP), and creatinine, up to the day before the branch.

- **Delta features:** to capture temporal changes, adjacent-day differences and whole-period differences are generated automatically.
- **Institutional information:** the total number of patients at the institution and the number of patients by disease are input as continuous values.

2) *Statistical Approach: L2-Regularized Logistic Regression:* In this method, the execution of a branch ($y = 1$) is used as the objective variable, and global, linear factors are estimated as odds ratios by binomial logistic regression. In medical data, the correlations among test values are strong, and coefficients tend to be unstable owing to multicollinearity. We therefore introduce L2 regularization (Ridge) [9] to stably separate the contributions of patient-condition and institutional factors. To preserve the temporal order of causation, we use data up to the day before the branch. For factor evaluation, we use the standardized coefficient β^* and the odds ratio, regarding factors with a large $|\beta^*|$ as strong contributors [16]. The clinical validity is confirmed afterward using group-wise execution rates and odds ratios and p-values from 2×2 contingency tables. However, when the sample size of the abnormal group is small, the odds ratio can be inflated by quasi-complete separation [12]. We therefore report the odds ratio together with the group-wise sample sizes and execution rates (Table III), so that odds ratios inflated by quasi-complete separation can be recognized from those figures rather than accepted at face value.

3) *Explainable Learning-Based Approach: Predictive Model and SHAP:* To achieve explainable estimation of branching factors, we combine a predictive model with SHAP [4, 10]: the predictive model captures linear and nonlinear effects, whereas SHAP attributes each feature’s contribution to the prediction, yielding explainable factors with decision thresholds. The key idea of this approach is to identify, from the dependence of the SHAP values on feature values, the point at which the sign of the contribution reverses and interpret it as a model-derived decision boundary. This presents, in a clinically interpretable form, the test value or change at which a medical action branches. The specific choice of the predictive model and the handling of missing values are determined in Section IV-C2 through a comparison of several candidate models.

IV. EVALUATION EXPERIMENT

A. Experimental Data

We used an anonymized EMR dataset based on the Next Generation Medical Infrastructure Act, collected from 27 medical institutions between 2015 and 2024 (approved by the Life Data Initiative Ethics Committee, No. 2024_MIL_0004_A001). This dataset consists of 58,598 patients with nine diseases, including breast, lung, and gastric malignancies and heart disease (see [7] for

the breakdown by disease). The extraction and evaluation of frequent patterns, the selection of branching points, and the comparison of the prediction accuracy of three models were conducted on all nine diseases, while the detailed estimation of branching factors was performed on 5,961 patients with gastric malignancy as a case study. All 27 institutions were included without filtering, with institutional info used solely as an explanatory variable. As no facility attributes (e.g., bed count) are retained, facility size is operationalized as the number of patients an institution contributes to this dataset—a proxy, not a direct measure. Of the 27 institutions, 24 contributed gastric malignancy cases (median 221, range 4–983). The gastric malignancy cases analyzed here all share a single top-level procedure code (K653), so the cohort is narrower than a disease-level grouping. Anonymization removes the subcategory digits that would distinguish variants within this code, so finer procedure-level differences cannot be observed.

B. Experimental Method

First, medical order sequences with elapsed days were created for each disease from the 27 institutions, and frequent patterns were extracted following prior work [6, 7]. Next, by comparison with each disease’s standard clinical pathway, patterns in the top and bottom quartiles of the F-measure distribution were selected as candidate branching points, and a dataset was constructed with the selected medical order as the objective variable y and the patient state and institutional information as the explanatory variables x . On this dataset, three models—logistic regression (LR), random forest (RF), and XGBoost [17]—were evaluated by 5-fold stratified cross-validation, and the predictive model’s performance was confirmed by the area under the curve (AUC; Section IV-C2). Branching factor estimation then used L2-regularized LR (Section IV-C3) and XGBoost with SHAP (Section IV-C4).

C. Experimental Results (Gastric Malignancy)

Gastric malignancy was chosen for factor estimation because it has many patterns consistent with the standard clinical pathway.

1) *Accuracy Evaluation of the Extracted Frequent Medical Order Patterns:* Varying minSup in stages, we computed the recall, precision, and F-measure of the frequent patterns extracted from the nine diseases, and confirmed their consistency with the standard clinical pathway (Table I). The precision was high at 0.74–0.97 across all settings, indicating that the extracted patterns include the medical orders of the standard clinical pathway. The recall, on the other hand, was low at 0.52–0.54. This is because the loss of time information through anonymization causes orders specified for Day 1 in the standard to be extracted with a shift to Day 2–3 in the actual data; correcting these date shifts is a future task. Lowering minSup to 30% increases the number of patterns to 1,547 but reduces

TABLE I
EVALUATION RESULTS OF FREQUENT MEDICAL ORDER PATTERN
EXTRACTION: AVERAGES OVER THE NINE DISEASES.

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

TABLE II
COMPARISON OF PREDICTIVE PERFORMANCE (MEAN AUC BY DAY
GROUP OVER ALL NINE DISEASES).

Day group	N	LR	RF	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

the F-measure to 0.61 (owing to the inclusion of less relevant, fine-grained patterns). With minSup = 40%, in contrast, foundational patterns such as surgery on Day 0 and intravenous drips (antibiotics and blood substitutes) on Day 0 and Day 1 could be extracted stably without an excessive increase. For individual core patterns, the precision and recall reached 1.00.

2) *Validation of the Predictive Model:* As a prerequisite for factor estimation, the three models (LR, RF, and XGBoost) were evaluated on all nine diseases (81 targets) by 5-fold stratified cross-validation, and their predictive validity was assessed by AUC (Table II). Whereas LR remained at an average of 0.941, the nonlinear RF and XGBoost both achieved averages above 0.99. The gap widens for strongly nonlinear Day groups: LR drops to 0.812 (Day -1), 0.859 (Day 2), and 0.828 (Day 4), while XGBoost maintains 0.95–1.00. On Day 0 and Day 7, LR is also high at 0.98 or above, and these branches are linearly predictable. Thus, capturing feature interactions and nonlinear effects is important, and we adopt XGBoost for factor estimation—not because of the accuracy difference but because of its NaN-aware splits, which treat missingness as a signal (informative missingness), and its suitability for SHAP-based decision threshold extraction.

3) *Branching Factor Estimation Results by Logistic Regression:* For gastric malignancy, we extracted factors satisfying $|\beta^*| \geq 0.9$ and $p < 0.05$ by L2-regularized LR ($\lambda = 1.0$), obtaining six candidate factors from four of the nine targets (Table III). Clinical specialists judged each factor on a three-level scale of $\bigcirc$ (valid), $\triangle$ (no direct causation but indirectly related), and $\times$ (invalid), yielding zero valid, three indirectly related, and three invalid.

All extracted factors are “downward trends.” All six were downward trends in creatinine, TP, or CRP (transitions to low abnormal values over the few preoperative

days), and the other indicators (vital signs and absolute values or daily variations of test values) did not satisfy $|\beta^*| \geq 0.9$. This is considered to be because linear LR strongly captures the linear relationship between the low-value group and the normal group.

Imbalance of group samples and quasi-complete separation. In all six cases, the low-abnormal-trend group was small at 1–81 patients; in particular, creatinine and TP on Day -1 are unstable with two patients or fewer. The CRP downward trend in the Day 1 infection test has an OR exceeding 20,000, but the execution rates of the two groups are almost the same at 99% versus 94% (in fact lower in the abnormal group); this is an apparent value produced by quasi-complete separation due to the near-universal execution (positive rate of 99%), indicating the breakdown of LR estimation.

Institutional factors were more consistently identified. Facility size (by disease) was significantly extracted for four targets (the hemostatic function, renal function, and blood count tests on Day -1, and the blood substitute on Day 2), but as it is an institutional operational factor rather than a clinical trigger, it was excluded from judgment. The institutional factor is robust, with a large sample and $p \ll 0.001$, in contrast to the aforementioned unstable downward trends.

Evaluation by clinical specialists. The three invalid cases lacked medical causality: the creatinine decrease was judged “a low value itself is not regarded as a problem and is hard to consider a factor,” and the Day 1 infection test CRP decrease “performing an infection test on the day after surgery is not normally expected.” For the three indirectly related cases, TP was “an indicator of nutritional status, suggesting confounding due to set-test operation,” and the Day 1 blood count CRP decrease “a factor is hard to find because routine blood sampling covers almost all cases.”

Overall, LR yielded no clinically valid branching factors, owing to its linearity assumption, dependence on extreme downward trends, and quasi-complete separation.

4) *Explainable Estimation of Branching Factors by XGBoost and SHAP:* Using XGBoost with SHAP, we extracted a total of 35 candidate factors, excluding institutional factors, from eight targets, and clinical specialists judged their validity. The breakdown was 16 valid, 6 indirectly related, and 13 invalid; compared with LR (zero valid), the medically supported findings increased substantially. Many of the invalid cases stem from confounding due to routine tests or from quasi-complete separation, indicating the need for filtering by clinical knowledge. We highlight two key observations below.

Dominant influence of institutional factors. Facility size (total patients and by disease) ranked in the top 10 of Gain Importance for all nine targets, and the total patient count ranked first for six targets. It is thus a dominant predictor of perioperative order decisions, alongside individual clinical indicators—a finding obtained

TABLE III
LOGISTIC REGRESSION FACTOR ESTIMATION RESULTS AND CLINICAL SPECIALISTS' JUDGMENTS.

Target	Associated factor	Normal range	Normal-trend group (n, rate)	Low-abnormal-trend group (n, rate)	Odds Ratio	p-value	Judg.
Day -1 renal function test	Creatinine downward trend (Day -1) [mg/dL]	0.0–2.0	125 (2%)	2 (50%)	4.89	0.047	×
	TP downward trend (Day -1) [g/dL]	6.0–9.0	75 (3%)	1 (100%)	2.50	0.039	△
Day -1 blood count test	Creatinine downward trend (Day -1) [mg/dL]	0.0–2.0	125 (2%)	2 (50%)	5.08	0.047	×
	TP downward trend (Day -1) [g/dL]	6.0–9.0	75 (3%)	1 (100%)	2.54	0.039	△
Day 1 infection test	CRP downward trend (Day -1) [mg/dL]	0.0–1.0	1,737 (99%)	81 (94%)	23,940.54	0.009	×
Day 1 blood count test	CRP downward trend (Day -1) [mg/dL]	0.0–1.0	1,737 (99%)	81 (95%)	26,279.61	0.014	△

Normal range denotes the clinical reference interval of each test item, which defines the two groups: the normal-trend group stays within the range, whereas the low-abnormal-trend group deviates below it. Judg., judgment by clinical specialists: ○, valid; △, no direct causation but indirectly related; ×, invalid. Abbreviation: TP, total protein.

thanks to the multi-institutional data from the 27 medical institutions—though it may proxy unobserved institution-level case mix, including the procedure mix, which we cannot separate from operational practice here.

Differences in physicians' judgments by target. Whereas the Day 2 blood substitute (6/7 valid) and the Day 1 renal function test (4/6 valid) showed high conceptual consistency, no factor was judged valid for the Day -1 12-lead electrocardiogram or the Day 1 infection test, and their findings are likely to originate from confounding.

Below, we describe the Day 1 renal function test in detail as a representative example (Table IV). Of the six extracted factors, five are findings related to blood pressure or pulse, and four of these were judged valid.

As a representative finding, Fig. 2 shows the SHAP dependence plot for the Day -1 → Day 0 change in pulse. It shows a clear negative relationship in which the SHAP value decreases as the change in pulse becomes larger, with the sign reversing from positive to negative around 26.37 bpm (the decision threshold). A clinical specialist evaluated that “a pulse change of 26.37 bpm is quite large, and patients with hypertension or arteriosclerosis show large perioperative fluctuations,” supporting the view that pulse variation can be involved in the decision to order a renal function evaluation.

In the SHAP summary plot of Fig. 3, a two-layer structure was observed: facility size (by disease and total patients), which occupies the top ranks, acts globally with a wide distribution of SHAP values, while the blood pressure and pulse factors split in sign according to the patient's values and locally modulate the decision. The blood-pressure-related factors were evaluated by a clinical specialist as “hypertension tends to accompany renal complications, so a renal function test on the first post-operative day can be justified,” and can be consistently interpreted as a postoperative renal function evaluation against a background of hypertension and arteriosclerosis.

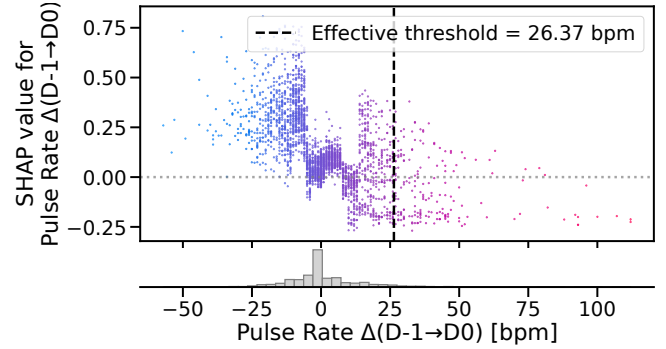


Fig. 2. Dependence plot for pulse [Day -1 → Day 0 change].

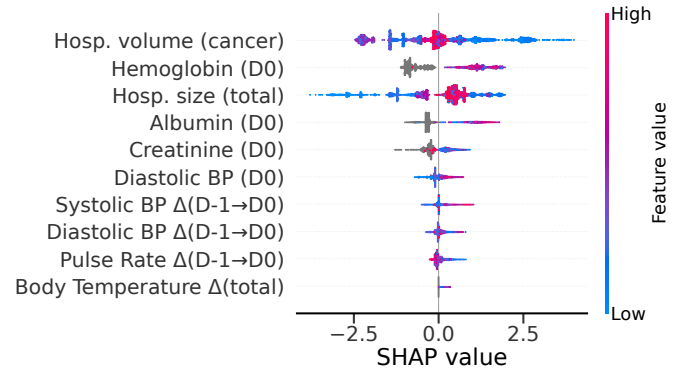


Fig. 3. SHAP summary plot for the Day 1 renal function test (top 10 features by Gain Importance). Hosp., hospital; BP, blood pressure; D0, day of surgery; Δ, change between the indicated days.

Regarding the invalid Alb (Day 0), a specialist commented that “malnourished patients tend to have concurrent renal dysfunction and one would want to draw blood, but the direction of the data is the opposite”; its Gain Importance is high, but no decision threshold was extracted and its class balance was strongly skewed.

From the above, XGBoost with SHAP can visualize non-linear, threshold-based branching factors, and it quantitatively demonstrated two structural findings: the dominant influence of institutional factors, and the strong condition—

TABLE IV
SHAP FACTOR ESTIMATION RESULTS FOR THE DAY 1 RENAL FUNCTION TEST AND CLINICAL SPECIALISTS' JUDGMENTS.

Associated factor	Decision boundary	Below-threshold group (n , rate)	At-or-above-threshold group (n , rate)	Odds Ratio	p-value	Judg.
Pulse [Day -1 → Day 0 change]	26.37 bpm	3,329 (62%)	178 (38%)	2.61	<0.001	○
Diastolic blood pressure (Day 0)	68.57 mmHg	553 (49%)	3,277 (61%)	0.59	<0.001	○
Diastolic blood pressure [Day -1 → Day 0 change]	1.00 mmHg	1,755 (55%)	1,869 (62%)	0.76	<0.001	○
Systolic blood pressure (Day 0)	123.53 mmHg	685 (52%)	3,142 (61%)	0.68	<0.001	○
Systolic blood pressure [Day -1 → Day 0 change]	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	×

The decision boundary is the SHAP sign-reversal point, i.e., a value at which the model contribution changes sign. [†]: no SHAP boundary extracted; the lower limit of the normal range is used instead. Odds Ratio: below- vs. at-or-above-boundary group. Judg. (by clinical specialists): ○, valid; △, indirectly related; ×, invalid. Alb, albumin.

medical-order association in the Day 2 blood substitute and the Day 1 renal function test.

V. CONCLUSION

In this study, we proposed a branching factor estimation method that selects branching points based on F-measure consistency with a standard clinical pathway and estimates factors with two approaches: a statistical one (L2-regularized LR) and an explainable learning-based one (predictive model with SHAP; XGBoost adopted via model comparison). In an evaluation experiment using data from 58,598 patients across nine diseases and 27 medical institutions, the validity rate in clinical specialists' judgments reached 45.7% for XGBoost with SHAP versus 0% for LR; for gastric malignancy, we extracted factors using diastolic blood pressure and pulse-rate change as decision thresholds in the Day 1 renal function test (4/6 valid), supporting the effectiveness of the proposed method and the greater effectiveness of the explainable learning-based approach in this case study. We also obtained a finding specific to multi-institutional analysis: institutional factors were dominant predictors across all nine branching targets identified for gastric malignancy.

Two limitations remain: the admission year is not modeled, so pathway revisions over 2015–2024 may be absorbed into the institutional factors; and anonymization removes the procedure subcategories. Future work includes modeling the admission year, inferring the procedure type from order sequences, and extending the detailed branching-factor estimation to all nine diseases, with qualitative and quantitative evaluation by clinicians.

ACKNOWLEDGMENT

Part of this research was supported by a Grant-in-Aid for Scientific Research (#24K02943) from the Japan Society for the Promotion of Science.

REFERENCES

- [1] S. Arikat et al. Evolving role of clinical laboratories in precision medicine. *Journal of Laboratory and Precision Medicine*, 9:17, 2024.
- [2] J. K. Tan et al. Machine learning-based prediction for high health care utilizers by using a multi-institutional diabetes registry. *JMIR AI*, 3:e58463, 2024.
- [3] H. H. Le et al. Methods for analyzing medical-order sequence variants in sequential pattern mining. *ACM Transactions on Computing for Healthcare*, 4(1):3:1–3:28, 2023.
- [4] S. M. Lundberg and S.-I. Lee. A unified approach to interpreting model predictions. In *Advances in Neural Information Processing Systems*, volume 30, pages 4765–4774, 2017.
- [5] T. Maruyama et al. Explainable machine learning for predicting postoperative length of stay after gastrectomy: a nationwide study using XGBoost and SHAP. *Frontiers in Medical Technology*, 7:1732580, 2025.
- [6] M. Sugitani et al. Extracting and visualizing frequent medical instruction patterns with statistical insights from multi-institutional electronic medical record data. In *IEEE International Symposium on Computer-Based Medical Systems*, pages 514–515, 2025.
- [7] M. Sugitani et al. Statistical analysis and visualization of medical instruction patterns in multimedia-rich electronic medical records. In *IEEE International Conference on Multimedia Information Processing and Retrieval*, pages 376–382, 2025.
- [8] J. W. Tukey. *Exploratory Data Analysis*. Addison-Wesley, 1977.
- [9] A. E. Hoerl and R. W. Kennard. Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, 12(1):55–67, 1970.
- [10] S. M. Lundberg et al. From local explanations to global understanding with explainable AI for trees. *Nature Machine Intelligence*, 2(1):56–67, 2020.
- [11] Y. Hua et al. Clinical risk prediction with logistic regression: Best practices, validation techniques, and applications in medical research. *Academic Medicine & Surgery*, page 131964, 2025.
- [12] Y. Guan and G.-H. Fu. A double-penalized estimator to combat separation and multicollinearity in logistic regression. *Mathematics*, 10(20):3824, 2022.
- [13] S. Tou et al. Identifying key variances in clinical pathways associated with prolonged hospital stays using machine learning and ePath real-world data. *JMIR Medical Informatics*, 13:e71617, 2025.
- [14] F. Yi et al. XGBoost-SHAP-based interpretable diagnostic framework for Alzheimer's disease. *BMC Medical Informatics and Decision Making*, 23(1):137, 2023.
- [15] S. Hur et al. Comparison of SHAP and clinician friendly explanations reveals effects on clinical decision behaviour. *npj Digital Medicine*, 8:578, 2025.
- [16] P. Schober et al. Correlation coefficients: Appropriate use and interpretation. *Anesthesia & Analgesia*, 126(5):1763–1768, 2018.
- [17] T. Chen and C. Guestrin. XGBoost: A scalable tree boosting system. In *ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 785–794, 2016.