複数医療機関間の電子カルテデータを用いた統計情報付き頻出医療指示パターンの抽出と可視化

杉谷　美和[†]　　松尾　亮輔[†]　　山崎　友義[†]　　荒木　賢二[†]　　小口　正人[†]

横田　治夫[††]　　Le Hieu Hanh[†††]

† お茶の水女子大学理学部　〒 112–0012 東京都文京区大塚 2-1-1

†† 城西大学理学部数学科　〒 102-0093 東京都千代田区平河町 2-3-20

††† お茶の水女子大学共創工学部　〒 112–0012 東京都文京区大塚 2-1-1

E-mail: †{g2120520,oguchi,le}@is.ocha.ac.jp, ††matsuo@ldi.or.jp,

†††{yamazaki.cp,araki6925,yokota.h.aa}@gmail.com

あらまし　電子カルテの普及によりカルテ情報の分析が進んでいるが，異なる医療機関の電子カルテデータは標準化されておらず，形式や表記も異なるため，単一医療機関のデータ解析が主流の従来の方法では，医療機関間の比較や特徴抽出は困難である．また，複数医療機関のデータから疾患ごとの医療指示パターンを抽出することは診療プロセスの標準化や改善に有用であるが，疾患ごとの特徴や治療方針の違いを考慮する必要があり，解析は複雑である．本論文では，データの標準化および名寄せ技術を適用して複数医療機関のデータを統合し，共通基準で解析可能な手法を提案する．提案手法では，電子カルテデータから疾患ごとの医療指示シーケンスを抽出し，日付情報を基に時間間隔を考慮した頻出医療指示シーケンスとして頻出クローズドパターンを抽出する．さらに，解析過程でシーケンス ID を保持することで，頻出医療指示パターンに該当する患者の検査結果や投薬情報を参照可能とし，医療機関ごとの特徴を分析可能にする．評価実験では，人工データを用い，頻出医療指示シーケンスの分岐，統計情報，検査結果の異常値発生率の変動を可視化し，意思決定支援や診療プロセス改善の有用性を確認する．

キーワード　医療・ヘルスケア，パターン・相関ルールマイニング，データ統合技術，時系列データ分析

1　はじめに

近年，電子カルテの普及が進んでおり，大規模病院を中心に導入が進められている．現在では中小規模の病院にも電子カルテが拡大しており，医療現場のデジタル化が加速している．また,「千年カルテプロジェクト」 [1] をはじめとする医療データの集約および二次利用の取り組みも活発化しており，大量の医療データを活用した新たな診療支援や研究の可能性が広がっている．

医療データ解析の重要性も増しており，医療指示シーケンスの解析は診療プロセスの標準化や改善，診療精度の向上に貢献している．標準化された診療計画により，医療の質の均一化と患者の治療効果の向上が期待される．一方で，患者の年齢や健康状態によって異なる医療指示が求められるため，個別化された診療支援の必要性も指摘されている [2,3].

さらに，複数医療機関でのデータ共有・解析の重要性が高まっている．他院のデータを参照することで，自院の診療プロセスの改善が期待される．

本研究は，複数医療機関から提供された匿名加工済みの電子カルテデータを前提に，それらデータを統合し，疾患ごとの頻出医療指示パターンを抽出・分析することを目的とする．これにより，医療機関間の診療プロセスの差異や共通点を明らかにし，診療の標準化や質の向上を図る．

診療プロセスの解析に関する研究は数多く行われているが，単一医療機関のデータに基づくものが多く，異なる医療機関間での診療プロセスの比較は十分に行われていない [4,5]．また，複数医療機関のデータを解析する試みとして，特定疾患に焦点を当てた研究 [6,7] も存在するが，疾患ごとの違いや複数疾患にまたがる診療プロセスの解析は限定的である．さらに，時間間隔を考慮した医療指示パターンの抽出に取り組んだ研究もあるが [8,9]，頻出医療指示パターンに関連する検査結果や検査結果の異常値の統計情報の抽出は十分ではなく，有効な可視化ツールも存在していない．

本研究では，これらの課題を踏まえ，複数医療機関から提供された匿名加工済みの電子カルテデータを前提に，それらデータを統合し，疾患ごとに頻出する医療指示パターンを抽出・分析する．加えて，医療指示間の時間的な要素を考慮した T-PrefixSpan アルゴリズム [10]，を用いることで，診療プロセスの時間的特徴も明らかにする．これにより，医療機関間の診療プロセスの違いを可視化し，診療の標準化や質の向上に寄与することを目指す.

評価実験では，複数の医療機関から提供された電子カルテデータを参考に作成した 10,000 人分の医療指示シーケンスを含む人工データセットを用い,「頻出医療指示パターンの差異が可視化されるか」「統計情報付きの頻出医療指示パターンを適切に抽出できるか」「頻出医療指示パターンに関連する検査結果の異常値発生率を算出できるか」の 3 点に着目して実験を行

い，本研究の有用性を検証する．

本稿は以下の構成で記述される． 2 節では，本研究に関連する技術や概念を説明し，関連研究の詳細を述べる． 3 節では，電子カルテデータを解析し，頻出医療指示パターンを抽出・可視化する手法を述べる． 4 節では，提案手法を用いた評価実験の結果を示す．最後に， 5 節では本研究のまとめと今後の課題について述べる．

2 背景知識と関連研究

2.1 背 景 知 識

2.1.1 シーケンス

定義 2.1 (アイテムセット).
アイテムセット I を以下のように定義する．

$$I = \{i_1, i_2, \ldots, i_n\}$$

ここで $i_j \in I$ はアイテムを表し，それぞれがユニークな要素として扱われる．

定義 2.2 (シーケンス).
アイテムセット I に対して，シーケンス S を以下のように定義する．

$$S = (\{s_1, s_2, \ldots, s_m\}, \prec_S)$$

ここで $s_j = (id, i), i \in I$ であり， id はシーケンス内でユニークなインデックスである．また，$\prec_S$ は S 上の全順序関係であり，

$$\forall s_i, s_j \in S, s_i \prec_S s_j \vee s_j \prec_S s_i$$

が成り立つ．

2.1.2 シーケンシャルパターンマイニング (SPM)

SPM は Agrawal らによって提案されたシーケンシャルデータベース (以下，SDB) から頻出シーケンスを抽出する手法である [11]． 特に，牧原ら [12] は，電子カルテのアクセスログから頻出シーケンスを抽出することで，医療指示シーケンスの解析を行っている．アイテムの順列をシーケンスと呼び，SDB はあるシーケンス集合に属するシーケンスと，そのシーケンスを一意に決める識別子 (以下，sid) を組みとする要素からなる．

定義 2.3 (シーケンスデータベース).
シーケンスデータベース (SDB) は以下のように定義される．

$$SDB = \{(sid_1, S_1), (sid_2, S_2), \ldots, (sid_m, S_m)\}$$

ここで sid_i は各シーケンスを識別する識別子 (Sequence ID) であり，S_i はシーケンスである．

定義 2.4 (頻出シーケンス).
シーケンスデータベース SDB に対し，最小支持度 minSup を設定した際，支持度が minSup 以上のシーケンスを頻出シーケンスと定義する．

定義 2.5 (タイムアイテム).
アイテム集合 I の要素 $i \in I$ に対し，アイテムが発生した時刻を t としたとき，タイムアイテム (i， t) は以下のように定義される．

$$(i, t)$$

ここで， i はアイテム， t は発生時刻を表す．

定義 2.6 (タイムシーケンス).
タイムアイテムから構成される順列 s をタイムシーケンスと呼び，以下のように表す．

$$s = \langle (i_1, t_1), (i_2, t_2), \ldots, (i_n, t_n) \rangle$$

ここで， s は長さ n のタイムシーケンスであり，各要素

$$(i_k, t_k)$$

はアイテムとその発生時刻の組である．

定義 2.7 (シーケンスバリアント (Sequence Variant，SV)).
タイムシーケンス $s = \langle (i_1, t_1), (i_2, t_2), \ldots, (i_n, t_n) \rangle$ が存在する場合，同一のアイテム集合 I に対し，時間間隔やアイテムの順序が異なるシーケンス $s' = \langle (i'_1, t'_1), (i'_2, t'_2), \ldots, (i'_m, t'_m) \rangle$ をシーケンスバリアント (SV) と定義する．
シーケンスバリアント $SV(s)$ は以下のように表される．

$$SV(s) = \{s' | s' \neq s, s' \in SDB\}$$

ここで，s' はシーケンスデータベース SDB 内の s に関連するシーケンスであり，バリアントは頻出シーケンスの時間的・構造的な多様性を示す．

定義 2.8 (クローズドパターン).
シーケンス S の中で，同じ支持度を持ち，他の頻出パターンに包含されない最大の部分シーケンスをクローズドパターンと定義する．

2.2 関 連 研 究

本節では，本研究に関連する研究を紹介する．これらの研究は，診療プロセスの標準化や診療の質向上に寄与しており，本研究においても重要な基盤となる知見を提供している．

2.2.1 単一医療機関での頻出医療指示パターン抽出と可視化

本田らの研究 [5] では，単一医療機関の電子カルテログを用いて，医療指示間の時間情報を含む頻出医療指示シーケンスの抽出が行われている．頻出する医療指示のバリアント（SV）を検出・可視化し，時間間隔を考慮した診療プロセスの分岐や多様性が直感的に把握できる仕組みを構築している．また，医療従事者が視覚的に結果を確認できる対話型グラフィカルインターフェースシステムを開発した．

さらに，山田らの研究 [4] では，電子カルテに記録された医療指示の流れをシーケンスデータとして捉え，シーケンシャルパターンマイニング（SPM） [11] を用いて頻出する医療指示列の抽出を行っている．この研究では，シーケンス ID（SID）を保持することで，シーケンスごとのバリアント（SV）の統計

評価や指標算出が可能となっている．

しかし，これらの研究は単一医療機関のデータのみに基づいて解析が行われており，複数医療機関間でのデータ標準化や統合という観点が考慮されていない．そのため，医療機関ごとの診療プロセスの違いや施設間での比較が困難であるという課題が存在する．

本研究では，複数の医療機関から収集されたデータを統合し，共通する頻出医療指示パターンを抽出することに焦点を当てている．これにより，異なる医療機関間での診療プロセスの比較が可能となり，診療の標準化や施設間連携の促進に寄与することを目指している．

2.2.2 複数医療機関での頻出医療指示パターン抽出と可視化

趙らの研究 [7] では，複数の医療機関における COVID-19 の診療プロセスを比較し，頻出治療パターンの違いを分析している．これにより，各施設の特徴やバリアントが明確化され，診療の標準化および医療連携の促進が期待されている．

また，安光らの研究 [6] では，複数の医療機関の電子カルテデータを対象とし，COVID-19 に関する頻出医療指示の抽出を行っている．抽出した医療指示パターン（SV）の距離を定義し，病院間で階層的クラスタリングを行うことで，各病院の診療プロセスを比較した．さらに，クラスター内での医療指示パターンを併合し，共通するパターンを作成することで，診療プロセスの標準化および病院間の医療連携の促進を図っている．

しかし，これらの研究は複数の医療機関を対象としているものの，単一の疾患に焦点を当てて頻出医療指示パターンの抽出が行われている．そのため，複数疾患にまたがる診療プロセスの解析や，異なる疾患間での頻出医療指示パターンの比較が行われていない点が課題として挙げられる．

本研究では，疾患や治療単位にとらわれず，医療指示全体を対象とした頻出医療指示パターンを抽出することで，幅広い診療プロセスの解析を可能とする．

2.2.3 時間間隔を考慮した頻出医療指示パターン抽出手法

佐々木らの研究 [9] では，タイムインターバル SPM を適用し，医療指示の時間的な間隔を考慮した頻出医療指示パターン抽出が行われている．これにより，診療プロセスの時間的な要素を踏まえた分析が可能となり，医療従事者の意思決定支援に寄与している．

また，Le らの研究 [8] では，電子カルテから頻出医療指示パターンの抽出とバリアント解析を実施している．効率的な SPM である CSpan [13] を拡張した T-CSpan [14] を使用し，医療指示間の時間間隔を考慮したパターン抽出を実現した．T-CSpan はタイムインターバルの統計情報を収集できる手法であり，頻出医療指示パターンのより詳細な解析が可能となる．さらに，算出したタイムインターバルの統計情報を比較し，分岐要因の推定を行う．

しかし，これらの研究は時間間隔に焦点を当てているものの，具体的な検査結果や投薬情報の解析が十分でないという課題がある．その結果，医療プロセスの変化や検査結果の異常発生頻度に関する詳細な評価が難しく，診療成果に与える影響の分析が限定的となっている

本研究では，頻出医療指示パターン内の検査結果や投薬情報の統計を算出し，検査結果の異常発生率を分析することで，より詳細な診療プロセスの改善を目指す．時間間隔だけでなく，医療指示全体にわたる情報を考慮し，診療成果への影響を可視化することが特徴である．

3 提 案 手 法

本研究では，電子カルテデータを活用して統計情報付き頻出医療指示パターンを抽出し，検査結果の異常値発生情報を付与することで，診療プロセスの標準化および医療従事者の意思決定を支援することを目的とする．提案手法は以下の 4 つのステップで構成される．

3.1 医療指示シーケンスの作成

患者の診療履歴を時系列で整理し，実際に行われた医療指示の流れを可視化することは，診療プロセスの分析や医療の標準化において重要な役割を果たす．本研究では，電子カルテデータから各患者の診療履歴を収集し，医療指示を時系列順に整理することで患者ごとの医療指示シーケンスを作成する．具体的には，診療記録内の医療指示や検査項目を抽出し，医療的に重要な医療指示を行なった日を基準日 (Day 0) として，各医療指示を経過日数に基づいて配置する．また，同じ患者であっても入院を繰り返すなどの理由で入院日が異なる場合は，各入院ごとに独立した診療プロセスとして整理する．この方法により，患者ごとの診療プロセスの流れを可視化し，医療指示の標準化や診療の改善に向けた分析が可能となる．

定義 3.1 (同日の医療指示の並び替えルール).
医療指示の実行時刻がないという前提で，同一日に複数の医療が記録されている場合は，以下の順序で並び替えを行う．

$$\begin{bmatrix} 1. & \text{手術} \\ 2. & \text{投薬} \\ 3. & \text{検査} \\ 4. & \text{診療行為} \end{bmatrix}$$

さらに，同一カテゴリ内で複数の医療指示が存在する場合は，事前に作成した辞書に基づき辞書順（あいうえお順）で並び替える．このルールにより，医療指示の順番が統一され，頻出医療指示パターン抽出の精度向上が期待される．

定義 3.2 (医療指示シーケンス).
患者 p_i に対する医療指示のシーケンスを以下のように定義する．

$$S_{p_i} = \langle s_1, s_2, \ldots, s_n \rangle$$

s_j は j 番目の医療指示を表し，$s_j = (t_j, a_j)$ である．なお，t_j は医療指示が実施された経過日数を表し，a_j は医療指示の種類（例：手術，投薬，検査など）を表す．

定義 3.3 (経過日数付き医療指示シーケンス).
医療指示シーケンスの上で医療的に重要な医療指示が行われた

日（例えば手術日）を基準（Day 0）とし，すべての医療指示に経過日数 d_j を付与する．

$$S_{p_i} = \langle (d_1, a_1), (d_2, a_2), \ldots, (d_n, a_n) \rangle$$

t_0 は基準日を示し，すべての医療指示が基準日と相対的に再配置される．$d_j = t_j - t_0$ は経過日数を表し，a_j は医療指示の種類を表す．なお，同じ d_j に複数の医療指示が含まれる場合もある．
また，手術を行っていない患者については，入院日を基準（Day 0）とし，すべての医療指示を入院日を基準として再配置する．

このように整理された医療指示シーケンスのデータは，患者ごとに一意な識別子 SID が付与され新しいデータテーブルに格納される．医療指示シーケンスの蓄積と解析により，診療プロセスの傾向を把握し，医療従事者が診療方針を決定する際の参考情報として活用される．

3.2 統計情報付き頻出医療指示パターンの抽出

医療指示シーケンスを基に，疾患ごとの頻出医療指示パターンを抽出する．本研究では，T-PrefixSpan アルゴリズム [10] を用いて，医療指示の順序だけでなく，指示間の時間的要素を考慮したパターン抽出を行う．

T-PrefixSpan は， SPM の手法を拡張したものであり，浦垣らによって提案された I-PrefixSpan [15] の問題点を解消し，より柔軟な時間間隔の取り扱いを可能にしたアルゴリズムである．

従来の PrefixSpan [16] では，シーケンス内のアイテムの順序のみを考慮して頻出パターンを抽出するが，T-PrefixSpan ではアイテム間の時間間隔が重要視され，時間依存性のある頻出パターンを抽出する点に特徴がある．

定義 3.4 (医療シーケンスデータベース， MSDB).
医療シーケンスデータベース (MSDB) は複数患者の医療指示列から構成される．MSDB 内のデータ集合 D は以下のように定義される．

$$D = \{(SID_1, S_{p_1}), (SID_2, S_{p_2}), \ldots, (SID_m, S_{p_m})\}$$

ここで，SID_i は i 番目の患者の識別子であり， S_{p_i} はその患者における医療指示シーケンスを示す．

定義 3.5 (T-PrefixSpan による頻出シーケンス).
時間間隔を考慮した頻出シーケンス fs は以下の形式で表される．

$$fs = \langle (a_1, x_1), (a_2, x_2), \ldots, (a_n, x_n) \rangle$$

ここで，

$$x_j = t_{j+1} - t_j$$

は医療指示 a_j と a_{j+1} の間隔を表す．
T-PrefixSpan では，医療指示の順序に加えて時間的な要素を保持するため，診療プロセスのタイミングやリズムを解析でき，時間依存性を伴う頻出医療指示パターンの抽出が可能となる．

このアプローチにより，時間間隔を含む医療指示列が抽出され，実際の診療プロセスにおける重要なシーケンスが精度高く得られる．

さらに，ここで抽出された頻出医療指示パターンに対して，各医療指示が行われた際の検査結果や投薬量を収集し，統計情報を算出する．検査結果と投薬量は，疾患の進行や治療効果を評価し，診断支援や治療の適正化に役立つ重要な指標である．

定義 3.6 (統計情報付き頻出医療指示パターン).
頻出医療指示パターン fs に関連付けられる検査結果や投薬量をの集合 $X(\alpha)$ は，次のように定義される．

$$X(\alpha) = \{x_1, x_2, \ldots, x_n\}$$

ここで，α は特定の検査を，x_j は頻出医療指示パターン fs に対応する検査値または投薬量を表す要素であり， k は観測値の総数とする．これに基づき，統計情報は以下のように計算される．

$$\begin{aligned}
&\text{平均値:} \quad \mathrm{Mean}(X(\alpha)) = \tfrac{1}{k} \textstyle\sum_{j=1}^{k} x_j \\
&\text{中央値:} \quad \mathrm{Median}(X(\alpha)) = x_{\lceil \frac{k}{2} \rceil} \\
&\text{最大値:} \quad \mathrm{Max}(X(\alpha)) = \max(X(\alpha)) \\
&\text{最小値:} \quad \mathrm{Min}(X(\alpha)) = \min(X(\alpha))
\end{aligned}$$

3.3 検査結果の異常値発生率の算出

本研究では，特定の疾患における 1 つの頻出医療指示パターンごとに，検査結果の異常値発生率を算出する．異常値の検出は，対象疾患に関連するすべての医療指示シーケンスを参照し，頻出医療指示パターンに「含まれる」群と「含まれない」群に分けて行う．その後，それぞれの群に対して検査結果を抽出し，正常範囲内，異常（高値または低値）の割合を経過日数ごとに算出する．

定義 3.7 (検査結果の異常値発生率).
異常値発生率とは，特定の検査 α における検査結果が正常範囲 $[B(\alpha), T(\alpha)]$ を逸脱する割合を指し，以下の高異常率，低異常率，正常率によって具体的に表される．頻出医療指示パターン fs における検査結果の異常値発生率は，対象疾患に関連するすべての医療指示シーケンスを次のように分類して算出する．
まず，対象疾患に関連するすべての医療指示シーケンス

$$D = \{S_{p_1}, S_{p_2}, \ldots, S_{p_m}\}$$

を参照し，頻出医療指示パターン fs のに含まれる群 D_{in} および含まれない群 D_{out} に分離する．

$$D_{\text{in}} = \{S_{p_i} \in D | fs \subseteq S_{p_i}\}$$

$$D_{\text{out}} = D \setminus D_{\text{in}}$$

次に，それぞれの群に対して経過日数 d_j ごとの検査結果を収集する．検査結果 x_j は，検査項目の正常範囲 $[B(\alpha), T(\alpha)]$ に対して以下のように分類される．

$$C(x_j) = \begin{cases} -1 & (x_j < B(\alpha)) \quad \text{(低値)} \\ 0 & (x_j \in [B(\alpha), T(\alpha)]) \quad \text{(正常)} \\ 1 & (x_j > T(\alpha)) \quad \text{(高値)} \end{cases}$$

経過日数 d_j ごとの高異常率 (HR)，低異常率 (LR)，正常率 (NR) は以下の式で定義される．

$$HR(\alpha, d_j) = \frac{|x_i \in X(\alpha, d_j) \mid C(x_i) = 1|}{|X(\alpha, d_j)|} \quad \text{(高異常率)}$$

$$LR(\alpha, d_j) = \frac{|x_i \in X(\alpha, d_j) \mid C(x_i) = -1|}{|X(\alpha, d_j)|} \quad \text{(低異常率)}$$

$$NR(\alpha, d_j) = 1 - (HR(\alpha, d_j) + LR(\alpha, d_j))$$
$$= \frac{|x_i \in X(\alpha, d_j) \mid C(x_i) = 0|}{|X(\alpha, d_j)|} \quad \text{(正常率)}$$

ここで，$X(\alpha, d_j)$ は経過日数 d_j における検査結果の集合を表す．

3.4 可 視 化

本研究では，頻出医療指示パターンおよび検査結果の異常値発生率の解析結果を可視化し，診療プロセスの特性および分岐点を視覚的に示す手法を提案する．

本可視化手法では，各ノードが医療指示を表し，同じ頻出医療指示パターンのノード同士はエッジで繋がっている．ノードのサイズは一律であり，サポート値 $Sup(\alpha)$ に応じて垂直方向に配置される．サポート値が高いノードは上部に，サポート値が低いノードは下部に位置し，診療プロセス内で特に重要なパターンが強調されるように設計されている．

頻出医療指示パターンの可視化に加えて，検査結果の異常値発生率も棒グラフ形式で可視化される．特定の異常値発生率のバーにカーソルを合わせることで，ツールチップに異常値発生情報や該当する頻出医療指示パターンのサポート値が表示される．また，ホバーされた異常値発生率の棒グラフと同様のパターンに対応するノードのみが強調され，視覚的に際立つレイアウトとなる．これにより，診療プロセス内で特定の異常が集中している箇所を迅速に特定できる．

さらに，可視化画面では特定の検査項目を選択して表示を絞り込む機能も備えており，診療プロセス内で異常が集中している特定の検査項目を個別に確認することが可能である．これにより，特定の検査における異常傾向や時系列的な変動が視覚的に明確になり，異常の発生原因を詳細に解析できる．

この可視化手法は，診療プロセスの解析結果を迅速に把握し，検査結果の異常が集中する箇所や重要な診療フローを特定するための支援ツールとして機能する．

4 評 価 実 験

本節では，提案手法の有用性および実現可能性を検証するために行った評価実験について述べる．

4.1 実験の目的

本研究の評価実験では，提案手法の有用性および実現可能性を検証することを目的とする．特に，電子カルテデータを用いて頻出医療指示パターンを抽出し，それに関連する検査結果や投薬データの統計情報が正確に算出されるかを確認する．また，頻出医療指示パターン間の差異（バリアント）が適切に可視化されるかについても評価を行う．これにより，提案手法が診療プロセスの解析および標準化に寄与する可能性を検証する．

本実験では，提案手法の有効性を以下の観点から検証する．

4.1.1 頻出医療指示パターン間の差異 (バリアント) の可視化

頻出医療指示パターン間の差異が適切に可視化されるかを検証する．診療フローにおける分岐点やパターンのバリアントが適切に表現されることで，医療従事者が診療の流れや異常箇所を直感的に把握できるかを確認する．

4.1.2 統計情報付き頻出医療指示パターンの抽出

統計情報付き頻出医療指示パターンが適切に抽出できるかを確認する．頻出医療指示パターンの抽出精度は，提案手法の根幹をなす要素であり，医療指示のシーケンスマイニングが正確に行われるかが問われる．

4.1.3 検査結果の異常値発生率の算出

頻出医療指示パターンに関連する検査結果の異常値発生率が正確に算出されるかを評価する．抽出された頻出医療指示パターンに紐づく検査データが適切に集計されることで，診療プロセスにおける異常値の発生傾向を定量的に把握できる．さらに頻出医療指示パターンに含まれる患者と含まれない患者の異常値発生率を比較することで，患者群ごとの特徴や傾向を明確にする．

4.2 実験データ

本実験では，複数の医療機関から提供された電子カルテデータを参考に 10,000 人分の人工データセットを作成し，提案手法の検証を行った．電子カルテデータの分析結果をもとに，患者情報，投薬情報，検査結果などの標準的な医療データ項目を抽出し，これらを組み合わせたデータスキーマを設計した．作成したデータスキーマは，実際の診療プロセスを再現することを意図しており，入院日や退院日が設定されるとともに，手術，投薬，検査などの医療指示が時系列的に適切に配置されるように調整されている．さらに，検査結果には正常値および異常値が含まれるように設計し，一定のランダム性を加えることで，多様な診療データパターンを再現する人工データを生成した．

なお，本実験で用いた人工データセットには医療機関の識別情報は含まれていないが，将来的には医療機関情報を含むデータセットを使用し，医療機関ごとの頻出医療指示シーケンスを分析することで，施設間での診療プロセスの違いや共通点を明確にする予定である．これにより，異なる医療機関間での診療プロセスの標準化や，施設ごとの特異なパスの発見が期待される．

表 1 に示すのは，作成した人工データセットのスキーマである．患者 ID を中心に各テーブルが連携し，傷病情報や検査情報，手術情報などが管理される設計となっている．この構造により，患者ごとに入院，検査，手術，投薬といった診療プロセスが日付順に記録され，診療の流れを分析する基盤データとして機能する．

表 2 には，作成した人工データセットに含まれる疾患ごとの患者数を示している．疾患ごとの患者数が適切に分布されており，解析の際に特定の疾患に偏らず，多角的な視点で頻出医療

表 1　人工データセットのスキーマ

テーブル名	データスキーマ
患者情報	患者 ID, 年齢, 性別, BMI
日付情報	患者 ID, 日付, 入院日, 退院日
傷病情報	患者 ID, 入院日, ICD-10 コード
手術情報	患者 ID, 手術日, DPC コード+K コード
薬剤情報	患者 ID, 投与日, 投与の合計数, 薬効分類コード
検査情報	患者 ID, 測定日, 検査名, 検査値, 単位, 検査回数
診療行為情報	患者 ID, 実施日, レセプト電算コード
バイタル情報	患者 ID, 測定日, バイタルサイン項目名, バイタルサイン測定値, 単位

表 2　人工データ：疾患ごとの患者数

疾患名	患者数 (名)
肺炎等	1,101
狭心症, 慢性虚血性心疾患, 経皮的冠動脈形成術等	1,100
肝・肝内胆管の悪性腫瘍, 肝悪性腫瘍ラジオ波焼灼療法	1,081
肺の悪性腫瘍手術, 肺葉切除又は 1 肺葉を超えるもの等	1,081
膀胱腫瘍, 膀胱悪性腫瘍手術, 経尿道的手術	1,052
椎間板変性, ヘルニア, 内視鏡下椎間板摘出 (切除) 術	1,013
急性心筋梗塞, 再発性心筋梗塞, 経皮的冠動脈形成術等	1,012
胃の悪性腫瘍, 内視鏡的胃, 十二指腸ポリープ, 粘膜切除術	1,011
脳梗塞	602
子宮頸, 体部の悪性腫瘍, 子宮悪性腫瘍手術等	484
乳房の悪性腫瘍, 乳腺悪性腫瘍手術, 乳房部分切除術	463
合計	10,000

表 3　実 験 環 境

項目	**内容**
プログラミング言語	Python
ライブラリ	PrefixSpan 実装
データベース	PostgresSQL ver. 16.6
データベースドライバー	Psycopg

指示シーケンスを検証できる．

4.3　実 験 環 境

本研究では，頻出医療指示パターンの抽出に PrefixSpan を拡張し，T-PrefixSpan アルゴリズムに基づく独自実装を用いた．既存の T-PrefixSpan には公式ライブラリが存在しないため，PrefixSpan ライブラリを改良し，T-PrefixSpan と同様の動作を実現した．これにより，順序だけでなく医療指示の発生間隔を反映したパターン抽出が可能となった．

実験環境を表 3 に示す．

4.3.1　データベースアクセス

人工データセットは PostgreSQL データベースに格納し，Python を用いて処理を行った．データベースアクセスには Psycopg ライブラリを使用し，大規模データに対して高速なクエリ処理を実現した．これにより，複数疾患のデータや医療指示シーケンスを効率的に処理し，頻出医療指示パターン抽出のパフォーマンス向上を図った．

4.3.2　頻出医療指示パターン抽出

頻出医療指示パターンの抽出には，PrefixSpan ライブラリ [16] を改良し，T-PrefixSpan [10] と同様の機能を実装した．T-PrefixSpan では，基準日からの経過日数と医療指示のセットを 1 つのアイテムとして扱い，頻出医療指示パターンを抽出する．このアプローチにより，診療プロセスにおける時間的要素が保持され，異なるタイミングで発生する医療指示も頻出医療指示パターンとして抽出可能となる．

本研究では，ICD-10 コードごとに医療指示シーケンスをグループ化し，それぞれに対して頻出医療指示パターンの抽出を行った．頻出医療指示パターンの抽出基準としては，疾患ごとのシーケンス数に対する一定割合を閾値（minSup）として設定し，この閾値を超えるパターンのみを抽出対象とした．

4.4　実 験 方 法

評価実験は以下の手順で実施される．まず，実験データセットから疾患ごとの医療指示シーケンスを作成する．次に，T-PrefixSpan を用いて頻出医療指示パターンを抽出する．抽出されたパターンに対して，検査結果や投薬情報の統計を算出し，各経過日数における検査結果の異常値発生率を求める．その後，得られた頻出医療指示パターンおよび検査結果の異常値発生率を可視化し，パターン間の差異や診療プロセスの特異点が視覚的に明確になるかを確認する．これにより，提案手法が診療プロセスの解析と改善に寄与するかを総合的に検証する．

4.5　実 験 結 果

本節では，提案手法に基づき抽出した頻出医療指示パターンおよび検査結果の異常値発生率の可視化結果について示す．検証項目で述べた「頻出医療指示パターン間の差異 (バリアント) の可視化」,「統計情報付き頻出医療指示パターンの抽出」,「検査結果の異常値発生率の算出」が，可視化によってどのように表現されるかを確認する．

4.5.1　頻出医療指示パターンの可視化

図 1 に示すのは，頻出医療指示パターンの可視化結果である．各ノードは診療プロセスの特定の医療指示を表し，ノードを繋ぐエッジは医療指示の流れを示している．サポート値の高い頻出医療指示パターンは上部に配置され，診療プロセス内で重要な流れが直感的に把握できる．サポート値の低い分岐は下部に配置されており，例外的な診療フローや特殊なケースも視覚的に明確である．

図 1 は,「胃の悪性腫瘍，内視鏡的胃，十二指腸ポリープ・粘膜切除術」に関する医療指示パターンの可視化例である．この結果は，対象疾患の全シーケンス数に対して閾値（minSup）を 10% に設定し，クローズドパターンとして抽出された結果を反映している．

本結果は，**「統計情報付き頻出医療指示パターンの抽出」**に関連し，提案手法によって頻出医療指示パターンやそれに伴う統計情報が適切に抽出されたことを確認できる．重要な診療フローが明確に視覚化され，例外的な診療プロセスも下位に配置

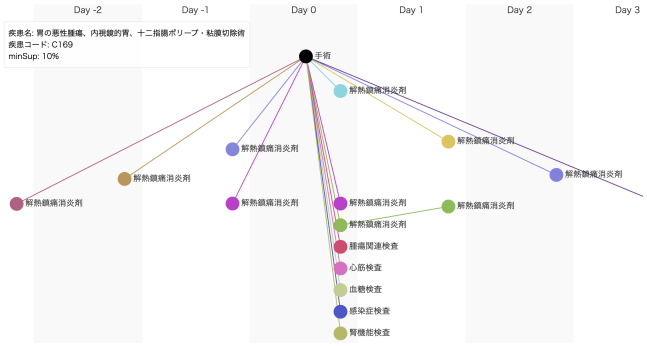

図 1　頻出医療指示パターンの可視化

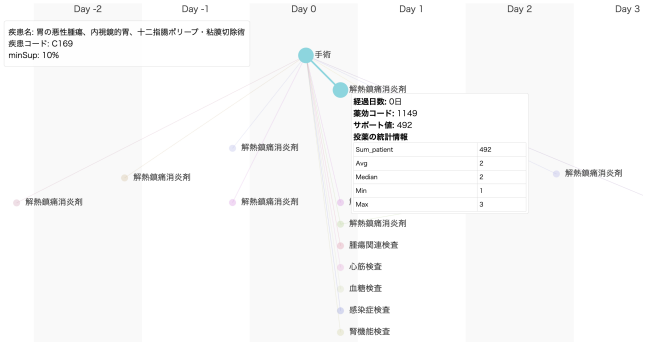

図 2　ノードクリック時の統計情報表示と強調機能の例

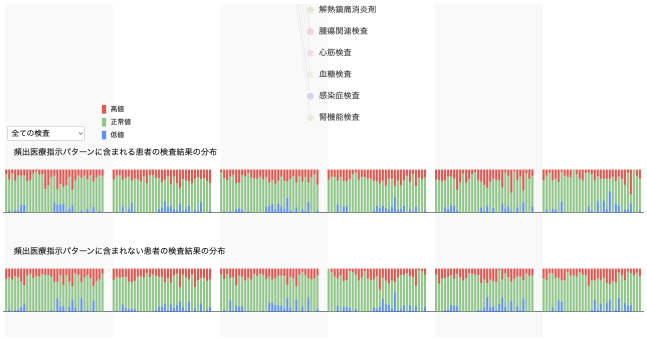

図 3　検査結果の異常値発生率の可視化（デフォルト状態）

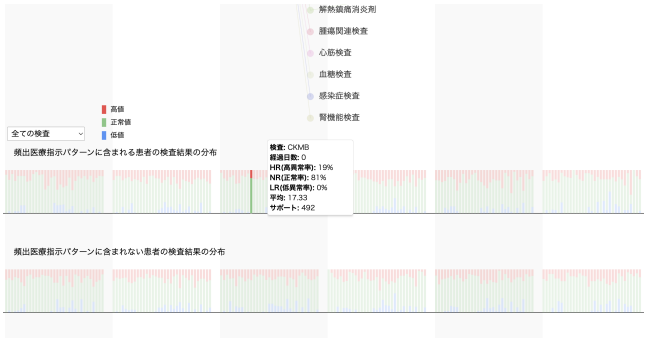

図 4　検査結果の異常値発生率の可視化（ホバー状態）

されることで，診療プロセスの分岐が的確に表現されている．これにより，**「パターン間の差異 (バリアント) の可視化」**の検証にも寄与している．

図 2 はノードをクリックした際の様子を示している．クリックしたノードに関連する検査結果や経過日数などの統計情報がツールチップとして表示され，即座に詳細を確認できる．

ツールチップには，ホバーしたノードの基準日からの経過日数，薬効コード，そのパターンのサポート値（Support）という基本情報が記載されている．さらに,「投薬の統計情報」として，そのパターンに関連する投薬データが表示されており，該当する患者数（Sum patient）が 492 名であること，投薬の平均値（Avg）が 2.0，中央値（Median）が 2.0，最小値（Min）が 1.0，最大値（Max）が 3.0 であることが確認できる．この情報により，ノードが示す医療指示の頻度や関連する患者数が直感的に把握できるだけでなく，診療プロセス内での患者の行動パターンや医療指示の特徴が一目で理解できる．

同時に，ホバーしたノードが含まれる頻出医療指示パターン全体が強調され，診療プロセスの分岐や頻出医療指示パターンの特徴が視覚的に強調される．

4.5.2　検査結果の異常値発生率の可視化

図 3 は，検査結果の異常値発生率を経過日数ごとに示したものである．バーは各経過日数に対応して配置されており，診療プロセスに沿った時系列的な視覚化が可能である．検査結果の異常値発生率は，高値が赤，低値が青，正常値が緑で示され，それぞれの割合が一目でわかるように設計されている．また，経過日数ごとに頻出医療指示パターンに含まれるものと含まれないものを分けて表示しており，診療過程の特定のタイミングで異常値がどの程度発生しているかが明確に把握できる．

本結果は，**「統計情報の算出精度」**に関連し，検査結果における異常値の発生頻度が正確に算出されていることを示している．異常値が特定の頻出医療指示パターンと関連している場合は，そのノードも強調されるため，頻出医療指示パターンと異常値の関連性を直感的に把握できる．

図 4 では，特定のバーにカーソルを当ててホバーした状態を示している．ホバー操作により，該当する頻出医療指示パターンのサポート値や検査結果の異常値の割合がツールチップとして表示され，詳細な情報を確認することができる．この例では，ツールチップに「検査： HbA1c」と表示されており，血糖値コントロールを目的とした血液検査の結果から異常値が検出されていることを示している．この機能により，異常値発生の傾向が把握しやすくなり，診療プロセスの改善に向けた具体的な指針が得られる．

これらの結果は，提案手法が**「検査結果の異常値発生率の算出」**の検証を満たし，検査結果の異常値の発生率が適切に可視化されていることを示している．この可視化機能により，異常値発生の分布や診療プロセス内のリスクポイントが視覚的に特定され，診療の質向上やリスク低減に向けた適切な対策が可能となる．可視化ツールは，単なる結果の表示に留まらず，医療従事者がデータを迅速に解釈し，実際の診療プロセス改善につなげるための強力な支援ツールとして機能する．

5　終わりに

5.1　まとめ

本研究では，頻出パターンマイニングを用いて複数の医療機

関から収集した電子カルテデータを解析し，頻出する医療指示パターンを抽出・分析する手法を提案した．従来の電子カルテデータ解析では，単一医療機関のデータや特定の疾患に限定した解析が行われ，統計情報の抽出は十分に行われていなかった．本研究では，複数の医療機関から提供されたデータを統合し，複数の疾患を対象とすることで，より包括的な解析を実現した．さらに，シーケンス ID（SID）を保持することで，頻出医療指示パターンに該当するシーケンス全体を特定し，詳細な解析が可能となった．

さらに，頻出医療指示パターンに対して検査結果や投薬情報を紐づけ，異常値を含む医療指示の差異（バリアント）を可視化する仕組みを構築した．これにより，診療プロセスにおける標準的なパスだけでなく，異常の多い分岐点や例外的なプロセスも視覚的に把握できるようになった．

加えて，抽出したデータを JSON 形式で出力し，可視化および解析が容易に行える環境を整備した．本手法により，医療現場での診療プロセスの標準化や，個々の患者に応じた柔軟な診療方針の選択が促進されることが期待される．

5.2 今後の課題

今後の課題としては，まず本研究で構築した手法を実際の医療データに適用し，その有効性を検証することが挙げられる．特に，より大規模な医療データを対象に解析を行い，抽出された頻出医療指示パターンおよび検査結果の異常値の発生傾向の再現性や精度を確認する必要がある．

また，頻出医療指示シーケンスの分岐やバリアントが生じる要因の推定についても今後の重要な研究課題である．患者の年齢や疾患歴といった背景情報や，医療機関ごとの診療プロセスの違いが分岐点にどのように影響を与えるのかを詳細に分析することで，診療プロセスの標準化や改善に向けた新たな知見が得られると考えられる．

加えて，本研究で得られた結果を医療従事者と共有し，現場でのフィードバックを受けながら，より実用的で汎用性の高いシステムへと発展させることも求められる．最終的には，診療プロセスの最適化や診療の質の向上に資するツールとして，本システムを医療現場で活用が可能になることを目指す．

謝　　辞

本研究の一部は日本学術振興会科学研究費 (#24K02943) の助成からの支援によって行われた．

文　　献

[1] 吉原博幸. 千年カルテプロジェクト：本格的日本版 EHR と 医療データの 2 次利用に向けて, 第 60 巻. 2018.

[2] 鎌谷直之. 個別化医療とバイオインフォマティックス. 日本生体医工学会誌, Vol. 44, No. 3, pp. 422–428, 2006.

[3] 中谷中. オーダーメイド医療と臨床検査. 日本内科学会雑誌, Vol. 102, No. 12, pp. 3103–3109, 2013.

[4] 山田達大, 本田祐一, 萱原正彬, Le HieuHanh, 串間宗夫, 小川泰右, 松尾亮輔, 山崎友義, 荒木賢二, 横田治夫. Sid を保持するシーケンシャルパターンマイニングによるクリニカルパスバリアント分析. In *Proceedings of DEIM Forum*, No. D1-1, 2019.

[5] Yuichi Honda, Muneo Kushima, Tomoyoshi Yamazaki, Kenji Araki, and Haruo Yokota. Detection and visualization of variants in typical medical treatment sequences. In *Proceedings of the 3rd International Workshop on Data Management and Analytics for Medicine and Healthcare in conjunction with the 43rd International Conference on Very Large Data Bases (VLDB'17)*, pp. 88–101, Cham, 2017.

[6] 安光夕輝, Le Hieu Hanh, 松尾亮輔, 山崎友義, 荒木賢二, 横田治夫. クラスタリングを用いた多病院間の頻出医療指示パターン比較. In *Proceedings of DEIM Forum*, No. 5b-6-3, 2023.

[7] 趙子泰, Le Hieu Hanh, 松尾亮輔, 山崎友義, 荒木賢二, 横田治夫. Covid-19 の異なる医療機関と時期における頻出治療パターンの比較. 第 42 回医療情報学連合大会論文集, pp. 887–892, 2022.

[8] Hieu Hanh Le, Tatsuhiro Yamada, Yuichi Honda, Takatoshi Sakamoto, Ryosuke Matsuo, Tomoyoshi Yamazaki, Kenji Araki, and Haruo Yokota. Methods for analyzing medical-order sequence variants in sequential pattern mining for electronic medical record systems. *ACM Transactions on Computing for Healthcare*, Vol. 4, No. 1, March 2023.

[9] 佐々木夢, 荒堀喜貴, 串間宗夫, 荒木賢二, 横田治夫. 電子カルテシステムのオーダログデータ解析による医療行為の支援. In *Proceedings of DEIM Forum*, No. G5-1, 2015.

[10] Keishiro Uragaki, Tomoyuki Hosaka, Yoshitaka Arahori, Muneo Kushima, Tomoyoshi Yamazaki, Kenji Araki, and Haruo Yokota. Sequential pattern mining on electronic medical records with handling time intervals and the efficacy of medicines. In *Proceedings of the first IEEE Workshop on ICT Solutions for Health in conjunction with the 21st IEEE International Symposium on Computers and Communications*, pp. 20–25, 2016.

[11] Rakesh Agrawal and Ramakrishnan Srikant. Fast algorithms for mining association rules in large databases. *In Proceedings of the 20th International Conference on Very Large Data Bases*, pp. 487–499, 1994.

[12] 牧原健太郎, 荒堀喜貴, 渡辺陽介, 串間宗夫, 荒木賢二, 横田治夫. 電子カルテシステムの操作ログデータの時系列分析による頻出シーケンスの抽出. In *Proceedings of DEIM Forum*, No. F6-2, 2014.

[13] V. P. Raju and G. S. Varma. Mining closed sequential patterns in large sequence databases. *International Journal of Database Management Systems*, Vol. 7, No. 1, pp. 29–39, 2015.

[14] Hieu Hanh Le, Henrik Edman, Yuichi Honda, Muneo Kushima, Tomoyoshi Yamazaki, Kenji Araki, and Haruo Yokota. Fast generation of clinical pathways including time intervals in sequential pattern mining on electronic medical record systems. In *Proceedings of the 2017 International Conference on Computational Science and Computational Intelligence (CSCI)*, pp. 1726–1731. IEEE, 2017.

[15] Yen-Liang Chen, Mei-Ching Chiang, and Ming-Tat Ko. Discovering time-interval sequential patterns in sequence databases. *Expert Systems with Applications*, Vol. 25, pp. 343–354, 2003.

[16] Jiawei Pei, Jiawei Han, Behzad Mortazavi-Asl, Helen Pinto, Qiming Chen, Umeshwar Dayal, and Mei-Chien Hsu. Prefixspan: Mining sequential patterns by prefix-projected growth. In *Proceedings of the 17th International Conference on Data Engineering*, pp. 215–224, 2001.