頻度と実施時刻によるグループ化を採り入れたシーケンス解析に基づく医療指示推薦

坂本　任駿† 　小林　莉華† 　Le HieuHanh† 　松尾　亮輔†† 　山崎　友義††
荒木　賢二†† 　横田　治夫†

† 東京工業大学 情報理工学院　〒 152-8552 東京都目黒区大岡山 2-12-1
†† 宮崎大学 医学部附属病院 医療情報部　〒 889-1692 宮崎県宮崎市清武町木原 5200
E-mail: †{sakamoto,rika_kobayashi,hanhlh}@de.cs.titech.ac.jp, †yokota@cs.titeh.ac.jp
††{ryosuke_matsuo,yama-cp,taichan}@med.miyazaki-u.ac.jp

あらまし　電子カルテの二次利用として，蓄積された医療情報の解析による有効活用が期待されている．我々は二次利用として，入院から退院までの標準的な医療指示の流れの頻出医療指示シーケンスの抽出や，シーケンスの差異の検出を試みてきた．本研究では，抽出シーケンスの適正化と患者情報を用いた医療指示推薦を行う．具体的にはシーケンシャルパターンマイニングで抽出した頻出シーケンスに対し，医療従事者の情報に基づくデータの前処理による適正化，分岐先頻度と実施時刻情報を用いたグループ化による適正分岐候補選択を行い，頻出シーケンス分岐先の適正候補選択による併合シーケンスを生成する．また，分岐要因である患者情報の拡充し，併合シーケンスの分岐への要因適用による医療指示推薦を提案し，実際の電子カルテを用いて医療指示推薦を行い評価する．

キーワード　電子カルテ，医療指示推薦，シーケンシャルパターンマイニング

1 序　　論

1.1 研究背景

近年，紙のカルテに代わり電子カルテの普及が進み，今後さらに日本全国の電子カルテの普及率が増加していくことが予想される．また，国家レベルでの医療情報管理の必要性が認識され，電子カルテなどの医療・健康に関する記録を全国規模で一元的に集める「千年カルテプロジェクト」[1] が始まっている．これに伴い，蓄積された医療情報の二次利用による，根拠に基づく医療（Evidence-based medicine：EBM) [2] の実践が期待されている．二次利用の例として，特定の病気の患者に対して頻出医療指示シーケンスの抽出・検出による，典型的な医療行為の流れ「クリニカルパス」の生成支援が挙げられる．従来，クリニカルパスの作成は医療関係者自身の医学的経験に基づいて行われており，過去の事例が膨大であるため，人力でクリニカルパスバリアントを収集・分析して改善するのは容易ではなかった．そのため，計算機によって電子カルテをデータ工学の観点から分析し，医療行為改善の支援を行う研究が始まった．

電子カルテから頻出医療指示シーケンスの抽出を行い，従来のクリニカルパスの妥当性を確認することや新たなクリニカルパスの可能性を発見することは非常に有用であり，医療の質と効率の向上が見込まれる．

1.2 関連研究

牧原らの研究 [3] では，電子カルテのアクセスログから，ある患者に対して行った医療指示をアイテム，医療指示の流れをシーケンス，全ての患者の医療指示の流れをデータベースと考えることで，アプリオリアルゴリズム [4] を元にしたシーケンシャルパターンマイニング (以下，SPM) を用いて頻出シーケンスの抽出を行った．

佐々木ら [5] は，飽和オーダー列と呼ばれる概念 [6] の導入で出力の情報量を損なわずに出力数を削減し，タイムインターバル SPM [7](以下，TI-SPM) を PrefixSpan [8] に用いることによって，二つのアイテム間の時間間隔を考慮した抽出を行った．しかし，TI-SPM はタイムインターバル (以下，TI) を人為的に定めるため，最適な時間間隔を定めることが困難であるという問題点があった．また，これらの研究では注射のような薬剤情報を含む医療指示において，投与薬剤の種類の情報が含まれないという問題もあった．

浦垣ら [9] は，アイテムを (大別 Type，詳しい説明 Explain，薬効コード Code，薬剤名 Name) の四つ組によって構成することで，薬剤情報を取り入れた抽出を行った．また，薬剤の効果を表す薬効コードに基づくマイニングを行うことで，薬剤名でのマイニングでは得られなかった結果が抽出可能であると示した．さらにアイテム間の時間間隔を統計処理によって算出することで，予め最適な TI セットを求めなければならないという問題に対処した．

Le ら [10] は，TI-SPM を他の SPM よりも計算時間の短い CSpan [11] に用いた T-CSpan を提案し，T-PrefixSpan に比べて短時間で処理が行えることを示した．

保坂ら [12] は，三つのスコアリング手法を導入することで最適な TI セットを事前に決定し，頻出シーケンスの抽出を行った．しかし，これらの先行研究には医学的に解析の価値が高い，シーケンスバリアントに対しての処理がなされていない，出力結果が頻出シーケンスの文字列であるため，シーケンスバリアントを含む場合の視認性が悪いという二つの問題点があった．

山田ら [13] は，頻出シーケンスがどのシーケンスと対応関係にあるのかという情報を，シーケンスの識別子であるシーケンス ID（以下，sid）を保持しながらマイニングすることによって，頻出シーケンス抽出と同時に取得し，頻出シーケンスごとの安全性や効率性の指標の算出・SV 評価が円滑かつ正確に行える可視化手法を提案した．

本田ら [14] はこの二つの問題解決のため，シーケンス同士の共通部分を医療指示が出される日の情報を考慮して検出していくことによって，共通部分を持つシーケンスの差異「シーケンスバリアント」を検出し，グラフで表現して医療従事者への視認性の高い可視化ツールの提供を行った．加えて，SV に対しての分析ができていなかったため，本田ら [15] は SV によって生じるシーケンスの分岐の原因を多変量解析により要因を推定する研究を行った．

上記の研究では，生じたバリアンスの医療指示を独立なものとしてシーケンス中の分岐として扱っており，連続して医療指示が発生する場合について考慮されていない点，医学的に有用ではない医療指示を含み，医療指示が日付単位で昇順にソートされたデータに対して SPM を適用していたため，抽出した頻出シーケンスに医学的に有用でない医療指示を含んでおり，そして同じ日時で多数医療指示が出されていた場合に医療指示の順序関係が不明確であった点が課題点であった．ただし医学的に有用ではない医療指示とは，医療指示推薦の対象者と考える主治医であることを前提として重要でない医療指示，例えば「看護タスク」や「診療予約」のような医療指示や，対象とする症例に対して重みが低い医療指示である．

Wright ら [16] の研究では，zaki [17] が提案した SPADE(Sequential PAttern Discovery using Equivalence classes）に長さや時間の時間的制約を組み込んだ cSPADE [18] を用いて，頻出パターンを抽出している．SPADE とは ID-list という系列 ID(SID)，時間 (EID)，アイテム集合 (Items) からなるデータベースを構築し，クラス分けされたアイテム間の時間間隔を保持させてマイニングを行う手法で，候補シーケンスをグループごとに分割しグループをメインメモリに格納する事で高速化を実現している．Wright らは糖尿病患者に投薬された投薬履歴を薬効クラスと薬名クラスの場合で 90%をトレーニングセット，10%をテストセットに分割しトレーニングセットに対し，cSPADE で得られたルールを用いてテストセットに対し，投薬を推薦し評価した．しかしこの研究では，推薦をする際に頻出なシーケンスに対し分岐の考慮がされていない点，投薬のみに着目しておりその他医療行為に着目しておらず患者に沿った推薦がなされていない点が課題点として挙げられる．

1.3 本研究の目的

本研究は，電子カルテに記録された医療指示のデータセットから SPM で抽出した頻出シーケンスと患者情報を用いて，患者に沿った医療指示推薦を行うことを目的とする．アプローチとして医療従事者の情報に基づくデータの前処理による適正化と，分岐先頻度と実施時刻情報を用いたグループ化による適正分岐候補の選択を行い，併合シーケンス生成する．また，併合シーケンスの分岐への要因適用による医療指示推薦を行うために，分岐要因となる患者情報の拡充し，分岐に対して多変量解析のロジスティック回帰を用いて分岐の要因を推定する．先に生成した併合シーケンスと推定した分岐要因，患者情報に基づいて実際の電子カルテのデータを用いて医療指示推薦を行い評価する．

1.4 本稿の構成

本稿は以下の通り構成される．2 章では本研究に関連する概念を背景知識として説明する．3 章では提案手法である抽出シーケンスの適正化と患者情報を用いた医療指示推薦について述べる．4 章では，3 章の手法を用いて実際の医療指示データを解析すること医療指示推薦の評価を行う．最後に 5 章でまとめと今後の課題について述べる．

2 背景知識

2.1 SPM

Agrawal らによって提案された SPM は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 [4]．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，sid を組みとする要素からなる．

2.2 TI-SPM

当初 Agrawal らの提案手法 [4] は，アイテム間の時間間隔を考慮していない．例えば，2021 年 1 月 1 日に注射を行い，その翌日に手術を行うようなシーケンスと 2021 年 1 月 1 日に注射を行い，その 1 カ月後に手術を行うシーケンスを同一のシーケンスとみなしていた．この 2 つを異なるものとして扱うべきである医療指示のよう時間間隔が重要なシーケンスに対するマイニング手法として，Chen らは TI-SPM と呼ばれる手法を提案した [7]．TI-SPM は，時間間隔を含んだ SDB D，最小支持度 $MinSup(0 \leqq MinSup \leqq 1)$，事前に設定した時間間隔 TI-セットを入力とすることで，TI-頻出シーケンスを出力する．

2.3 T-PrefixSpan

TI-SPM である I-PrefixSpan [7] はアイテム間の時間間隔が固定となりやすいデータに対して用いられるアルゴリズムであり,TI-セットとして事前に設定した入力が必要となる．このためアイテム間の時間間隔は定めた TI-セットによって変化してしまい, 正確なものにならないという問題点が存在した．浦垣らは TI-SPM の問題点を解決するべく,T-PrefixSpan [9] を導入した．T-PrefixSpan は 2 アイテム間の時間間隔を外れ値処

理を含む統計情報を用いて表現を行った．これにより,TI-セットを事前に決定する必要なく, 時間間隔情報を含んだパターンの抽出を可能とした.

2.4 T-CSpan

T-PrefixSpan は PrefixSpan を元にしているため, 計算時間が長いという問題がある． Le らはその解決のため, CSpan を元にした, T-CSpan [10] を導入した．T-CSpan も T-PrefixSpan と同様に時間間隔は統計情報を用いているため, TI-セットを必要としない． 以下に T-CSpan に関する概念の定義と, T-CSpan の説明を行う.

定義 1. タイムアイテム (i,t)

アイテム集合 I が与えられ，アイテム $i \in I$ の発生時刻が t であるとき，i と t の組 (i,t) を**タイムアイテム**と定義する.

定義 2. タイムシーケンス s

タイムアイテムからなる順列 s を**タイムシーケンス**と定義し，以下のように表す.

$$s =< (i_1,t_1),(i_2,t_2),\cdots,(i_n,t_n) >$$

また，タイムシーケンス s の長さ $length(s)$ を $length(s) \equiv n$ とし，シーケンス $O_s =< i_1,i_2,\cdots,i_n >$ を s の**オリジナルシーケンス**と呼ぶ.

定義 3. 時間間隔 TI_k

タイムシーケンス $s =< (i_1,t_1),(i_2,t_2),\cdots,(i_n,t_n) >$ において, **時間間隔** TI_k を以下のように定義する.

$$TI_k \equiv t_{k+1} - t_k (k = 1,2,\cdots,n-2,n-1)$$

定義 4. タイム SDB D

タイムシーケンス集合 S が与えられたとき，**タイム SDB**D を以下のように定義する.

$$D \equiv \{(sid,s) \mid sid \text{ は識別子}, s \in S\}$$

ただし，任意の 2 要素の識別子 sid は異なる値を持つ.

タイム SDB に含まれる全てのタイムシーケンスのオリジナルシーケンスからなる SDB を**オリジナル SDB** と定義したとき，タイム SDB から抽出されるタイム頻出シーケンス及びを飽和タイム頻出シーケンスを以下のように定義する.

定義 5. タイム頻出シーケンス P

最小支持度 $MinSup$ $(0 \leqq MinSup \leqq 1)$, タイム SDBD が与えられたとき, $P =< i_1,X_1,i_2,\cdots,i_{n-1},X_{n-1},i_n >$ ($\forall j$ i_j はアイテム, $\forall k$ X_k は 5 つの値の組 $(min_k,mod_k,ave_k,med_k,max_k)$) の $O_P =< i_1,i_2,\cdots,i_{n-1},i_n >$ を考えた時，O_P が D のオリジナル SDB の $MinSuup$ において頻出シーケンスであれば，**タイム頻出シーケンス**とする．ただし，$min_k,mod_k,ave_k,med_k,max_k$ は以下に示す.

オリジナルシーケンスを構成したとき，O_P をサブシーケンスとするような D に存在する全てのタイムシーケンス $S =< i'_1,t_1,i'_2,t_2,\cdots,i'_{m-1},t_{m-1},i'_m >$ において，$i_k = i'_{j_k}, i_{k+1} = i'_{j_{k+1}}$ を満たす $k = 1,2,\cdots,n-1$, $1 \leqq j_1 < j_2 < \cdots < j_{n-1} < j_n \leqq m$ を考えたとき，時間間隔 $TI_k = t'_{j_{k+1}} - t'_{j_k}$ の集合 Set_{TI_k} を構成できる．このとき，$X_k = (min_k,mod_k,ave_k,med_k,max_k)$ において，min_k を Set_{TI_k} における最小値，mod_k を Set_{TI_k} における最頻値，ave_k を Set_{TI_k} における平均値，med_k を Set_{TI_k} における中央値, max_k を Set_{TI_k} における最大値とする．ここで, 時間間隔 $X_j = (min_j,mod_j,ave_j,med_j,max_j)$ $(1 \leqq j < n)$ に対して,$min_j = max_j$ が成り立つとき，アイテム i_j 及び i_{j+1} の時間間隔は一定となる．特に $min_j = max_j = 0$ の場合は，同日に起こるとする．また, O_P **を** P **のオリジナルパターン**とする.

定義 6. 飽和タイム頻出シーケンス

タイム SDB D から抽出したタイム頻出シーケンス集合 Σ に属する A に対し，以下の条件を満たす $B \in \Sigma \setminus A$ が存在しないとき，A を**飽和タイム頻出シーケンス**と定義する.

（1） A,B のオリジナルパターンを A',B' としたとき，$A' \subseteq B'$ が成り立つ.

（2） (1) が成り立つとき，$A =< a_1,T_1,a_2,\cdots,a_{n-1},T_{n-1},a_n >$,$B =< b_1,T'_1,b_2,\cdots,b_{m-1},T'_{m-1},b_m >$ としたとき，$a_k = b_{j_k}, a_{k+1} = b_{j_{k+1}}$ となる $k = 1,2,\cdots,n-1$,1 $\leqq j_1 < j_2 < \cdots < j_n \leqq m$ が存在する．このとき，全ての $T_k = (min_k,mod_k,ave_k,med_k,max_k)$, $T'_{j_k} = (min'_{j_k},mod'_{j_k},ave'_{j_k},med'_{j_k},max'_{j_k})$ に対して，$min_k \geqq min'_{j_k}$ かつ $max_k \leqq max'_{j_k}$ が成立する.

（3） $Sup(A) \leqq Sup(B)$

ここでタイム頻出シーケンスのサポート値 $Sup(A)$ を $Sup(A) \equiv |\ \{s \mid s \subseteq S, (sid,S) \in D, sid \text{ は } S \text{ の識別子 }\}\ |$ と定義する.

T-CSpan は SDBD, 最低支持度 $MinSup$ を入力とする．まず D 中の全ての頻出な単一タイムシーケンスを決定する． 次に, 各 k-タイムシーケンスにおいて, 射影データベースを構成し, 射影データベース内で頻出なアイテムを見つけ, タイムシーケンスを生成し, 最後に生成された頻出タイムシーケンスに対して, 各アイテム間の時間間隔の結果を計算する． T-CSpan は, 発生チェックを利用して計算結果に飽和タイム頻出シーケンシャルパターンのみを追加するため, 効率的なアルゴリズムとなっている.

2.5 シーケンスバリアントの検出

シーケンスバリアントとは，複数の頻出シーケンスを比較した際に現れる，同一時刻における別医療指示のことを指す．本田ら [14] は，シーケンスバリアントを検出する方法として，飽和タイム頻出シーケンスのうち，相対処置日毎の医療指示数が同じもの同士の共通部分検出を行った．ただし，医療指示として手術を行った日を相対処置日の基準日 (0 日) とし，手術日と同日に行った医療指示の相対処置日は「0 日目」，手術前日に行った医療指示の相対処置日は「-1 日目」となる．以下に共通部分検出に関連する概念の定義を行う.

定義 7. タイムアイテムブロック B

n 個のタイムアイテム $(i_1,\ t_1)$, $(i_2,\ t_2)$, $\cdots$, $(i_n,\ t_n)$ が，$t_1 = t_2 = \cdots = t_n$ の場合，タイムアイテムの集合 B を**タイムアイテムブロック**と定義し，その発生時刻を t_B とする．このときの n 個のタイムアイテムに順序関係は存在しない．

定義 8. バリアントを考慮した頻出パス集合 e

k 個のタイムアイテムブロックの順列 $< B_1,\ B_2,\ \cdots,\ B_k >$ で任意の $0 \leqq l < m \leqq k$ なる l, m に対して $t_{B_l} \leqq t_{B_m}$ が成り立っているものを**バリアントを考慮した頻出パス集合** e と定義し，バリアントを考慮した頻出パス集合の集合を E と表す．このとき，頻出パスは e 内の全ての B の要素数が 1 であるものである．

2.6 多変量解析によるシーケンスの分岐要因推定

本田ら [15] は，SPM で抽出された頻出医療指示シーケンスに存在するバリアンスが，シーケンスを生成した際に生じる分岐にどのように影響したか要因を推定する手法を提案した．この手法は患者の年齢や入院時期などの静的情報と体温や血圧などの動的情報の両方を因子として取り入れた多変量解析であるロジスティック回帰を行い，分岐に影響した要因を推定するものとなっている．

2.7 SID を保持する SPM

山田ら [13] は，SPM によって得られたデータを分析する際に抽出元のデータベースの情報が必要である場合があることに着目し，SPM 時に SID を保持しながら行う手法を提案した．例として，表 1 のようなタイム SDB D において，最小支持度 $MinSup = 0.4$ におけるマイニングを考える．

表 1　タイム SDB D

sid	タイムシーケンス
1	$<(A,\ 1),\ (B,\ 3),\ (C,\ 7),\ (E,\ 10)>$
2	$<(A,\ 1),\ (B,\ 4),\ (E,\ 7)>$
3	$<(A,\ 2),\ (B,\ 6),\ (B,\ 9)>$
4	$<(A,\ 2),\ (B,\ 5),\ (F,\ 10)>$
5	$<(A,\ 2),\ (B7)>$

タイム頻出シーケンスは，$\{<A>,\ <1,\ 2,\ 3,\ 4,\ 5>\}$, $\{<B>,\ <1,\ 2,\ 3,\ 4,\ 5>\}$, $\{<E>,\ <1,\ 2>\}$, $\{<A,\ (2,\ 3,\ 3,\ 3,\ 5),\ B>,\ <1,\ 2,\ 3,\ 4,\ 5>\}$, $\{<B,\ (3,\ 5,\ 5,\ 5,\ 7),\ E>,\ <1,\ 2>\}$, $\{<A,\ (2,\ 2,\ 2,\ 2,\ 3),\ B,\ (3,\ 5,\ 5,\ 5,\ 7),\ E>,\ <1,\ 2>\}$ となる．また，飽和タイム頻出シーケンスは $\{<A>,\ <1,\ 2,\ 3,\ 4,\ 5>\}$, $\{<B>,\ <1,\ 2,\ 3,\ 4,\ 5>\}$, $\{<A,\ (2,\ 3,\ 3,\ 3,\ 5),\ B>,\ <1,\ 2,\ 3,\ 4,\ 5>\}$, $\{<A,\ (2,\ 2,\ 2,\ 2,\ 3),\ B,\ (3,\ 5,\ 5,\ 5,\ 7),\ E>,\ <1,\ 2>\}$ となる．これにより，SID 情報をもとに電子カルテデータから必要な情報を取得することが可能となる．

3 提 案 手 法

本章では，提案手法の医療従事者の情報に基づくデータの前処理による適正化，分岐先頻度と実施時刻情報を用いたグループ化による適正分岐先候補選択，患者情報を拡充した分岐の要因推定，併合シーケンス，患者情報に基づいた医療指示推薦について明記し本提案手法の概要を説明する．前提として [13] の用いられた SID を保持する SPM を行い，頻出シーケンスを抽出しているものとする．

3.1 医療従事者の情報に基づくデータの前処理による適正化

医療従事者の情報に基づいて以下のデータについて前処理を行う．

- 「手術」が「手術麻酔」の実施時刻後に存在

「手術」の実施時刻は実際に手術が行われた時間でなく，手術室を予約した時間が記載されていたため同じ時間に設定する．「手術麻酔」の実施時刻が不明な場合は「手術」の実施時刻はデータ通りの値を用いる．

- 医学的に有用でない医療指示

主治医を対象に医療指示推薦をする場合に重みが低い「看護タスク」，患者の状態によって処方される「頓用薬剤」について除外．

- 実施時刻不明の医療指示

実施時刻が'9999' と記載され，実施時刻が不明な医療指示については除外．

3.2 分岐先頻度と実施時刻情報を用いたグループ化による適正分岐候補選択を適用した併合シーケンス生成

分岐先頻度と実施時刻情報を用いたグループ化による適正分岐候補選択と併合シーケンス生成について説明する．SPM により抽出された頻出シーケンス集合に含まれる医療指示（以下，アイテム i）について患者数に対する頻度 ρ を算出する．頻出シーケンスで生じた差異中のアイテムの組み合わせ集合からなるアイテム $i_n \ldots$ について頻度閾値を ϕ として，

（1） 適正分岐先候補の選択

1. アイテム $i_n \ldots$ から成る組み合わせ集合をアイテム $i_g \ldots$ とし，2 つ以上のアイテムからなる要素を 1 つのアイテムとしてグループ化する．

2. 頻度 $\rho_{i_g \ldots}$ が $\phi < \rho_{i_g \ldots}$ の時，分岐とする．

3. 頻度 $\rho_{i_g \ldots}$ が $\rho_{i_g \ldots} < \phi$ の時，$i_g \ldots$ を除外．

（2） 実施時刻に基づいたグループ化

アイテムの順序関係，回数が明確でない場合，主処置の実施時刻とアイテムの平均回数をもとにグループ化を行い分岐とする．

という処理を行い，分岐先頻度と実施時刻情報を用いたグループ化による分岐先候補の選択を行う．例として，表 2 のよ

表 2　各アイテム i の頻度 ρ

アイテム i	頻度 ρ
A	0.9
B	0.3
C	0.9
D	0.1
E	0.9
F	0.1
G	0.1
$B,\ F$	0.6
$G,\ E$	0.8

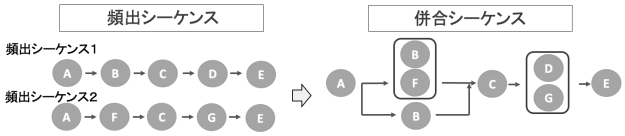

図 1　併合シーケンスの生成

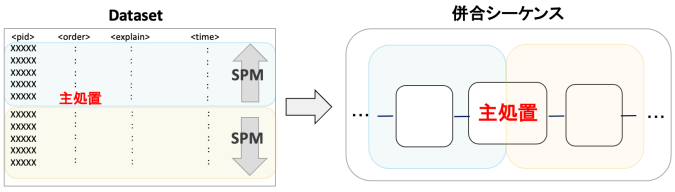

図 2　主処置から前方向と後ろ方向の SPM の組み合わせによる併合シーケンス生成の概要

うな各アイテム i の頻度 ρ の場合を考える．表 2 のような頻度 ρ の時，頻度閾値を $\phi = 0.2$ とし，各アイテムの実施時刻情報を用いることでグループ化されたアイテムの順序を決定する．併合シーケンスは図 1 のようになる．

3.3　主処置から前・後ろ方向への SPM の組み合わせによる併合シーケンス生成

データセットに対して順方向の SPM で頻出シーケンスの抽出が困難な場合に，主処置から前方向・後ろ方向へ SPM を適用する．頻出シーケンスの抽出が困難であるのは対象とする症例によってデータセット全体の医療指示の順序に一貫性がないためである．しかし，主処置の前後の医療指示は頻出かつ類似な順序関係であるために，主処置から主処置以前の医療指示に対する前方向，主処置以降の医療指示に対する後ろ方向へ SPM を適用し，それぞれで得られた併合シーケンスを結合することで全体の併合シーケンスを生成する．主処置から前方向と後ろ方向の SPM の組み合わせによる併合シーケンス生成の概要を図 2 に示す．

3.4　患者情報を拡充した分岐の要因推定

本研究では本田ら [15] の研究と同様，多変量解析による分岐の要因推定を行う．シーケンス中の情報である，シーケンスのどのアイテムを参照しても同じ値が得られる静的情報とシーケンスの参照する要素によって得られる値が異なる動的情報を用いて多変量解析を行うことで分岐の要因を推定する．多変量解

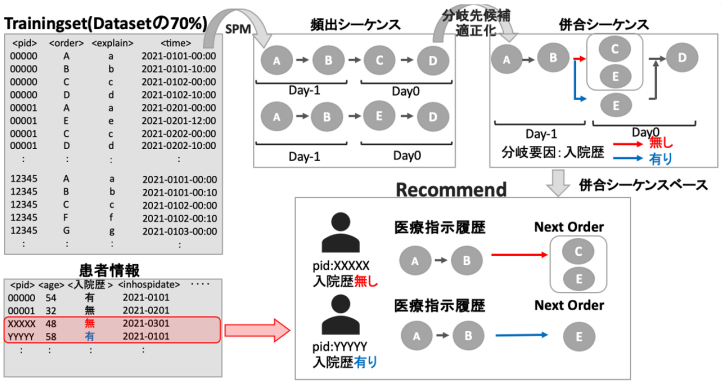

図 3　患者情報と医療指示履歴に基づいた医療指示推薦

表 3　実 行 環 境

CPU	Intel(R) Xeon(R) CPU E5-4650 @ 2.70GHz
Memory	64 GB
OS	Ubuntu 16.04.6 LTS
Java.ver	Java 11.0.3
R.ver	R 3.6.3

析の手段としては，ロジスティック回帰分析を用いる．有意水準を 0.05 とし，有意水準を下回った分岐を有意差を持つ分岐とする． 説明変数として扱う患者情報については，先行研究と同様，静的情報として患者年齢，入院時期，動的情報として体重，体温，収縮期血圧に加え，新たに静的情報に入院歴，動的情報に医療指示実施時刻情報を追加する．

3.5　併合シーケンスと患者情報に基づいた医療指示推薦

医療指示のデータセットを分割したトレーニングセットに対して SPM を行い，得られた頻出シーケンスに 3.2 の分岐先頻度と実施時刻情報を用いたグループ化による適正分岐候補選択を行った併合シーケンスを生成する．また，生成した併合シーケンスに生じた分岐に対して 3.3 の多変量解析による分岐の要因推定を行う．ここで得られた併合シーケンスと分岐の要因，そして患者情報と患者のこれまで適用された医療指示履歴に基づいて医療指示を推薦する．医療指示推薦の概要を図 3 に示す．

4　実　　験

本章では，宮崎大学医学部附属病院から提供された実際の電子カルテデータに対して提案手法を適用し，有効性を確認する．

4.1　実 験 内 容

電子カルテに記載された医療指示データに対し SPM を行い頻出シーケンスを抽出する．次に，前処理による適正化を行ったデータセットに対し提案手法を用いて併合シーケンスを生成し，分岐の要因推定を行う．得られた併合シーケンスと分岐要因，患者情報と医療指示履歴に基づいて医療指示を推薦し評価する．実験環境を表 3 に示す．

4.2　実験対象のデータ

本研究では宮崎大学医学部附属病院の電子カルテシステムに

表 4　TUR_Bt の実験に関する情報

延患者数	408
トレーニングセット (患者数)	286
テストセット (患者数)	122
頻出医療指示列数	22
平均医療指示数	49
平均在院日数	7 日
最小支持度	0.2
頻度閾値 ϕ	0.2

2005 年 1 月から 2018 年 3 月までに記録された，実際に使用されているクリニカルパスを元に行った医療指示データを対象とする．この医療指示データは宮崎大学医学部附属病院で使用されている電子カルテシステム WATATUMI [19](300GB) によって取得されており，個人情報保護の観点より患者を一意に特定できるような情報を含まない．ある患者に対して行った医療指示データを抽出する際には，連結不可能な匿名化患者 ID を用いている．なお，本研究で宮崎大学医学部附属病院の電子カルテデータを医療従事者支援に用いることは宮崎大学の HP [20] に記載がされており，宮崎大学の倫理審査委員会及び東京工業大学の人を対象とする研究倫理審査委員会の承認を得ている．

本実験では電子カルテシステムでクリニカルパス名が TUR と記載された膀胱癌の対して行われる膀胱悪性腫瘍手術 (TUR_Bt) を受けた患者の入院期間中に行われた医療指示のデータセット，クリニカルパス名が「RFA」，「RFA(2cm 以上)」，「RFA(2cm 以下) 」，「RFA>2cm(新) 金入院」，「RFA ≦ 2cm(新) 水入院」，「RFA>2cm(新) 水入院」，「RFA ≦ 2cm(新) 金入院」と記載された肝臓癌に対して行われる計 7 種のラジオ波焼灼療法 (RFA) を受けた患者の外来と入院期間中に行われた医療指示のデータセットを対象として 3 章の提案手法を適用する．

TUR_Bt は [13], [15] で対象とされた症例であること，RFA については複数の主処置が存在する肝臓癌のような症例に対しこれまで着目していなかった外来時の医療指示にも着目し，外来時から入院時にかけての頻出なパターンを抽出することを目的として選択した．RFA の実験では，外来時の医療指示と入院時の医療指示をシーケンス中で区別できるようデータセットに医療指示として「入院日」と「退院日」を加える．

TUR_Bt の実験に関する情報は表 4，RFA の実験に関する情報は表 5 に示す．TUR_Bt に関しては入院時の医療指示履歴のデータセットをトレーニングセット (70%) とテストセット (30%) に分割し，トレーニングセットを用いて，最小支持度を 0.2 として SPM で頻出シーケンス集合を抽出する．得られた頻出シーケンス集合に頻度閾値 $\phi = 0.2$ として提案手法を適用し併合シーケンスを抽出し，併合シーケンスを構成するアイテムに関して，テストセットを用いて分岐要因と患者情報，医療指示履歴に基づいて次に指示される医療指示について推薦する．

表 5　RFA の実験に関する情報

延患者数	224
平均医療指示数	23
平均在院日数	10 日
最小支持度	0.2
頻度閾値 ϕ	0.2

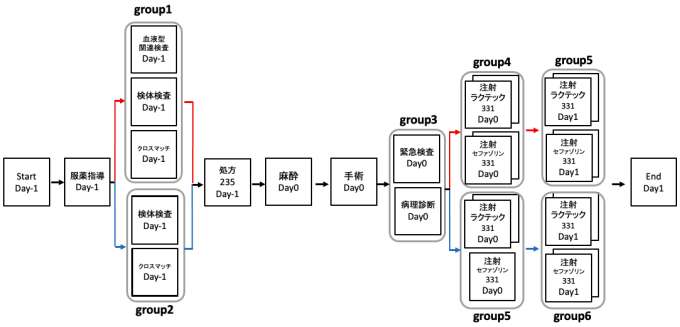

図 4　TUR_Bt の併合シーケンス

表 6　分岐のデータ分布

要因種別	TUR_Bt の入院歴の有無	
入院歴	無	有
患者数	120	163
p 値	0.0437	

4.3　実 験 結 果

4.3.1　併合シーケンス生成と分岐の要因推定

TUR_Bt については，「看護タスク」，「実施時刻が不明」，患者によって処方される「頓用薬剤」を除外して実験を行い，生成した併合シーケンスを図 4 に示す．図 4 に示した通り，入院歴を分岐要因とする分岐や手術時間の実施時刻を分岐要因とする分岐を検出した．図 5 は実験によって有意差あると推定された分岐，図 6 は group3 から group4，5 への分岐部分を拡大し，分岐要因と患者数を付加した図である．また，表 6 は，図 5 における要因種別，患者情報，患者数，p 値をまとめたものである．

図 5 に示した通り，TUR_Bt の入院歴の有無で有意差がある分岐が生じ，初めて TUR_Bt のために入院する患者のみ「血液型関連検査」を行うという結果となった．これは初入院の際の「血液型関連検査」により血液型を判定できるため，入院歴がある患者に対しては医療指示が出されていないと解釈できる．また図 6 に示した通り，手術時間が 13:00 以前・以降に行われるかによって Day0 の注射の平均回数が異なり，有意差は得られなかったが分岐として検出した．医療従事者に TUR_Bt の生成した併合シーケンス，検出した分岐について確認したところ医学的に有用であると評価を得た．

図 7 は RFA の併合シーケンスを示す．RFA は医療指示データを外来から入院，退院に順方向に全体を SPM を適用したところ TUR_Bt とは異なり，全体で医療指示の順序に一貫性がないため頻出なシーケンスを抽出する事が困難であった．そこで，医学的に有用でないデータである「看護タスク」，「実施時刻が不明」，「診療予約」等に加え，医療従事者の情報に基づい

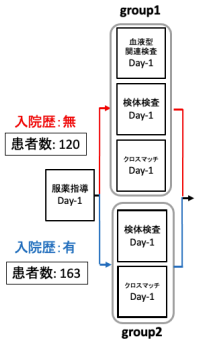

図 5　分岐要因が入院歴の分岐

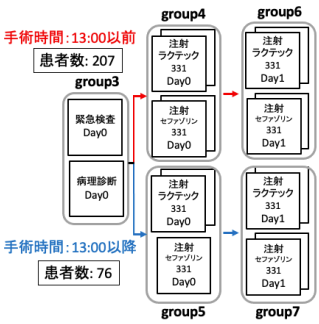

図 6　分岐要因が手術時間の分岐

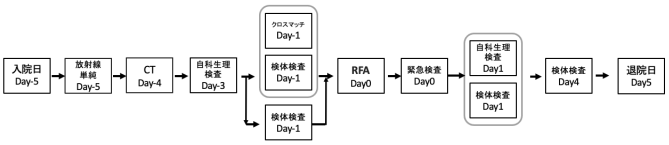

図 7　RFA の併合シーケンス

て「薬剤の処方」,「注射」を除外し，主処置から前・後ろ方向への SPM の組み合わせにより併合シーケンスを生成した．

主処置の医療指示である RFA から前方向の SPM は最小支持度が 0.2，RFA から後ろ方向の SPM は最小支持度が 0.5 で初めて頻出シーケンスが得られた．これは，RFA 以前は検査等の医療指示が多く，医療現場で CT 等の検査は予約可能な日時で医療指示が行われるため，順序に一貫性がない事が影響していると考える．それに比べ RFA 以降の医療指示は，大部分の患者で行われる医療指示が一致しているために頻出シーケンスの抽出が高い最小指示度で抽出できたと考える．

生成した併合シーケンスは Day-3 から Day4 までは RFA7 種類のクリニカルパス記載の医療指示と一致した医療指示となっているが，Day-5 の「放射線単純」と Day-4 の「CT」に関してはクリニカルパスに記載がない医療指示であった．この結果に対し医療従事者に確認したところ，診療オプションの医療指示であり，クリニカルパスと差異が生じていることはクリニカルパス改善に有用であると評価を得た．

4. 3. 2　医療指示推薦

推薦した医療指示を *Precision*，*Recall*，*F* 値 で評価する．ここで次に指示されるべき医療指示を正解セットとすると，正解セットのうち推薦できた医療指示数を TP，正解セットに含まれない医療指示を推薦した医療指示数を FP，正解セットのうち推薦できなかった医療指示数を FN と定義すると，*Precision* と *recall*，*F* 値 をそれぞれ以下のように算出する．

$$Precision = \frac{TP}{TP + FP}$$

$$Recall = \frac{TP}{TP + FN}$$

$$F\ 値 = \frac{2\mathrm{Recall} \cdot \mathrm{Precision}}{\mathrm{Recall} + \mathrm{Precision}}$$

表 7　TUR_Bt の推薦精度

医療指示間	Precision	Recall	F 値
Start to 服薬指導	0.35	0.35	0.35
服薬指導 to group1	0.72	0.70	0.71
服薬指導 to group2	0.65	0.52	0.58
group1 to 処方	0.51	0.48	0.49
group2 to 処方	0.33	0.33	0.33
処方 to 麻酔	0.21	0.20	0.20
麻酔 to 手術	0.60	0.59	0.59
手術 to group3	0.25	0.18	0.21
group3 to group4	0.72	0.47	0.57
group3 to group5	0.69	0.59	0.64
group4 to group6	0.81	0.30	0.45
group5 to group7	0.85	0.31	0.45
Average	0.57	0.42	0.46

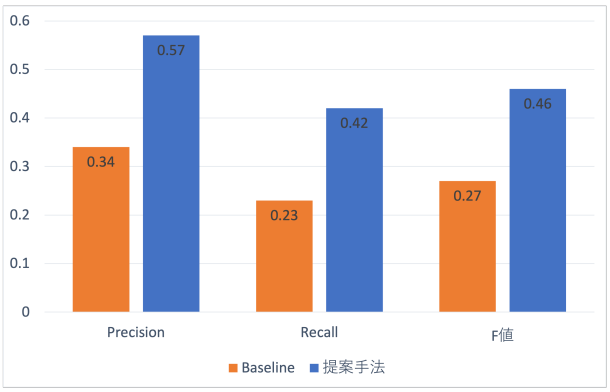

図 8　提案手法と Baseline の推薦精度比較

TUR_Bt を対象として生成した図 4 に示した併合シーケンスと分岐要因，患者情報，医療指示履歴に基づいて医療指示を推薦し，テストセットを用いて *Precision*，*Recall*，*F* 値 を算出したものを表 7 に示す．

表 7 より，服薬指導から group1 のように推薦精度が高い医療指示間が存在した．しかし，処方から麻酔のように，著しく推薦精度が低い医療指示間も存在している．これは併合シーケンスとして抽出できなかった個々の患者の持病のための投薬等の医療指示が実際のデータセットには多数存在していることが原因と考えられる．

次に提案手法と SPM で抽出した頻出シーケンスをベースに推薦を行う Baseline の推薦精度について比較する．この Baseline は関連研究同様に頻出シーケンスを用いて推薦を行う手法で，提案手法の併合シーケンスの生成，分岐の要因推定，患者情報の利用を行わない．提案手法と Baseline で TUR_Bt のデータセットに対し医療指示を推薦した際の *Precision*，*Recall*，*F* 値 を算出し平均したものを図 8 に示す．

図 8 から，Baseline に比べ，提案手法の頻度や医療指示の実施時刻を考慮した併合シーケンスの生成，分岐の要因推定を行い，併合シーケンスと分岐要因を用いて，患者情報を考慮した医療指示推薦が有用であると示せた．

5 結　　論

5.1 ま と め

本研究では電子カルテに記載された症例に対する医療指示のデータセットから，分岐先頻度と実施時刻を用いたグループ化による適正分岐候補選択を行い併合シーケンスを生成し，生成した併合シーケンスと患者情報・医療指示履歴に基づいて医療指示推薦を行う手法を提案し評価を行った．

医療従事者の情報に基づくデータの前処理による適正化と分岐先頻度と実施時刻を用いたグループ化による適正分岐先候補選択，患者情報を拡充した分岐の要因推定を行うことで実施時刻・入院歴を分岐要因とする分岐を検出することができ，医学的に有用な併合シーケンスを生成できた．また，順方向の SPM で困難な頻出シーケンスに対し，主処置から前・後ろ方向への SPM の組み合わせにより併合シーケンス生成を可能とした．

併合シーケンス・患者情報に基づいた医療指示推薦では，頻出シーケンスに基づいた推薦である Baseline に比べ提案手法は推薦精度が向上した．

5.2 今後の課題

今後の課題として，対象とする症例の患者が受けた医療指示のデータセットに対し，外来時の医療指示を含んだ頻出シーケンスの抽出し，外来時の医療指示と入院時の医療指示の相関関係の考察が必要である．加えて他病院のデータセットで得られる頻出シーケンスを比較して病院毎で行われる医療指示の差異を考察することも展望とする．

謝　　辞

本研究の一部は，日本学術振興会科学研究費補助金基盤研究 (B)(#20H04192) の助成により行われた．また，本研究は宮崎大学医学部附属病院の電子カルテデータを用いている．これは宮崎大学の HP [20] に記載されており，宮崎大学の倫理審査委員会及び東京工業大学の人を対象とする研究倫理審査委員会の承認を得ている．関係者各位の協力に感謝する．

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. 情報管理，vol.60，no.11，pp. 767-778，2018.

[2] G.Guyatt. Evidence-based medicine. ACP J Club，vol.114，no.2，pp.A16，1991.

[3] 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. DEIM Forum 2014，F6-2，2014.

[4] R. Agrawal，R. Srikant. Fast algorithms for mining association rules in large databases. Proceeding of the 20th International Conference on Very Large Data Bases，pp. 487-499，1994.

[5] 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. DEIM Forum 2015，G5-1，2015.

[6] X. Yan，J. Han，R.Afshar. CloSpan: Mining closed sequential patterns in large databases. Proc.SIAM Int'1 Conf. Data Mining (SDM '03)，pp. 166-177，May 2003.

[7] Y. Chen，M. Chiang，M. Ko. Discovering time-interval sequential patterns in sequence databases. Expert Systems with Applications 25，pp. 343-354，2003.

[8] J. Pei，J. Han，B. Mortazavi-Asl，H. Pinto，Q. Chen，U. Dayal，M. Hsu. PrefixSpan: Mining sequential patterns efficiently by prefix-projected pattern growth. Proceeding of 2001 international conference on data engineering，pp. 215-224，2001.

[9] K. Uragaki，T. Hosaka，Y. Arahori，M. Kushima，T. Yamazaki，K. Araki，H. Yokota. Sequential pattern mining on electronic medical records with handling time intervals and the efficacy of medicines. First IEEE workshop on ICT solutions for health，proc. 21st IEEE international symposium on computers and communications，pp. 20-25，2016.

[10] Hieu Hanh Le，Henrik Edman，Yuichi Honda，Muneo Kushima，Tomoyoshi Yamazaki，Kenji Araki，Haruo Yokota.，"Fast Generation of Clinical Pathways Including Time Intervals in Sequential Pattern Mining on Electronic Medical Record Systems." Proceeding of the fourth International Conference on Computational Science and Computational Intelligence (CSCI 2017)，，pp. 1726-1731，2017.12.

[11] V. P. Raju，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.

[12] 保坂智之，浦垣啓志郎，荒堀喜貴，串間宗夫，山崎友義，荒木賢二，横田治夫. 医療履歴の時系列解析におけるシーケンス間類似度評価による時間間隔調整の導入 DEIM Forum 2016，G7-5 2016.

[13] 山田達大，本田祐一，萱原正彬，Le Hieu Hanh，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.

[14] Y. Honda，M. Kushima，T. Yamazaki，K. Araki，H. Yokota. Detection and visualization of variants in typical medical treatment sequences. Proceeding of the 3rd VLDB workshop on data management and analytics for medicine and healthcare. Springer，pp. 88-101，2017.

[15] 本田祐一，山田達大，萱原正彬，Le Hieu Hanh，串間宗夫，小川泰右，松尾亮輔，山崎友義，荒木賢二，横田治夫. 患者の固有情報及び動的状況を考慮したクリニカルパス分岐要因推定. DEIM Forum 2019.

[16] Wright，Aileen P and Wright，Adam T and McCoy，Allison B and Sittig，Dean F. The use of sequential pattern mining to predict next prescribed medications. Journal of biomedical informatics，vol.53，pp.73-80，2015.

[17] Zaki，Mohammed J. SPADE: An efficient algorithm for mining frequent sequences. Machine learning，vol.42，pp.31-60，2001.

[18] Buchta C，Hahsler M，Buchta MC. Package 'arulesSequences'；2012

[19] 電子カルテシステム WATATUMI. http://www.corecreate.com/02_01_izanami.html

[20] 宮崎大学医学部附属病院医療情報部. http://www.med.miyazakiu.ac.jp/home/jyoho/

[21] 診療報酬情報提供サービス 基本マスター. http://www.iryohoken.go.jp/shinryohoshu/kaitei/

[22] Z. Huang，X. Lu，H. Duan. On mining clinical pathway patterns from mediacal behaviors. Artificial intelligence in medicine 56 (2012) 35-65，2012.

[23] DPC Web 辞書：診断群分類の解説. http://bone.jp/dpc/