

典型的医療指示シーケンスの共通部分検出とその可視化

本田 祐一[†] 串間 宗夫^{††} 山崎 友義^{††} 荒木 賢二^{††} 横田 治夫^{†††}

[†] 東京工業大学 情報工学科 〒152-8565 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 〒889-1601 宮崎県宮崎市清武町木原 5200

^{†††} 東京工業大学 情報理工学院 〒152-8565 東京都目黒区大岡山 2-12-1

E-mail: †honda.y.ai@m.titech.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、医療指示の履歴にシーケンシャルパターンマイニングを適用することで、医療指示の典型シーケンス抽出による医療従事者に対する支援を試みてきた。本研究では、先行研究によって得られる典型シーケンスの共通部分検出を行い、医療指示の変形パターン（バリエーション）を検出し、バリエーションパターンという形でグラフ表現する。さらに文字列による出力結果から視認性の高いWebアプリケーションによる可視化のツールの作成を行う。

キーワード シーケンシャルパターンマイニング, 電子カルテ, クリニカルパスバリエーション, クリニカルパス可視化

1. はじめに

1.1 研究背景

近年、病院での電子カルテの導入が進んでいる。電子カルテを利用することによって、対象となる患者の検索やオーダーと呼ばれる医療指示の履歴閲覧の高速化を可能とし、医療指示全体の標準化・改善に貢献している。また、電子カルテは診療記録として用いるだけでなく、二次利用としての情報活用が期待されている。

二次利用の例としては、ある病気の患者に対してクリニカルパスと呼ばれる医療指示（オーダー）列の典型例を抽出することが挙げられる。従来の典型クリニカルパスは、医療従事者の医学的経験に基づいて作成されていたが、人が作成するのは容易ではなかった。このような背景より、電子カルテをデータ工学の観点において分析し医療指示の改善の支援を目的とした研究が行われている。電子カルテのデータを分析する研究により、医療指示履歴から従来の典型クリニカルパスの妥当性の判断をすることは有用であり、また新たな医療指示履歴の変形パターン「クリニカルパスバリエーション」の検出は医療指示改善への支援となることが見込まれる。

1.2 先行研究

牧原らの研究 [1] では、電子カルテの履歴から、ある患者に対して行った医療指示をアイテム、医療指示の履歴をシーケンス、全ての患者の医療指示の流れをデータベースと捉えることで、アプリアリアルゴリズム [2] を元にしたシーケンシャルパターンマイニング（以下、SPM）を適用し、頻出なクリニカルパスの抽出を行った。牧原らは、手術といった重要な医療指示を基準として、その前後の部分シーケンスでそれぞれにマイニングを行った。これにより、基準の医療指示の前後で行うべき医療指示を抽出出来ることを示した。しかし、この手法には基準となる医療指示後の部分でのマイニング結果の出力数が膨大になってしまうことと、医療指示間の時間間隔が考慮されていないという二つの問題点があった。

佐々木ら [3] は、これら二つの問題点の解決のために、飽和オーダー列と呼ばれる概念 [4] 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 [5]（以下、TI-SPM）を PrefixSpan [6] に用いることによって、二つのアイテム間の時間間隔が考慮された抽出を行った。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に、定めた時間間隔によって結果が異なってしまうという問題点があり、最適な時間間隔を定めなければならない。また、これらの研究では注射のように薬剤情報を含むような医療指示において、どの薬剤を投与するのかの情報が含まれない、という問題もあった。

浦垣ら [7] は、これらの問題の解決のために、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を抽出可能であると示した。さらに、時間間隔の設定の問題に対しては、予め TI セットを決めるのではなく、統計処理によって定めることで問題解決を図った。保坂ら [8] は、三つのスコアリング手法を導入し、そのスコアの結果によって最適な TI セットを決定することで、最適な TI セットを事前に決定し、頻出なクリニカルパスの抽出を行った。

しかし、これらの先行研究の共通な問題点としてはクリニカルパスバリエーションに対しての処理がなされていない点が挙げられる。クリニカルパスの変形パターンであるバリエーションの検出により、従来の典型的なクリニカルパスの改善へ支援ができる可能性がある。さらに、先行研究の出力はどれも頻出なシーケンスの文字列であるために、クリニカルパスバリエーションを含む場合の視認性が悪く、バリエーション検出をするためには、複数の抽出したシーケンス同士の差異を目視による確認が必要で分析には時間を要することが挙げられる。

1.3 本研究の目的

本研究は、電子カルテシステムに記載されたある症例に対する医療指示から頻出シーケンシャルパターンを抽出し、飽和の

概念を導入して、出力シーケンスの削減を行ったものに対して、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り、クリニカルパスバリエーションの検出をすることを目的とする。また、クリニカルパスバリエーションをグラフ表現をし、医療従事者への視認性の高い可視化ツールの提供を目指す。

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. 章ではシーケンス同士の共通部分を検出する手法の導入とその可視化の手法を提案手法として述べる。4. 章では、3. 章の手法を用いて、実症例に対する医療指示データを解析し、得られたクリニカルパスバリエーションを確認と、実験で得られた結果を用いた可視化ツールの紹介を行う。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) から以下によって定義がされる頻出シーケンシャルパターンを抽出する手法である [2]。アイテムの順列をシーケンス、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定する識別子を組みとする要素からなる。

2.2 TI-SPM

当初 Agrawal らが提案手法 [2] では、アイテム間の時間間隔は考慮しない頻出シーケンシャルパターンの抽出であった。例えば、2017 年 1 月 1 日に処方を行い、その翌日に手術を行うようなシーケンスと 2017 年 1 月 1 日に処方を行い、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。医療指示のような順番だけでなく、アイテム間の時間間隔も重要な情報となるシーケンスに対するマイニングの手法として、Chen らは TI-SPM と呼ばれる手法を提案した [5]。この手法により、上記の二つのシーケンスを異なるシーケンスと捉えることが可能となる。

TI-SPM は、時間間隔を含んだ SDB である D、最小支持度 $MinSup(0 \leq MinSup \leq 1)$ 、TI-セットを入力として与えることで、TI-シーケンスから TI-頻出シーケンシャルパターンを得る。

定義 1. TI

$r - 1$ 個の定数 $T_1, T_2, \dots, T_{r-1}$ を元に、**TI** $I_k (k = 0, 1, \dots, r - 1, r)$ は以下により定義される

$$I_k \equiv \begin{cases} \{0\} & (k = 0) \\ \{t \mid 0 < t \leq T_1\} & (k = 1) \\ \{t \mid T_{k-1} < t \leq T_k\} & (k = 2, 3, \dots, r - 1) \\ \{t \mid T_{r-1} < t\} & (k = r) \end{cases}$$

そして、 $r - 1$ 個の定数 $T_1, T_2, \dots, T_{r-1}$ によって構成される $r + 1$ 個の **TI** の集合を **TI-セット** V とする。

定義 2. TI-シーケンス

アイテム集合 I 、TI-セット V が与えられたとき、以下の B を TI-シーケンスと定義する。

$$B \equiv \begin{cases} < b_1 > & (k = 1) \\ < b_1, \&_1, b_2, \&_2, \dots, b_k - 1, \&_{k-1}, b_k > & (k \geq 2) \end{cases}$$

ただし、 $\forall i = 1, 2, \dots, k$ について $b_i \in I$ とし、 $\forall v = 1, 2, \dots, k - 1$ について $\&_v \in V$ とする。

定義 3. TI-サブシーケンス

シーケンス $A = < (a_1, t_1), (a_2, t_2), \dots, (a_n, t_n) >$ と TI-シーケンス $B = < b_1, \&_1, b_2, \&_2, \dots, b_{m-1}, \&_{m-1}, b_m >$ に対して以下を満たす $1 \leq j_1 < j_2 < \dots < j_m \leq n$ となる整数列 $\{j_m\}$ が存在するとき、**B は A の TI-サブシーケンス** といい、 $B \subseteq A$ と表す。

$$(1) b_1 = a_{j_1}, b_2 = a_{j_2}, \dots, b_m = a_{j_m}$$

$$(2) t_{j_i} - t_{j_{i-1}} \in \&_{i-1} \quad (i = 2, 3, \dots, m)$$

定義 4. 射影シーケンス

シーケンス $A = < (a_1, t_1), (a_2, t_2), \dots, (a_n, t_n) >$ と A の TI-サブシーケンスである TI-シーケンス $B = < b_1, \&_1, b_2, \&_2, \dots, b_{m-1}, \&_{m-1}, b_m >$ が $m \leq n$ かつ $a_{i_k} = b_k (1 \leq k \leq m)$ を満たすとき、A のサブシーケンス $A' = < (a'_1, t'_1), (a'_2, t'_2), \dots, (a'_n, t'_n) >$ が次を満たすとき **A' は A の B に関する射影シーケンス** と定義する。

$$(1) n' = n + m - i_m \text{ を満たす } i_m \ 0 \leq i_m \leq n \text{ が存在する}$$

$$(2) B \text{ は } A' \text{ の TI-サブシーケンス}$$

$$(3) A' \text{ の後方 } n - i_m \text{ 個のアイテムと } A \text{ の後方 } n - i_m \text{ 個のアイテムが一致}$$

定義 5. TI-ポストフィックス

シーケンス $A = < (a_1, t_1), (a_2, t_2), \dots, (a_n, t_n) >$ の TI-シーケンス $B = < b_1, \&_1, b_2, \&_2, \dots, b_{m-1}, \&_{m-1}, b_m >$ に関する射影シーケンスを $A' = < (a'_1, t'_1), (a'_2, t'_2), \dots, (a'_n, t'_n) >$ とする。このとき、 $A'_B = < (a'_{m+1}, t'_{m+1}), (a'_{m+2}, t'_{m+2}), \dots, (a'_n, t'_n) >$ を **B に関する A の TI-ポストフィックス** とする。

2.3 T-PrefixSpan

TI-SPM である I-PrefixSpan[5] 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[7] 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。以下に T-PrefixSpan に関する概念の定義と、T-PrefixSpan の説明を行う

定義 6. 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 7. 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.

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

タイムシーケンス s の長さを $length(s)$ を $length(s) \equiv n$ とし, シーケンス $O_s = \langle i_1, i_2, \dots, i_n \rangle$ を s の**オリジナルシーケンス**と呼ぶ

定義 8. 時間間隔 TI_k

タイムシーケンス $s = \langle (i_1, t_1), (i_2, t_2), \dots, (i_n, t_n) \rangle$ において, 時間間隔 TI_k は次で定義される $TI_k \equiv t_{k+1} - t_k$ ($k = 1, 2, \dots, n-2, n-1$)

定義 9. 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** は以下で定義される.

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

ただし, D の任意の 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から構成される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シャルパターン及び飽和タイム頻出シーケンシャルパターンを以下のように定義する.

定義 10. タイム頻出シーケンシャルパターン P

最小支持度 $MinSup$ ($0 \leq MinSup \leq 1$), タイム SDB D が与えられたとき, $P = \langle i_1, X_1, i_2, \dots, i_{n-1}, X_{n-1}, i_n \rangle$ ($\forall j$ i_j はアイテム, $\forall k$ X_k は 5 つの値の組 $(min_k, mod_k, ave_k, med_k, max_k)$) について, シーケンス $O_P = \langle i_1, i_2, \dots, i_{n-1}, i_n \rangle$ を考えた時, O_P が D のオリジナル SDB の最小支持度 $MinSup$ において頻出シーケンシャルパターンであれば, **タイム頻出シーケンシャルパターン**とする

ただし, $min_k, mod_k, ave_k, med_k, max_k$ は以下に示すものとする. 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 $S = \langle i'_1, t_1, i'_2, t_2, \dots, i'_m, t_m \rangle$ において, $i_k = i'_{j_k}, i_{k+1} = i'_{j_{k+1}}$ を満たす, $k = 1, 2, \dots, n-1, 1 \leq j_1 < j_2 < \dots < j_{n-1} < j_n \leq m$ を考えたとき, 時間間隔 $TI_k = t'_{j_{k+1}} - t'_{j_k}$ の集合を 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, $min_k = \min Set_{TI_k}, mod_k$ 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, $mod_k = \max Set_{TI_k}$ とする. ここで, 時間間隔 $X_j = (min_j, mod_j, ave_j, med_j, max_j)$ ($1 \leq j < n$) 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, 同日に発生することを示す.

また, O_P を P の**オリジナルパターン**とする.

定義 11. 飽和タイム頻出シーケンシャルパターン

タイム SDB である D から抽出したタイム頻出シーケンシャルパターン集合 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シャルパターン**と定義する.

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

(2) 条件 (1) が成り立つとき,

$A = \langle a_1, T_1, a_2, T_2, \dots, a_{n-1}, T_{n-1}, a_n \rangle, B = \langle b_1, T'_1, b_2, T'_2, \dots, b_{m-1}, T'_{m-1}, b_m \rangle$ としたとき, $a_k = b_{j_k}, a_{k+1} = b_{j_{k+1}}$ となる $k = 1, 2, \dots, n-1, 1 \leq j_1 < j_2 < \dots < j_n \leq m$ 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, $min_k \geq min'_{j_k}$ かつ $max_k \leq max'_{j_k}$ を成立する.

(3) $Sup(A) \leq Sup(B)$

ここでタイム頻出シーケンシャルパターンのサポート値 $Sup(A)$ を $Sup(A) \equiv |\{s \mid s \subseteq A, (sid, s) \in D, sid \text{ は } S \text{ の識別子}\}|$ と定義される.

T-PrefixSpan は SDBD, 最低支持度 $MinSup$ を入力とする. 初めに D 中のタイム頻出シーケンシャルパターン (α) を求め, それを元に射影 SDB を構成. その後, α により構成された射影 SDB において, $Sup(\beta) \geq Size(D) \times MinSup$ なるタイム頻出シーケンシャルパターン (β) を求め, オリジナル SDB から α, β 間の X を求め, $\langle \alpha, X, \beta \rangle$ をタイム頻出シーケンシャルパターンとして出力する. その後, $\langle \alpha, X, \beta \rangle$ で射影を行う. という操作を繰り返していき, タイム頻出シーケンシャルパターンを求めるというアルゴリズムになる.

2.4 エピソードマイニング

Mannila らによって提案されたエピソードマイニング (以下, EM) は, 単一の長いシーケンスデータから頻出な反順序集合を抽出する手法 [9] である. 出力されるパターンはエピソードと呼ばれる. エピソードは順序が一意に決定はされず, シーケンシャルパターンよりも汎用性が高いものとなる. エピソードは三種類に分けて考えられる (図 1-a,b,c). (a) のシリアルエピソードは, イベント間の順序が一意に決定されるもの. (b) のパラレルエピソードは, イベント間の順序が全てないもの. (c) のジェネラルエピソードは, イベント間に反順序関係があるものである. シリアルエピソードとパラレルエピソードはジェネラルエピソードの特殊な場合であり, また, シリアルエピソードはシーケンシャルパターンと等価である.

本研究で扱う分岐パターンは, SPM によって得られた頻出シーケンシャルパターンに対して, 共通部分検出を行うもののため一般に用いられる EM とは異なる.

3. 提案手法

本章では従来研究にはなかったクリニカルパスバリエーションの検出手法とデータの可視化のためにエピソードの表現方法を参考にしたクリニカルパスバリエーションの表現手法について説明する.

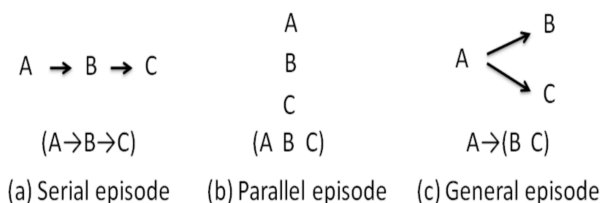


図1 エピソードの例

3.1 時間情報を含むシーケンシャルパターンの共通部分検出

3.1.1 医療指示の取り扱い

本研究では [7] 同様に、マイニングにおいてアイテムをアイテムを (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. 薬剤情報を含まない医療指示においては Code, Name は「null」となる. 例えば薬剤情報を含む「注射」の「静脈内注射」において、薬効コードが「331」の薬剤「ラクテック注」を投与した際は、(注射, 静脈内注射, 331, ラクテック注) となり、薬剤情報を含まない「看護タスク」の「全身清拭」は (看護タスク, 全身清拭, null, null) となる. ここで、薬効コードが「331」の場合の薬効は「血液代用剤」である.

Type が「処方」もしくは「注射」を含まない医療指示で扱っている薬剤は医学的には有用ではないとして、Code, Name は「null」とする.

3.1.2 従来研究の未達成点

従来研究において得られていた、頻出シーケンシャルパターンは各パターンにおいて順序が一意に決定されていたが、パターン同士の関係性については不明であり、分析には人の手により行わなければならないかった. しかし、マイニングの最小支持度によっては出力数は膨大となり、人の手による細かな分析は困難である.

本研究では、T-PrefixSpan[7] を用いて抽出した、タイム頻出シーケンシャルパターンの共通部分検出を行うことにより、タイム頻出シーケンシャルパターン同士の関係性の導出を行う.

3.2 クリニカルパスバリエーションの検出

3.2.1 検出実行の判断基準

クリニカルパスバリエーションとは、複数のクリニカルパ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. クリニカルパスバリエーションを分析することによって、従来医療指示の改善支援を行うことが出来る.

本研究では、飽和タイム頻出シーケンシャルパターンのうち、相対処置日毎の医療指示数が同じものの同士の共通部分検出を行う. 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.

図2は、図中左側のパターン X を図中右側のパターン A, パターン B のどちらと共通部分検出を行うかの例である. パターン X の各相対処置の医療指示数は「-1 日目: 1 / 0 日目: 2 / 1 日目: 2」である. この時パターン A の各相対処置の医療指

示数は「-1 日目: 1 / 0 日目: 1 / 1 日目: 2」と異なるのに対し、パターン B の各相対処置の医療指示数は「-1 日目: 1 / 0 日目: 2 / 1 日目: 2」とパターン X と同様であるため、パターン X はパターン B と共通部分検出を行う.

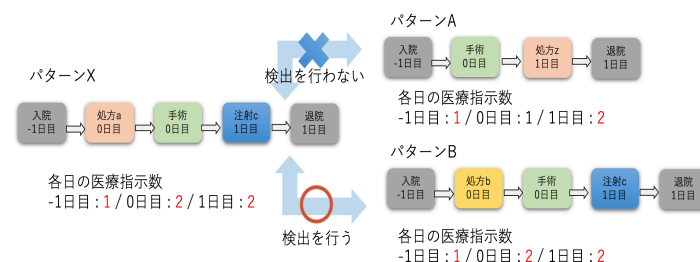


図2 バリエーションパターン例

これはクリニカルパスバリエーション内における差異が小さいバリエーション, 例えば同じ注射の医療指示であり注射する薬剤の効果が同じであるが注射される薬剤名が異なるものを抽出するためである.

3.2.2 検出方法

本研究では、前章で説明した、従来研究の未達成点を補うべく、クリニカルパスバリエーションを検出する方法として、タイム頻出シーケンシャルパターンの共通部分検出の導入を行う. まずはじめに、共通部分検出に関連する概念の定義を行い、共通部分検出の説明を行う. 以下のように、タイムアイテムブロック、バリエーションパターンを定義する.

定義 12. タイムアイテムブロック B

n 個のタイムアイテム $(i_1, t_1), (i_2, t_2), \dots, (i_n, t_n)$ が $t_1 = t_2 = \dots = t_n$ 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 13. バリエーションパターン e

k 個のタイムアイテムブロックの順列 $\langle B_1, B_2, \dots, B_k \rangle$ で任意の $0 \leq l < m \leq k$ なる l, m に対して $t_{B_l} \leq t_{B_m}$ が成り立っているものをバリエーションパターン e と定義する. バリエーションパターンの集合を E と表す. このとき、タイム頻出シーケンシャルパターンは e 内の全ての B の要素数が 1 であるものである.

本研究では、タイム SDB に T-PrefixSpan を適用して得られた飽和タイム頻出シーケンシャルパターンの集合に対して、共通部分検出によるクリニカルパスバリエーション作成のアルゴリズムを導入した. またこのときの、飽和タイム頻出シーケンシャルパターンの集合はパターン長の昇順に整列されているものとする. アルゴリズムは以下の Algorithm 1 の通りである.

Algorithm 1 共通部分検出によるクリニカルパスバリエーション作成

Input: 飽和タイム頻出シーケンシャルパターン集合 P' **Output:** バリエーションパターンの集合 E

```
1:  $e \leftarrow \{p \mid p \in P'\}$ 
2: for  $p \in P' \setminus e$  do
3:    $k, j = 1$ 
4:   while  $k < \text{length}(e)$  かつ  $j < \text{length}(p)$  do
5:      $B_e = \{B_i \in e\}$ 
6:      $B_p = \{B_j \in p\}$ 
7:     if  $t_{B_e} == t_{B_p}$  then
8:       for  $\{i \mid i \in B_e\}$  do
9:         if  $i_{B_p}$  が  $i_{B_e}$  のどれとも一致しない then
10:            $(i_{B_p}, t_{B_p})$  を  $B_e$  に加える
11:         end if
12:       end for
13:        $k = k + 1, j = j + 1$ 
14:     else if  $t_{B_e} > t_{B_p}$  then
15:        $B_e$  の直前に  $B_p$  を挿入する
16:        $k = k + 1, j = j + 1$ 
17:     else
18:        $k = k + 1$ 
19:     end if
20:   end while
21:   while  $j < \text{length}(p)$  do
22:      $B_j, B_{j+1}, \dots, B_{\text{length}(p)}$  を  $e$  の末尾に加える
23:   end while
24: end for
25:  $E \leftarrow e$ 
```

3.2.3 バリエーションパターンの記述

前節で求めたバリエーションパターンの記述方法について述べる。

従来研究で得られる出力は、順序が一意に決まっていることが前提だったため、本研究での出力となるバリエーションパターンの出力特に、タイムアイテムブロック内の要素が複数の場合には対応しておらず、表現出来ない場合が存在した。

そこで、本研究では EM で用いられるエピソードの表現方法を参考にした入れ子構造によって、バリエーションパターンを表現することにする。例えば、 A, B, C, D, E がタイムアイテムとしたとき、バリエーションパターン $e = [[A], [B], [C, D], [E]]$ は図 3 のようなグラフになる。

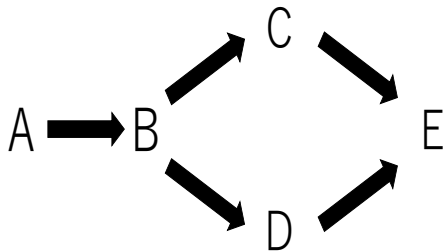


図 3 バリエーションパターン例

3.3 クリニカルパスバリエーションの可視化

3.3.1 文字列による表現の問題点と可視化の意義

先行研究では、結果である頻出シーケンシャルパターンを文字列によって表現していた。しかしこの出力形式は頻出シーケンシャルパターンよりもさらに複雑な構造を含むバリエーションパターンを表現する場合には、視認性が低く得られた結果に対して直感的な理解が難しいという欠点がある。そのため、データ利用者の能力によって得られる情報量が異なってしまうという問題がある。データを可視化することでバリエーションを含む情報の読み取りを正しく円滑に行えるようになり、分析の支援となる。また結果を利用するのは主に医療従事者のため、それに合った形式が求められる。

3.3.2 可視化アルゴリズム

図 2 のような表示を行うためには、バリエーションパターンのノード集合とエッジ集合を得る必要があり、そのアルゴリズムは以下の Algorithm 2 の通りである。

Algorithm 2 バリエーションパターンのノード、エッジ集合の導出

Input: バリエーションパターン e **Output:** e のノード集合 V , エッジ集合 E

```
1:  $k = 1$ 
2:  $count = 1$ 
3: for  $k < \text{length}(e)$  do
4:    $start_{B_k} = count$ 
5:   for  $\{i \mid i \in B_k\}$  do
6:      $i_{count} = i$ 
7:      $V \leftarrow i_{count}$ 
8:     for  $k \geq 2$  かつ  $start_{B_{k-1}} \leq j \leq end_{B_{k-1}}$  do
9:        $E \leftarrow (i_j, i_{count})$ 
10:    end for
11:     $count = count + 1$ 
12:  end for
13:   $end_{B_k} = count - 1$ 
14:   $k = k + 1$ 
15: end for
```

ノード集合には、バリエーションパターンに存在する全てのタイムアイテムが属し、エッジ集合には、タイムブロック B_{k-1} に属する全てのタイムアイテムから B_k に属する全てのタイムアイテムへのエッジが属する。可視化の際には、Algorithm 2 で得られたノード集合とエッジ集合からグラフを作成する。

4. 実験

これまでに本研究で用いるタイムシーケンシャルパターンの共通部分検出、その出力形式について説明した。本章では、宮崎大学医学部附属病院から提供される電子カルテデータに対して、提案手法を適用してどのような結果が得られるのかを観察する。

4.1 実験目的

実際の症例に対する医療指示から抽出した頻出シーケンシャルパターンから、正しくクリニカルパスバリエーションが検出されているのかを確認する。またクリニカルパスバリエーションを可視化ツールにより正しく表現されているのかを確認する。

4.2 実験内容

事前実験により求めた最小支持度を用いて、薬剤名称を用いてのマイニングを行い、その結果に対してタイムシーケンシャルパターンの共通部分検出を行う。さらに、実験結果をバリエーションの表現を考慮した形式での可視化も行う。マイニングにおいては T-PrefixSpan を用いてタイム頻出シーケンシャルパターンを求め、その後に飽和タイム頻出シーケンシャルパターンのみを求めて、それを共通部分検出の入力とした。

実験では、マイニングによって得られたタイム頻出シーケンシャルパターンから、バリエーションパターンを構成出来ていることを確認する。実行環境を表 1 に示す。

表 1 実行環境

CPU	3 GHz Intel Core i7
Memory	8 GB
OS	OS X Yosemite ver.10.10.5 64bit
Java.ver	Java 1.8.0_45

4.3 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 1991 年 11 月 19 日から 2015 年 10 月 4 日までに記録された、実際に使わ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。この医療支持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[10](49GB)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[11] 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。

電子カルテシステムに記録された TUR_Bt という医療指示を含む医療指示データを対象のデータセットとして、3 章で説明したシーケンシャルパターンの共通部分検出を行う。TUR_Bt は術後の医療指示の流れがあまり定まっていないものであるため、クリニカルパスバリエーションの検出が行いやすいのではないかと予測しデータセットとして選んだ。

4.4 クリニカルパスバリエーションの検出

対象データセットの TUR_Bt に関する医療指示の詳細情報を表 2 に示す。

表 2 対象データセット

のべ対象患者数	381
患者一人に対する医療指示の最大数	1999
患者一人に対する医療指示の平均数	85.67

今回の実験では入院期間が 7 日間であるもののみを用いて実験を行い、抽出する頻出クリニカルパス長は 17 以上でかつ飽和なパターンのみとした薬名分類を行った。これにより得られた頻出クリニカルパスの情報を表 3 に示す。

この頻出クリニカルパスに対して、共通部分検出によるクリニカルパスバリエーションの作成を行う。

表 3 頻出クリニカルパス

最小支持度	0.1
入院期間	7 日間
クリニカルパス数	147
最大クリニカルパス長	21
平均クリニカルパス長	17.755

4.5 実験結果

実験の結果、前節で得られた 147 のクリニカルパスから 74 のバリエーションを考慮したグラフを検出することが出来た。1 グラフを構成するのに使用されたクリニカルパス数の情報を表 4 に示す。

表 4 1 バリエーションパターンを構成するクリニカルパス数

1 グラフを構成する クリニカルパス数	1	2	3	4	5	6	7	8	全グラフ
グラフ数	38	20	7	3	2	3	0	1	74

得られた頻出クリニカルパスとそれを元に作成されたバリエーションパターンの一部をそれぞれ図 4、図 5 に示す。バリエーションパターンは図 4 の頻出クリニカルパス $\langle a, b, c, d, e \rangle$ の共通部分検出によって生成されたものを出力している。

```
<頻出クリニカルパス a>
...[{"order_type": "手術", "order_explain": "膀胱悪性腫瘍手術", "order_name": "", "code": "", "day": "01"},
[{"order_type": "注射", "order_explain": "静脈内注射", "order_name": "ラクテック注", "code": "331", "day": "01"},
[{"order_type": "注射", "order_explain": "静脈内注射", "order_name": "セファゾリン N a 点滴静注用", "code": "613", "day": "01"},
[{"order_type": "病理診断", "order_explain": "T?M", "order_name": "", "code": "", "day": "01"},
[{"order_type": "緊急検査", "order_explain": "WBC", "order_name": "", "code": "", "day": "01"}] ...

<頻出クリニカルパス b>
...[{"order_type": "手術", "order_explain": "膀胱悪性腫瘍手術", "order_name": "", "code": "", "day": "01"},
[{"order_type": "注射", "order_explain": "点滴注射", "order_name": "ブイーンD注", "code": "331", "day": "01"},
[{"order_type": "注射", "order_explain": "静脈内注射", "order_name": "ラクテック注", "code": "331", "day": "01"},
[{"order_type": "注射", "order_explain": "静脈内注射", "order_name": "セファゾリン N a 点滴静注用", "code": "613", "day": "01"},
[{"order_type": "緊急検査", "order_explain": "WBC", "order_name": "", "code": "", "day": "01"}] ...

<頻出クリニカルパス c>
...[{"order_type": "手術", "order_explain": "膀胱悪性腫瘍手術", "order_name": "", "code": "", "day": "01"},
[{"order_type": "手術麻酔", "order_explain": "閉鎖循環式全身麻酔", "order_name": "", "code": "", "day": "01"},
[{"order_type": "注射", "order_explain": "静脈内注射", "order_name": "ラクテック注", "code": "331", "day": "01"},
[{"order_type": "注射", "order_explain": "静脈内注射", "order_name": "セファゾリン N a 点滴静注用", "code": "613", "day": "01"},
[{"order_type": "緊急検査", "order_explain": "WBC", "order_name": "", "code": "", "day": "01"}] ...

<頻出クリニカルパス d>
...[{"order_type": "手術", "order_explain": "膀胱悪性腫瘍手術", "order_name": "", "code": "", "day": "01"},
[{"order_type": "手術麻酔", "order_explain": "閉鎖循環式全身麻酔", "order_name": "", "code": "", "day": "01"},
[{"order_type": "注射", "order_explain": "静脈内注射", "order_name": "ラクテック注", "code": "331", "day": "01"},
[{"order_type": "緊急検査", "order_explain": "WBC", "order_name": "", "code": "", "day": "01"}] ...

<頻出クリニカルパス e>
...[{"order_type": "手術", "order_explain": "膀胱悪性腫瘍手術", "order_name": "", "code": "", "day": "01"},
[{"order_type": "注射", "order_explain": "静脈内注射", "order_name": "ラクテック注", "code": "331", "day": "01"},
[{"order_type": "注射", "order_explain": "静脈内注射", "order_name": "セファゾリン N a 点滴静注用", "code": "613", "day": "01"},
[{"order_type": "緊急検査", "order_explain": "WBC", "order_name": "", "code": "", "day": "01"}] ...
```

図 4 頻出クリニカルパス例 (一部抜粋)

```
...[{"order_type": "処方", "order_explain": "外用薬剤", "order_name": "グリセリン洗眼「オクタ」", "code": "235", "day": "01"},
[{"order_type": "手術", "order_explain": "膀胱悪性腫瘍手術", "order_name": "", "code": "", "day": "01"},
[{"order_type": "注射", "order_explain": "静脈内注射", "order_name": "ラクテック注", "code": "331", "day": "01"},
[{"order_type": "注射", "order_explain": "点滴注射", "order_name": "ブイーンD注", "code": "331", "day": "01"},
[{"order_type": "手術麻酔", "order_explain": "閉鎖循環式全身麻酔", "order_name": "", "code": "", "day": "01"},
[{"order_type": "注射", "order_explain": "静脈内注射", "order_name": "セファゾリン N a 点滴静注用", "code": "613", "day": "01"},
[{"order_type": "注射", "order_explain": "静脈内注射", "order_name": "ラクテック注", "code": "331", "day": "01"},
[{"order_type": "病理診断", "order_explain": "T?M", "order_name": "", "code": "", "day": "01"},
[{"order_type": "注射", "order_explain": "静脈内注射", "order_name": "セファゾリン N a 点滴静注用", "code": "613", "day": "01"},
[{"order_type": "緊急検査", "order_explain": "WBC", "order_name": "", "code": "", "day": "01"}] ...
```

図 5 バリエーションパターン例 (一部抜粋)

図 5 の色字部分がそれぞれのバリエーションとなっており、同一

のタイムアイテムブロックに属していて、医療大別が異なる、もしくは医療大別と薬効が同じであるが注射されている薬剤名が異なっていることが確認出来る。

4.6 可視化実験

図4の得られたクリニカルパスバリエーションの一部を可視化ツールによって表現したものが図6(全体図)、図7(拡大図)である。ツールの作成に関してはD3.js[12]を用いて、Webアプリケーションとして開発している。各医療指示をグラフのノードとし、シーケンスに従ってノードを矢印で表現することで可視化を実現している。ツールには図4で示した、バリエーションの文字列による記述方法の入力を与えている。

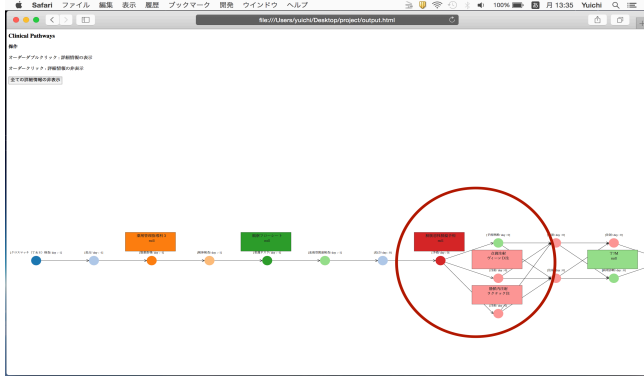


図6 可視化例 (全体概要)

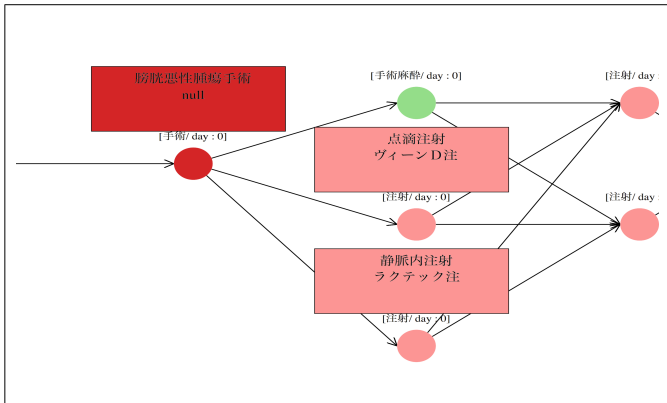


図7 可視化例 (拡大図)

医療指示の(大別 type)は全体の流れを把握する上で必要な情報となるため、大別毎にノードに色の設定を行い色分けをし、視認性を高めた。同様に、(処置日 day)も全体を把握するために必要な情報であるため、ノード上部に大別と共に常に表示するようにしている。さらに、各医療指示の細かな分析のために必要と思われる補足情報として、ノード上部の四角の枠内に(詳しい説明 Explain, 薬剤名 Name)が記載されている。この部分はクリック等のマウス操作により表示と非表示の切り替えが可能である。これにより、ノードと矢印を追っていくことでの全体の流れの把握、及び医療指示の補足情報を表示することでの医療指示単位の情報の把握が可能であり、従来の文字列による出力(図4)よりも理解がしやすく、扱いやすい形式となった。

5. まとめと今後の課題

5.1 ま と め

本研究では、医療指示履歴のデータ SDB から T-PrefixSpan によって得られた典型シーケンスに対して共通部分検出を行うことにより、医療指示の変形パターンであるバリエーションを、バリエーションパターンというグラフ表現をする手法を説明した。提案手法を用いた実験の結果、TUR_Bt 術における典型クリニカルパス同士の対応関係が把握出来るようになったことが確認できた。また、従来の文字列による出力よりも視認性を高めるため、可視化ツールの作成を行い、実験によって得られたデータの表示を行った。

5.2 今後の課題

本提案手法では、SPM で抽出されなかったバリエーションパターンも結果として出力としてしまう場合がある。そのため、SPM で抽出された頻出クリニカルパス上に存在しないようなエッジに対する扱いをどのようにするのか検討の必要がある。また、本手法はバリエーションパターンとなっている医療指示間の優先度(支持度)の情報を含んでいない。実際のクリニカルパスの分析のためには、重要な情報であるため、考慮するような方法の検討が必要である。

今回の実験で行った共通部分検出を行う頻出クリニカルパスの基準では、他の他のクリニカルパスとも共通部分検出を行わないものが多く得られてしまう結果となってしまうために、本研究に合うようなクリニカルパス同士の比較基準を考える必要がある。

今回は確認を行ったのは TUR_Bt のみのため、他のクリニカルパスに対しても同様の実験を行う必要がある。

そして、このクリニカルパスバリエーションの情報を含む出力がどの程度医学的に有益なのか医療従事者と確認し、提案手法の評価を行う必要がある。

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の助成により行われた。また、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指示支援に用いることは宮崎大学の HP[11]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] 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時系列分析による頻出シーケンスの抽出” DEIM Forum 2014, F6-2, 2014.
- [2] Rakesh Agrawal and Ramakrishnan 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.
- [3] 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.
- [4] X. Yan, J. Han and R.Afshar., “CloSpan: Mining closed sequential patterns in large databases.” Proc.SIAM Int '1

Conf. Data Mining (SDM '03), pp. 166-177, May 2003.

- [5] Yen-Liang Chen, Mei-Ching Chiang and Ming-Tat Ko., "Discovering time-interval sequential patterns in sequence databases." Expert Systems with Applications 25, pp. 343-354, 2003.
- [6] Jian Pei, Jiawei Han, Behzad Mortazavi-Asl, Helen Pinto, Qiming Chen, Umeshwar Dayal, Mei-Chun Hsu., "PrefixSpan: Mining Sequential Patterns Efficiently by Prefix-Projected Pattern Growth." Proceeding of 2001 International Conference on Data Engineering, pp. 215-224, 2001.
- [7] 浦垣啓志郎, 保坂智之, 荒堀喜貴, 串間宗夫, 山崎友義, 荒木賢 二, 横田治夫., "電子カルテの投薬履歴における薬効に着目した 医療行為パターンの抽出" DEIM Forum 2016, G7-5 2016.
- [8] 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.
- [9] Heikki Mannila, Hannu Toivonen, and A. Inkeri Verkamo., "Discovery of frequent episodes in event sequences" Data Mining and Knowledge Discovery, Vol. 1, pp. 259-289, 1997.
- [10] 電子カルテシステム WATATUMI., http://www.corecreate.com/0201_izanami.html
- [11] 宮崎大学医学部附属病院医療情報部., <http://www.med.miyazaki-u.ac.jp/home/jyoho/>
- [12] D3.js., <https://d3js.org>