

患者の固有情報及び動的状态を考慮したクリニカルパス分岐要因推定

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[†]

[†] 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

^{††} 東京工業大学 情報工学科 〒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 病院 IR 部 〒889-1692 宮崎県宮崎市清武町木原 5200

E-mail: †honda@de.cs.titech.ac.jp

あらまし 電子カルテデータの二次利用として、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。先行研究では、医療従事者への支援として、医療指示の履歴に時系列データマイニングを適用することで、医療指示の頻出クリニカルパス集合の抽出を行い、抽出した頻出クリニカルパス集合からパス中で部分的に異なる医療指示であるクリニカルパス分岐を検出した。本研究では、このクリニカルパス分岐の要因推定を行い、より各患者に合った医療の支援を試みるために、SPM で生成された頻出シーケンスの分岐要因推定をシーケンスの静的及び動的情報の両方を考慮して行う手法を提案する。実験では宮崎大学医学部附属病院で使用されている実電子カルテのデータを用いて、クリニカルパス分岐に対して静的情報として年齢や入院時期を、動的情報として体重、体温と血圧を用いてロジスティック回帰分析によるクリニカルパス分岐要因推定を行った。その結果、有意な差を持つクリニカルパス分岐とその要因の推定をすることが出来た。

キーワード 電子カルテ、多変量解析、クリニカルパス分岐

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、千年カルテプロジェクトが開始され複数の医療機関の電子カルテデータを一元的に管理されることが計画されている。そしてこのプロジェクトには、管理されたデータをプライバシーを考慮してから研究機関に渡し、研究などによる二次利用を行うことも含まれており、電子カルテデータの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、またパス中で部分的に異なる医療指示であるクリニカルパス分岐を検出し、分岐に影響した要因を推定することは、より患者にあった医療指示の実行の支援となることが見込まれる。

1.2 先行研究

Le ら [1] 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、アイテム間の時間間隔を考慮した、タイムインターバルシーケンシャルパターンマイニング (以下、TI-SPM) を適用した。この方法により、電子カルテから頻出なクリニカルパスの抽出を行った。

以前の研究 [2] では、TI-SPM によって得られた頻出クリニカルパス集合の共通部部分検出によって、パス同士の共通部分と非共通部分を区別することにより、パス中で部分的に異なる医療指示であるクリニカルパス分岐を検出する方法を提案し、実際の電子カルテに対して実験を行い、分岐を検出を行った。しかし、この研究の課題として分岐の発生要因 (例えば、患者の年齢や入院時期等) については言及をしていないことが挙げられる。分岐の発生要因を推定することによって、より患者にあった医療指示の実行の支援となる可能性がある。

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クリニカルパスから SPM によって抽出された頻出クリニカルパスに存在するクリニカルパス分岐に対して、分岐に影響した要因を推定することを目的とする。そのため、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取り入れた多変量解析を行い、分岐に影響した要因の推定手法を提案する。

1.4 本稿の構成

本稿は以下の通り構成される。2. 章では電子カルテの二次利用に関する研究について述べる。3. 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。4. 章では提案手法として, SPM で生成された頻出シーケンスの分岐要因推定をシーケンスの静的及び動的情報の両方を考慮して行う手法について述べる。5. 章では, 提案手法を電子カルテに適用する場合の各種データの扱いについて説明する。6. 章では提案手法を用いて, 実症例に対して実験を行い結果の考察を行う。最後に 7. 章でおわりにと題してまとめと今後の課題について述べる。

2 関連研究

電子カルテの二次利用に関する研究としては, 以下のようなものがある。若宮ら [3] は, 電子カルテの二次利用に必要な機能についてを提案した。また, 平野ら [4] は典型らしさ指標を用いて医療指示のクラスタリングとクリニカルパスの候補を生成する研究を行った。

また, 電子カルテからシーケンシャルパターンマイニング (以下, SPM) を用いて, 頻出クリニカルパスの抽出を行う研究も行われている。牧原ら [5] の研究では, 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, アプリオリアルゴリズム [6] を元にした SPM を適用し, 頻出なクリニカルパスの抽出を行った。

佐々木ら [7] は, 飽和オーダー列と呼ばれる概念 [8] 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 [9] (以下, TI-SPM) を PrefixSpan [10] 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い, 医療情報には重要な時間の概念を取り入れた。

浦垣ら [11] 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を抽出可能であると示した。さらに, 予め時間間隔集合を設定せず統計処理によって, 決定する T-PrefixSpan を提案した。保坂ら [12] 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。

Le ら [1] は, TI-SPM を他の SPM よりも計算時間の短い CSpan [14] 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。

3 背景知識

3.1 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 [6]。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。

3.2 TI-SPM

当初 Agrawal らが提案手法 [6] 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 [9]。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。

3.3 T-PrefixSpan

TI-SPM である I-PrefixSpan [9] 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 の問題点を解決すべく, T-PrefixSpan [11] 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が可能とした。

3.4 T-CSpan

T-PrefixSpan は PrefixSpan を元に行っているため, 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 [1] 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。

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

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 3. 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 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** 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 5. 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 6. 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 $Sup(A) \equiv |\{s \mid s \subseteq S, (sid, S) \in D, sid \text{ は } S \text{ の識別子}\}|$ と定義される。

T-CSpan は SDBD, 最低支持度 $MinSup$ を入力とする。初めに 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。

3.5 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、データベースのシーケンス ID(以下、SID とする)を保持しながら行う方法を提案した。

4 提案手法

本章では、SPM で生成された頻出シーケンスの分岐要因推定をシーケンスの静的及び動的情報の両方を考慮して行う手法について説明する。前提として、オリジナルシーケンス集合から SPM により頻出シーケンス集合が抽出されており、さらにその頻出シーケンス集合から分岐検出により分岐を考慮した頻出シーケンスが作成されているものとする。

その分岐を考慮した頻出シーケンスに対して適用する本提案手法は、大きく 3 つのステップに分けられる。

1. 分岐に対応するオリジナルシーケンス集合の特定
 2. 多変量解析に必要な静的及び動的情報の抽出
 3. 多変量解析による有意差を持つ分岐の特定
- 以下、ステップ毎に説明を行う。

4.1 分岐に対応するオリジナルシーケンス集合の特定

分岐に対応するオリジナルシーケンス集合の特定は、さらに二つの手続きに分けられる。

- (1) 分岐に対応する頻出シーケンスを特定
- (2) その頻出シーケンスに対応するオリジナルシーケンス集合の特定

である。これらの手続きは共に対応する識別子の取得によって行われる。

(1) は、分岐検出時における分岐後の要素に対応する頻出シーケンスの識別子の取得する

(2) は、SID を保持する SPM [13] によって付与された頻出

		<シーケンス情報区分>					
シー ケ ン ス 情 報 V	静的情報		IC1	IC2	IC3	IC4	...
		attr1	V1	V1	V1	V1	...
	動的情報	attr2	V2	V2	V2	V2	...
		attr3	V3a	V3a	V3b	V3c	...
		attr4	V4a	V4b	V4b	V4b	...
		...	...	...	...	...	...

図 1 シーケンス情報テーブル

シーケンスに対応するオリジナルシーケンス集合の識別子を取得する

これにより、頻出シーケンスの分岐に対応するオリジナルシーケンス集合を特定することが出来る。

4.2 多変量解析に必要な静的及び動的情報の抽出

頻出シーケンスの分岐要因と考えられるものとして、シーケンスの静的情報と動的情報が挙げられる。以下に、その定義を示す。

定義 7. シーケンス静的情報及び動的情報

シーケンスの静的情報とは、シーケンスのどの要素を参照しても同じ値が得られる情報を指す。

シーケンスの動的情報とは、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情報を指す。

また本研究において単にシーケンス情報といった場合には、シーケンスの静的情報と動的情報の両方を指すものとする。

静的・動的情報の両方が分岐要因となる可能性があるため、要因推定を行う際には静的・動的情報の両方を多変量解析の説明変数として利用する必要がある。

本節では、シーケンス情報の表現方法の導入を行い、抽出方法についてを説明をする。本研究において、シーケンス情報はシーケンスの各属性を行として、シーケンス情報区分を列としたシーケンス情報テーブルによって表現する(図1)。

シーケンス情報区分とは、その情報区分に属する全ての要素の全シーケンス情報の値が等しくなるように、シーケンスの要素を分けた時の区分である。

この時、静的情報の値はシーケンス情報区分によらず一定のため行内で同じ値を示す(図1中 attr1,2)。一方動的情報の値は情報区分によって変わるため、行内で値が変化する(図1中 attr3,4)。

多変量解析の説明変数としてシーケンス情報(特に、動的情報)を利用するには、分岐時のオリジナルシーケンスの情報区分を求めることが必要である。しかし、オリジナルシーケンスの分岐箇所を求めて情報区分を得ることは難しいため、以下のような方法をとる。

まず、全シーケンスに共通な基準要素を設定する。次に、その基準要素の情報区分と頻出シーケンス上の分岐位置の情報区分から距離を求め、最後にそれを用いてオリジナルシーケンス上の基準要素の情報区分と求めた距離からオリジナルシーケンス上の分岐時のシーケンス中の情報区分を類推する。

SID	情報区分	対応頻出シーケンス	attr1	attr2	attr3	attr4	...
SID(n)	IC1	FS1	V1	V2	V3a	V4a	...
SID(n)	IC2	FS1	V1	V2	V3a	V4b	...
SID(n)	IC3	FS1	V1	V2	V3b	V4c	...
...	...	...	...	...	...	...	...
SID(m)	IC1	FS2	V'1	V'2	V'3a	V'4a	...
SID(m)	IC2	FS2	V'1	V'2	V'3b	V'4b	...
SID(m)	IC3	FS2	V'1	V'2	V'3b	V'4c	...
...	...	...	...	...	...	...	...
...	...	...	...	...	...	...	...

図 2 解析情報テーブル

これにより、分岐時のオリジナルシーケンスの情報区分を求める。

4.3 多変量解析による有意差を持つ分岐の特定

前節で得られたシーケンス情報を用いて多変量解析を行うことで有意差を持つ分岐の特定が可能となる。まず、全オリジナルシーケンスのシーケンス情報テーブル、オリジナルシーケンスとその対応頻出シーケンスの情報から解析情報テーブルを作成する(図2)。その解析情報テーブルから、分岐毎に目的変数と説明変数セットを選ぶことによって多変量解析を行う。

多変量解析では説明変数としてオリジナルシーケンスに対応する頻出シーケンスを用いる。これは、直感的には分岐の方向を示す指標となっている。また、説明変数セットは、4.2節で得られた分岐時の情報区分を用いて取得するシーケンス情報となる。この多変量解析をシーケンス中の各分岐に対して行うことで、有意差を持つ分岐とその要因を特定する。図2では、赤く囲まれた部分が目的変数、緑色で囲まれた部分が説明変数セットとして選ばれている状態を示す。

5 電子カルテへの適用

本章では、前章で説明した提案手法を電子カルテデータに対して適用する際のデータの扱いを説明する。

5.1 医療指示の取り扱い

本研究では[11]と同様に、各医療指示を(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。

この各医療指示がシーケンスにおける要素に、クリニカルパスがシーケンスにあたる。

5.2 医療指示実施日の取り扱い

頻出クリニカルパス集合における医療指示実施日は、パスの

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
R.ver	R 3.3.1

表 2 TUR_Bt 詳細情報

患者数	342
平均医療指示数	48.8
平均在院日数	7 日
最小支持度	0.2

主医療指示の実施日を 0 日とした, 主医療指示との相対日時で表現する. 主医療指示ではないある医療指示において, 実施日が「-1 日目」である場合は, 主医療指示の前日に行う医療指示であることを指し, 同様に実施日が「2 日目」である場合は, 主医療指示の二日後に行う医療指示であることを指す. 主医療指示が基準要素に, 医療指示実施日がシーケンス情報区分となる.

5.3 患者情報の種別

本研究で用いるシーケンス情報は患者情報となる. また患者情報は, (1) 患者固有情報と (2) 患者状況情報の二つに分けられる.

(1) 患者固有情報とは, クリニカルパス中で一般に固定値である情報を指す. 患者の年齢などのクリニカルパス内で値の変化しない情報が該当する.

(2) 患者状況情報とは, クリニカルパス中で変化する情報を指す. 患者の体温などの日々値が更新される情報が該当する.

本研究においては, 単に患者情報といった場合は, 患者固有情報と患者状況情報の両方を指す.

また, 電子カルテの規格である「様式 1」と呼ばれるものに記載される情報や入院中に行われた検査項目とその結果などが患者情報の対象となる. また, 長い周期で定期的に変化する属性は, クリニカルパス中では一定であると考えられるため, 患者固有情報に含めることが出来る. 今回の例では年齢がこれにあたる.

6 実験

これまでに本研究で用いる, クリニカルパス分岐の要因推定に関する導入を行った. 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実験と考察を行う.

6.1 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集合から, クリニカルパス分岐を検出し, 提案手法によって得られた動的状況を考慮したクリニカルパス分岐要因推定について考察を行う.

表 3 TUR_Bt の最長の飽和タイム頻出シーケンシャルパターン情報

患者数	342
平均医療指示数	48.8
平均在院日数	7 日
最小支持度	0.2

表 4 利用する患者情報一覧

患者固有情報	年齢, 入院時期
患者状況情報	体重, 体温, 収縮期血圧

6.2 実験内容

電子カルテデータに対して TI-SPM と分岐検出を行い, 分岐を考慮した頻出シーケンスを抽出する. TI-SPM においては T-CSpan を用いて飽和タイム頻出シーケンシャルパターンを求める.

そして得られる分岐を考慮した頻出シーケンスに提案手法を適用し得られた要因推定の結果に対して考察を行う.

実験環境を表 1 に示す.

6.3 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 [15](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 [16] 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, 4 章の提案手法を適用する. 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, 今回の要因推定の対象として適切であると考えデータセットとして選んだ. TUR_Bt に関する詳細情報は表 2 に示す.

また, この TUR_Bt データセットに対して頻出シーケンスを抽出する T-CSpan の最小支持度と抽出された飽和タイム頻出シーケンシャルパターンのうち, 今回の実験で扱う最長の飽和タイム頻出シーケンシャルパターンの情報を表 3 に示す. 最長のパターンのみを扱うのは, パターンとして情報量の多いものを利用するためである.

また多変量解析の手段としては, 医学情報の解析手法として一般的なロジスティック回帰分析を用いた. また有意水準は 0.05 とし, この有意水準を下回った分岐を有意差を持つ分岐とした. 多変量解析の説明変数として扱う患者情報については, 患者固有情報として患者年齢と入院時期の二種類, 患者状況情報として体重, 体温, 収縮期血圧の三種類の計 5 種類 (表 4) を用いた.

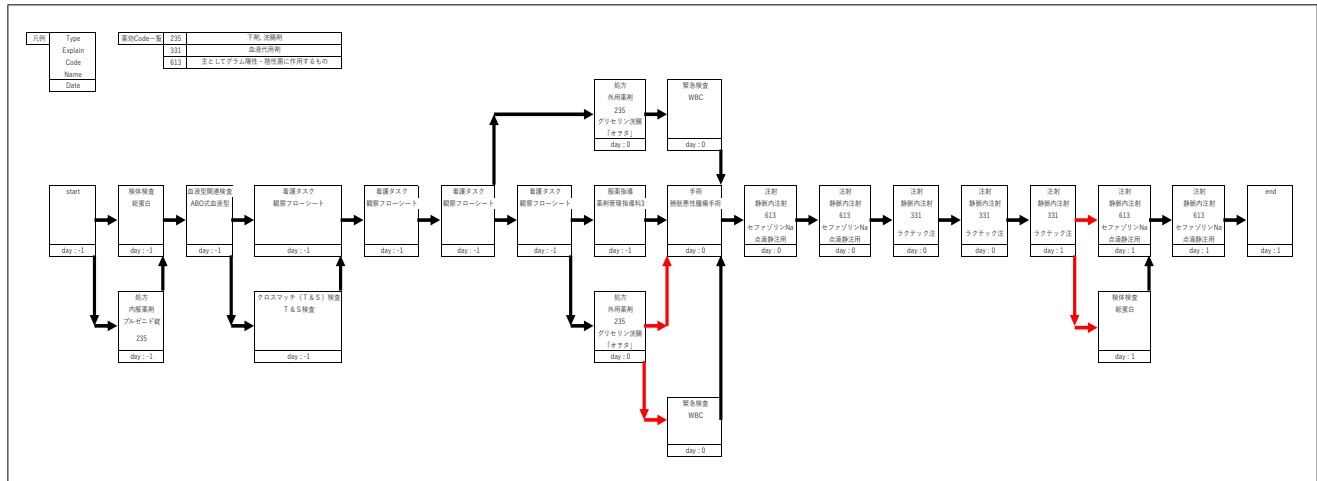


図 3 TUR_Bt 全体概要図

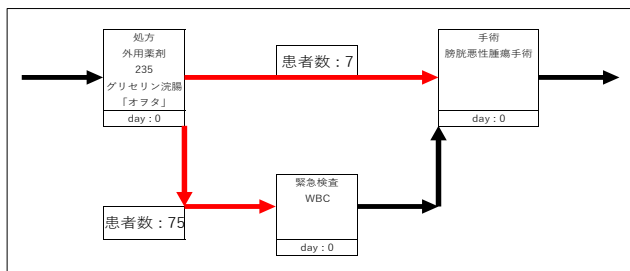


図 4 クリニカルパス分岐 (A)

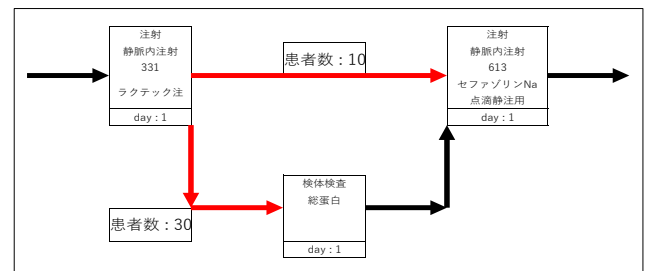


図 5 クリニカルパス分岐 (B)

表 5 クリニカルパス分岐 (A) データ分布

要因種別	入院時期	
患者数	7	75
95%信頼区間	2012.8 ~ 2013.5	2013.5 ~ 2013.9
データ平均	2013.1	2013.7

表 6 クリニカルパス分岐 (B) データ分布

要因種別	患者年齢	
患者数	10	30
95%信頼区間	56.6 ~ 74.0	54.5 ~ 61.3
データ平均	65.3	57.9

6.4 実験結果

事前実験で得たデータセットに対して、分岐グラフを作成し提案手法によってクリニカルパス分岐の要因推定を行った結果を図 3 に示す。各四角ノードが医療指示を示し、矢印で医療指示の流れを示している。図中の赤い矢印が推定の結果、分岐要因が存在すると推定された部分になる。今回の実験では、六ヶ所存在するクリニカルパス分岐の中で、二カ所のクリニカルパス分岐において有意差が存在すると推定された。

図 45 (クリニカルパス分岐 (A) 及び (B)) は、実験によって有意差があると推定されたクリニカルパス分岐部分 (図 3 の赤矢印が含まれる部分) を拡大し、推定要因とそれぞれの患者数を付加したものである。また、表 5,6 は、それぞれにおける要因となった患者情報の”患者情報種別”, ”患者数”, ”95%信頼区間”と”平均値”をまとめたものである。

クリニカルパス分岐 (A) では、主として手術前に「緊急検査」を行う場合が多いが一部行わない場合が存在する。同様に、クリニカルパス分岐 (B) では、「セファゾリン Na の点滴注射」を行う前に「検体検査」を行っているが、一部行わない場合が存在している。

有意差が生じたクリニカルパス分岐 (A) 及び (B) は、ある

医療指示を行うか行わないかの選択肢がある部分であるという点において共通点がある。実医療においては、基礎となる典型的なクリニカルパスは安全やより多くの患者に当てはまるように考慮して、ある程度多くの医療指示を入れておくことが自然である。実際に行われる医療指示は、そこから医療従事者の経験をもとに、対象の患者に不必要であると考えられる医療指示を削除したものとなる。この点に関して、提案手法によってクリニカルパス分岐 (A) 得られた及び (B) のような部分に有意差が表れる結果になったことと実医療に差異はないと考えられる。

表 5 及び 6 を観察すると、クリニカルパス分岐 (A) では、入院時期が新しいほど「緊急検査」を行う傾向にあり、クリニカルパス分岐 (B) では、患者の年齢が高いと「検体検査」を行わない傾向にあることが分かる。

また、一部のクリニカルパス分岐では有意差が生じなかった。これは、今回の実験で用いた TUR_Bt のデータセットにおける患者数 (データ数) が少なくデータの分布に特徴が出るまで行かなかったこと、また今回扱った患者情報以外のものが分岐の要因となっていることが考えられる。患者数を増やして実験を行う、または扱う患者情報の種類を増やすした実験を行うこと

で新たに有意差が生じる場合があると考えられる。

7 おわりに

7.1 ま と め

本研究では、SPM と分岐検出によって生成された分岐を考慮した頻出シーケンスに分岐要因を推定する方法を提案した。本提案は、頻出シーケンスに含まれる静的情報だけでなく、動的情報も要因推定の対象とすることが可能となっている。

また、提案手法を実際に使用されている電子カルテに記載された TUR_Bt という症例に対するデータセットについて適用した。実験の結果、有意差のあるクリニカルパス分岐とその要因を得ることが出来た。

7.2 今後の課題

今回の実験で説明変数として利用した患者情報は限定的なものであるため、より多くの患者情報を用いた時にどのような結果が得られるかを観察する必要がある。同様に、他症例においても実験を行い考察をする必要がある。

そして、今回の結果がどの程度医学的に有益なのか医療従事者と確認する必要がある。

また、本提案はクリニカルパス分岐の要因についてのみに言及しているが[13]では、分岐の安全性についても言及している。クリニカルパス分岐の要因と安全性の両方を考慮することで、より医療従事者へ有益な情報を提供することができる可能性がある。

謝 辞

本研究の一部は、厚生労働行政推進調査補助金政策科学推進研究事業「ナショナルデータベース (NDB) データ分析における病名ロジック作成のための研究」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 [16] 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] Hieu Hanh Le, Henrik Edman, Yuichi Honda, Muneeo 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.
- [2] Y. Honda, M. Kushima, T. Yamazaki, K. Araki, H. Yokota. “Detection and Visualization of Variants in Typical Medical Treatment Sequences” the 3rd VLDB Workshop on Data Management and Analytics for Medicine and Healthcare (DMAH). Springer, pp. 88-101,2017.
- [3] Shunji Wakamiya and Kazunobu Yamauchi. What are the standard functions of electronic clinical pathways? International Journal of Medical Informatics 78, pp. 543-550, 2009.
- [4] Shoji Hirano and Shusaku Tsumoto. Clustering of order

- sequences based on the typicalness index for finding clinical pathway candidates. IEEE International Conference on Data Mining (ICDM Workshops), 2013.
- [5] 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.
- [6] 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.
- [7] 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.
- [8] X. Yan, J. Han and R. Afshar., “CloSpan: Mining closed sequential patterns in large databases.” Proc.SIAM Int’l Conf. Data Mining (SDM ’03), pp. 166-177, May 2003.
- [9] 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.
- [10] 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.
- [11] Keishiro Uragaki, Tomoyuki Hosaka, Yoshitaka Arahori, Muneeo Kushima, Tomoyoshi Yamazaki, Kenji Araki and Haruo 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.
- [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エーション分析” DEIM Forum 2019, 2019.
- [14] 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.
- [15] 電子カルテシステム WATATUMI., http://www.corecreate.com/0201_izanami.html
- [16] 宮崎大学医学部附属病院 IR 部, <http://www.med.miyazaki-u.ac.jp/home/jyoho/>