T2R2 東京科学大学 リサーチリポジトリ
Science Tokyo Research Repository

論文 / 著書情報
Article / Book Information

論題(和文)	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
Title(English)	
著者(和文)	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
Authors(English)	Keishirou Uragaki, Tomoyuki Hosaka, Yoshitaka Arahori, Muneo Kushima, Tomoyoshi Yamazaki, Kenji Araki, Haruo Yokota
出典(和文)	第8回データ工学と情報マネジメントに関するフォーラム論文集, , , G7-5
Citation(English)	, , , G7-5
発行日 / Pub. date	2016, 3

Powered by T2R2 (Science Tokyo Research Repository)

DEIM Forum 2016 G7-5

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

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††

† 東京工業大学 工学部情報工学科 〒 152-8565 東京都目黒区大岡山 2-12-1
†† 東京工業大学 大学院情報理工学研究科 計算工学専攻 〒 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
E-mail: †uragaki.k.aa@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, 医療行為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, 医療行為の典型的な流れである「クリニカルパス」抽出に向けた支援を試みてきた. 本研究では, 先行研究で扱って来なかった薬剤の情報を解析に取組むことを試みる. 医療現場で実際に投与される薬剤の種類は多く, 単純に薬剤名を含めてマイニングを行うことはパターンを抽出する上で得策ではない. 我々は投与された薬剤の薬効に着目し, 医学的に有益なパターンを得るために利用する. 薬効を用いるか否かで出力がどのように変化するのかを比較することで手法の評価を行い, 医師が経験をもとに作成したクリニカルパスとどの程度一致しているのか確認する.

キーワード データマイニング, シーケンシャルパターンマイニング, 電子カルテ, 薬剤情報, 薬効

1. はじめに

1.1 研究背景

大規模病院において広く普及している電子カルテは, 従来の紙のカルテ比べて, 高速に検索・閲覧を可能とし, 医療行為の標準化に貢献している. 近年, 電子カルテはカルテとしての利用のみに留まらず, 二次利用が期待されている.

二次利用の例として, 特定の病気の患者に対しての典型的な医療行為の流れ「クリニカルパス」を抽出することが挙げられる. 従来, クリニカルパスは医療関係者自身の医学的経験に基づいて作成されていたが, 人の手による作成は容易ではなかった. そのような背景のもとで, 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・抽出し医療行為改善の支援を目的とした研究が現れ始めた. 電子カルテデータを分析する研究によって, 医療行為履歴からクリニカルパスが適切であると判断することは有用であり, 新たな医療行為の分岐「バリアント」の発見によりさらなる医療行為の改善が見込まれる.

1.2 先行研究

牧原らの研究 [1] では, 電子カルテのアクセスログから, ある患者に対して行った医療行為をアイテム, 医療行為の流れをシーケンス, すべての患者の医療行為の流れをデータベースとみることで, アプリオリアルゴリズム [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) により, 頻出シーケンシャルパターンの抽出を行った. 牧原らは, 手術といった特定の重要な医療行為を「基準イベント」と定め, 基準イベントの前後の部分シーケンスで独立にマイニングを行った. この手法によって, 医療現場の都合でシーケンスの順序が変化したとしても, 基準イベントの前後で行うべき医療行為を抽出することができた.

しかし, 牧原らの手法には, 基準イベントの後の部分でマイニングを行った場合のパターン数が膨大となってしまうことと医療行為間の時間間隔を考慮していないことの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を解決した. パターン数の削減を行うため, 飽和オーダ列と呼ばれる概念 [4] を導入した. 2 つのパターン A,B を比較した時に, A が B を含み, B のサポート値が A のサポート値以下であれば, B は飽和ではないという飽和の性質に基づき, すべての 2 パターンに対して比較を行っていき, 飽和でないパターンを出力から削除するという手法をとった. この結果, 出力の情報量を損なわずに出力数を減らすことができた. 医療行為間の時間間隔については, Chen らが提案した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, 2 アイテム間の時間間隔を考慮した抽出を行った. この手法により医療行為間の大まかな時間間隔を得ることができ, 従来より情報量の多いパターンを抽出することができた.

しかし, TI-SPM は人為的に定めたタイムインターバル (以下, TI) という特別な時間間隔内に注目する 2 アイテム間の時間間隔が収まっているのかを確認するというアルゴリズムであるために, 定めた時間間隔によって結果が変わるという問題点があり, その結果, 最適な時間間隔を定める必要性があった.

また, 前述の二つの先行研究におけるアイテムは薬剤情報を含んでいないものもしくは薬名の一部のみを含んでいるものだけであった. このため, 先行研究において, 注射といった薬剤情報を含んだ医療行為において, どの薬剤を投与するのかがわからないという問題点があった.

1.3 本研究の目的

本研究は, 電子カルテシステムに記録されたある症例に対する医療行為から頻出シーケンシャルパターンを抽出し, 飽和と

いう性質を用いることで出力パターンの絞込を行い, クリニカルパスの作成補助を目的とする. 宮崎大学医学部附属病院における電子カルテシステムに記録された医療行為データを用いることで, 従来研究では扱っていなかった薬剤情報を含んだ医学的に有益な頻出シーケンシャルパターンの抽出を行う. 単純に薬剤名のみを用いてのマイニングは, 実際に患者に対して投与される薬剤の種類はマイニングを行う上で多く, アイテムの種類が増大してしまうため, 注射や処方といった薬剤情報を含んだ医療行為がパターンに現れにくい. その為, 本研究では投与された薬剤の薬効に注目し, 医学的に有益なパターンの抽出を行う. 前節で問題点とした, 2 アイテム間の時間間隔については外れ値処理を含んだ統計情報を提示する手法で解決を目指す.

実験では, まず患者に対する医療行為データを用いて頻出シーケンシャルパターンの抽出を行い, 出力パターン数, 平均パターン長, 薬剤が絡んだ医療行為を含んだパターンの割合の 3 つの指標に対して, 薬効を用いるか否かでどのような変化が現れるのか観察する. さらにその後, 抽出により得られた典型的な流れと医師が経験をもとに作成したクリニカルパスとがどの程度一致しているのか確認する.

1.4 本稿の構成

本稿は以下の通り構成される. 2. 章では本研究の関連概念を背景知識と題して説明する. 3. 章で薬剤情報の取り扱い方法及び薬剤投与の正確な時間間隔を求める手法の導入を提案手法として述べる. 4. 章では, 3. 章の手法を用いて, ある症例における医療行為データを解析し, 薬効を用いた場合と用いない場合の抽出でどの程度の差が出力に現れるのか比較実験を行う. さらにその後, 抽出により得られた典型的な流れと医師が経験を元に作成したクリニカルパスとがどの程度一致しているのか確認する. 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といい, 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定める識別子を組とした要素からなる. 頻出シーケンシャルパターンの定義を行う前にサブシーケンス, シーケンスとシーケンス集合の包含関係の定義を行う. さらに, 頻出シーケンシャルパターンの包含関係と密接な関係にある飽和頻出シーケンシャルパターン [4] と呼ばれる概念について説明する. 飽和でない頻出シーケンシャルパターンを削除することによって, 冗長なパターンを含まない出力を得ることができる.

定義 1. サブシーケンス

2 つのシーケンス A $= < a_1, a_2, ..., a_n >$ (ただし, a_i はアイテム. $i = 1, 2, ..., n$) , B $= < b_1, b_2, ..., b_m >$ (ただし, b_i はアイテム. $i = 1, 2, ..., m$) に対して, 以下が成り立つとき, **A を B のサブシーケンス**といい, **A⊆B** と表す.

（1） $a_1 = b_{j_1}, a_2 = b_{j_2}, ..., a_n = b_{j_t}$

（2） $n \leqq t \leqq m$

（3） $1 \leqq j_1 < j_2 < ... < j_t \leqq m$

定義 2. シーケンスとシーケンス集合の包含関係

シーケンス $A = < a_1, a_2, ..., a_n >$ (ただし, a_i はアイテム. $i = 1, 2, ..., n$) に対して $A \subseteqq B$ となるシーケンス B がシーケンス集合 $\sum$ 中に存在するとき, A **は** $\sum$ **に含まれている**といい, $A \subseteqq \sum$ と表す.

定義 3. 頻出シーケンシャルパターン

最小支持度 $MinSup (0 \leqq MinSup \leqq 1)$, SDB である D が与えられ, シーケンス $A = < a_1, a_2, ..., a_n >$ (ただし, a_i はアイテム. $i = 1, 2, ..., n$) において, $| \{Seq | A \subseteqq Seq, (sid, Seq) \in D, sid$ は Seq の識別子 $\} | \geqq Size(D) \times MinSup$ が成り立つとき, シーケンス A を D の最小支持度 $MinSup$ における**頻出シーケンシャルパターン**という.

ただし, $Sup(a_i)$ はアイテム a_i の D におけるサポート値, $Size(D)$ は D 中に存在するシーケンス数とする.

定義 4. 飽和頻出シーケンシャルパターン

SDB である D から抽出した頻出シーケンシャルパターン集合 $\sum$ に属する A に対して, 以下の条件を満たす $B \in \sum \backslash A$ が存在しないとき, A を**飽和頻出シーケンシャルパターン**であるという.

（1） $A \subseteqq B$

（2） $Sup(A) = Sup(B)$

ここで頻出シーケンシャルパターンのサポート値 $Sup(A)$ を $Sup(A) \equiv | \{s \mid s \subseteqq Seq, Seq \in D\} |$ と定義する.

2.2 TI-SPM

当初 Agrawal らが提案した手法 [2] は, 2 アイテム間の時間間隔を考慮してい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, 2015 年 1 月 11 日に検査を行い, 同日に手術を行うシーケンスと, 2015 年 1 月 11 日に検査を行い, 1 年後に手術を行うシーケンスを同じ情報を持ったシーケンスと見なしていた. こうした背景から Chen らは 2 アイテム間の時間間隔を考慮した TI-SPM と呼ばれる手法を提案した [5]. この手法によって, 例として挙げた二つのシーケンスを異なるシーケンスと区別することができるようになった.

TI-SPM は, 時間情報を含んだ SDB である D, 最小支持度 $MinSup (0 \leqq MinSup \leqq 1)$, TI-セットを入力として与えることによって 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. 以下のように, TI, TI-セット, TI-シーケンス, TI-サブシーケンス, TI-頻出シーケンシャルパターンは定義される.

定義 5. TI

$r-1$ 個の定数 $T_1, T_2, ..., T_{r-1}$ を元に, **TI** $I_k (k = 0, 1, ..., r-1, r)$ は以下によって定義される.

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

定義 6. TI-セット
$r-1$ 個の定数 $T_1, T_2, ..., T_{r-1}$ によって構成される $r+1$ 個の TI の集合を **TI-セット** V と定義する.

TI-SPM においては, シーケンスの要素をアイテムとそのアイテムの発生した時刻との組で表し, 同じ時刻に発生したアイテムは辞書順に並ぶものとした.

定義 7. 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.

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

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

定義 8. TI-サブシーケンス
シーケンス $A = < (a_1, t_1), (a_2, t_2), ..., (a_n, t_n) >$ と TI-シーケンス $B = < b_1, \&_1, b_2, \&_2, ..., b_{m-1}, \&_m, b_m >$ について, 以下の条件を満たす $1 \leqq j_1 < j_2 < ... < j_m \leqq n$ となるような整数列 $\{j_m\}$ が存在するとき, B **は** A **の TI-サブシーケンスである**といい, $B \subseteqq A$ と表す.

（1） $b_1 = a_{j_1}, b_2 = a_{j_2}, ..., b_m = a_{j_m}$
（2） $t_{j_i} - t_{j_{i-1}} \in \&_{i-1} \quad (i = 2, 3, ..., s)$

定義 9. TI-頻出シーケンシャルパターン
SDB D, 最小支持度 $MinSup$ $(0 \leqq MinSup \leqq 1)$ が与えられたとき, TI-シーケンス α が $|\{(sid, s) \mid (sid, s) \in D, \alpha \subseteqq s\}| \geqq Size(D) \times MinSup$ を満たすとき, α を **TI-頻出シーケンシャルパターン**と定義する.

これらの概念を定義し, Chen らは PrefixSpan [6] を, TI を考慮するように拡張した, I-PrefixSpan というアルゴリズムを提案した [5]. 以下では, I-PrefixSpan における概念とともに, I-PrefixSPan のアルゴリズムを述べる.

定義 10. TI-プレフィックス
シーケンス $A = < (a_1, t_1), (a_2, t_2), ..., (a_n, t_n) >$, TI-シーケンス $B = < b_1, \&_1, b_2, ..., b_{m-1}, \&_{m-1}, b_m >$ が次の条件を満たすとき, **B を A の TI-プレフィックス**と定義する.

（1） $m \leqq n$
（2） $a_i = b_i \ \ (1 \leqq i \leqq m)$
（3） $t_i - t_{i-1} \in \&_{i-1} \ \ (1 < i \leqq m-1)$

定義 11. 射影シーケンス
シーケンス $A = < (a_1, t_1), (a_2, t_2), ..., (a_n, t_n) >$, A の TI-サブシーケンスであるような TI-シーケンス $B = < b_1, \&_1, b_2, ..., b_{m-1}, \&_{m-1}, b_m >$ が $m \leqq n$ かつ $a_{i_k} = b_k \ \ (1 \leqq k \leqq m)$ を満たすとき, A のサブシーケンス $A' = < (a'_1, t'_1), (a'_2, t'_2), ..., (a'_{n'}, t'_{n'}) >$ が次の条件を満たすとき, A' **は** A **の** B **に関する射影シーケンスである**と定義する.

（1） $n' = n + m - i_m$ を満たす $i_m \ \ (0 \leqq i_m \leqq n)$ が存在する.
（2） B は A' の TI-サブシーケンスである.
（3） A' の後方 $n - i_m$ 個のアイテムと A の後方 $n - i_m$ 個のアイテムが一致する.

定義 12. TI-ポストフィックス
シーケンス $A = < (a_1, t_1), (a_2, t_2), ..., (a_n, t_n) >$ の TI-シーケンス $B = < b_1, \&_1, b_2, ..., b_{m-1}, \&_{m-1}, b_m >$ に関する射影シーケンスを $A' = < (a'_1, t'_1), (a'_2, t'_2), ..., (a'_{n'}, t'_{n'}) >$ とする. このとき, $A'_B = < (a'_{m+1}, t'_{m+1}), (a'_{m+2}, t'_{m+2}), ..., (a'_{n'}, t'_{n'}) >$ を B **に関する** A **の TI-ポストフィックス**と定義する.

定義 13. 射影 SDB
SDB D が与えられた時, D 中のすべてのシーケンスの α に関する TI-ポストプレフィックスの集合を**射影 SDB** $D\,|_\alpha$ と定義する.

I-PrefixSpan は SDB D, 最低支持度 $MinSup$ と TI-セット I を入力とする. まず初めに D 中の頻出シーケンシャルパターンを求め, それらを元に射影 SDB を構成する. その後, α で射影することによって構成された射影 SDB において, すべてのアイテム β に対し, I 中の TI 毎のサポート値を求め, 特定の TI & における値が $Size(D) \times MinSup$ 以上であれば, $<\alpha$ & $\beta>$ を TI-シーケンシャルパターンとして出力する. その後, $<\alpha$ & $\beta>$ で射影を行うという操作を繰り返し, TI-シーケンシャルパターンを求めるというアルゴリズムである.

3. 提 案 手 法

本章では従来研究にはなかった薬剤情報を取り入れる手法と医療行為間の時間間隔を統計的に導出する手法について説明する.

3.1 医療行為の取り扱い

本研究ではマイニングにおけるアイテムを (大別 Type, 詳しい説明 Explain, 薬効コード Code, 薬剤名 Name) の 4 つ組によって構成する. 薬剤情報に絡まない医療行為の場合は Code 及び Name は「null」と記述する. 薬効コードによって薬効が一意に定まる. 例えば, 「処方」である「内服薬剤」において, 薬効コードが「613」の薬剤「セフゾン細粒小児用１０％」を投与した時, (処方, 内服薬剤, 613, セフゾン細粒小児用１０%) と表され, 薬剤情報に絡まない医療行為である「看護タスク」の「シーツ交換」は (看護タスク, シーツ交換,null,null) と表される. ここで, 薬効コードが「613」の場合, 薬効は「主としてグラム陽性・陰性菌に作用するもの」となる.

Type 中の文字列に「処方」もしくは「注射」を含まない医療行為で扱っている薬剤は医学的に有用でないとして, Code 及び Name を「null」とする. さらに, Type の内容が「検体検査」である場合は Code 及び Name に加え Explain も有用ではないとして, 「null」とする.

Explain に関して, 電子カルテシステムに記録された医療行為の説明文をそのままマイニングに用いるとアイテムの種類数が増大してしまうことから, 佐々木ら [3] の研究で用いられ「短縮オーダ」と呼ばれる概念を導入することで, 説明文の短縮化を行った. 短縮オーダは電子カルテシステムに記録された説明

文の前半部分にその医療行為を特徴づける記述が行われていることに着目し, スペースや'・', '。' といった不定の区切り文字の前半部分のみを抽出する手法である. 例えば, 本来電子カルテシステム中に「皮膚レーザー照射療法（色素レーザー照射療法)」および「皮膚レーザー照射療法（色素レーザー照射療法)，皮膚レーザー照射療法（色素レーザー照射療法)，フィシザ」と記録された説明文をどちらも同じ「皮膚レーザー照射療法」とみなすことができ, マイニングの効率化を図ることができる.

最後に, あるシーケンスに属するすべてのアイテムの Type を見た時, クリニカルパスにおいて重要な医療行為である「手術」と一致する Type が存在しなければ, 医学的に有益なシーケンスでないとして削除してよい.

Code 及び Name に関しては後述する.

3.2 薬　　効

医療現場では異なる患者に同一の薬効, 異なる薬剤名称の薬剤を投与することが多く, Type, Explain, Code, Name の 4 つの属性すべてを用いてアイテムの一致判定を行う「薬名分類」では抽出できないパターンが存在する.

ここで, 薬剤情報を用いてマイニングを行う関連研究として, Wright らの研究が存在する [8]. この研究では, 薬剤分類に着目し, 薬剤分類が同一であれば同一アイテムとみなす手法により, 糖尿病患者に投与した薬剤をアイテムとするシーケンスによって構成される SDB から頻出シーケンシャルパターンの抽出を行った. その結果, 薬剤分類に着目した方法の方が, 薬剤名が同一であれば同一アイテムとみなす素朴な方法よりも, 高い確率で次に投与するアイテムを予測できた. 本研究ではこの結果に基づき, 薬剤の薬効に着目し, 薬剤が絡んだアイテムについて, Name が異なっていても Type,Explain,Code が同一であれば同一アイテムとみなす手法「薬効分類」を提案し, 「薬名分類」では抽出できないパターンの抽出を試みる.

3.3 薬剤投与における時間間隔の導出

3.3.1 TI-SPM の問題点

薬剤を投与する際の時間間隔は医学的に重要であるため, 正確に求める必要がある. しかし, 従来研究で用いられた 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用いられるアルゴリズムであり, TI-セットを構成するにあたって人為的な入力を与える必要がある. そのため, アイテム間の時間間隔が人為的に定めた TI-セットによって変化してしまい, 正確なものとならない問題点がある. そこで, 本研究では次節で説明する外れ値処理を含んだ統計情報を用いて 2 アイテム間の時間間隔を求める手法を導入することで薬剤投与における正確な時間間隔の導出を行う.

3.3.2 T-PrefixSpan

本研究では, 前章で説明した TI-SPM の問題点を解消するべく, Huang [7] らが提案した手法を参考に T-PrefixSpan を導入する. Huang らの手法との相違点はアイテム間の時間間隔の最小値と最大値の他に, 最頻値, 中央値, 平均値の 3 つの指標を加えて導出する点である. まずはじめに, T-PrefixSpan に関連する概念の定義を行い, T-PrefixSpan の説明を行う. 以下のようにタイムアイテム, タイムシーケンス, 時間間隔, タイムサブシーケンス, タイム SDB を定義する.

定義 14. 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 $i \in I$ の発生した時刻が 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 15. 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**とし定義し, 以下で表す.
$s = < (i_1,t_1),(i_2,t_2),...,(i_n,t_n) >$
同じ時刻に発生したタイムアイテムは辞書順に並ぶものとする. 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 $O_s = < i_1,i_2,...,i_n >$ を s の**オリジナルシーケンス**と呼ぶ.

定義 16. 時間間隔 TI_k
タイムシーケンス $s = < (i_1,t_1),(i_2,t_2),...,(i_n,t_n) >$ において, **時間間隔** TI_k を次で定義する.
$TI_k \equiv t_{k+1} - t_k \ (k = 1,2,...,n-2,n-1)$

定義 17. タイム SDB D
タイムシーケンス集合 S が与えられた時, **タイム SDB** D を以下で定義する.

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

ただし, D の任意の 2 要素の識別子 sid は異なる値を持つこととする.

タイム SDB に含まれるすべ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を以下のように定義する. さらに, 本研究において飽和タイム頻出シーケンシャルパターンを定義する.

定義 18. タイム頻出シーケンシャルパターン P
最小支持度 $MinSup$ $(0 \leqq MinSup \leqq 1)$, タイム SDB D が与えられたとき, $P = < i_1,X_1,i_2,X_2,...,i_{n-1},X_{n-1},i_n >$ $(\forall j \ \ i_j$ はアイテム, $\forall k \ \ X_k$ は 5 つの値の組 $(min_k,mod_k,ave_k,med_k,max_k)$ について, シーケンス $O_P = < i_1,i_2,...,i_{n-1},i_n >$ 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**と定義する.
ただし, $min_k,mod_k,ave_k,med_k,max_k$ は以下で示すものとする. オリジナルシーケンスを構成した時, O_P をサブシーケンスとするような D に存在するすべてのタイムシーケンス $S = < i'_1,t_1,i'_2,t_2,...,i'_{m-1},t_{m-1},i'_m >$ において, $i_k = i'_{j_k}, i_{k+1} = i'_{j_{k+1}}$ を満たす $k = 1,2,...,n-1$, $1 \leqq j_1 < j_2 < ... < j_{n-1} < j_n \leqq m$ を考えた時, 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,$min_k = \min Set_{TI_k}$, 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, $max_k = \max Set_{TI_k}$ とする. ここで, 時間間隔

$X_j = (min_j, mod_j, ave_j, med_j, max_j)$ $(1 \leqq j < n)$ 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してよく, 特に $min_j = max_j = 0$ のときは同日に起こるとして良い.
また, O_P を P **のオリジナルパターン**とする.

定義 19. 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 $\sum$ に属する A に対して, 以下の条件を満たす $B \in \sum \backslash A$ 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**であるという.

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

（2） 上の条件 (1) が成り立つとき,
$A = < a_1, T_1, a_2, T_2, ..., a_{n-1}, T_{n-1}, a_n >$,
$B = < b_1, T'_1, b_2, T'_2, ..., b_{m-1}, T'_{m-1}, b_m >$ と表した時, $a_k = b_{j_k}, a_{k+1} = b_{j_{k+1}}$ となる $k = 1, 2, ..., n-1$, $1 \leqq j_1 < j_2 < ... < j_n \leqq m$ が存在する. この時, すべての $T_k = (min_k, mod_k, ave_k, med_k, max_k), T'_{j_k} = (min'_{j_k}, mod'_{j_k}, ave'_{j_k}, med'_{j_k}, max'_{j_k})$ に対して, $min_k \geqq min'_{j_k}$ かつ $max_k \leqq max'_{j_k}$ を成立する.

（3） $Sup(A) \leqq Sup(B)$

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 $Sup(A) \equiv |\{s \mid s \subseteqq S, (sid, S) \in D, sid \text{ は } S \text{ の識別子}\}|$ と定義する.

例えば, 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
10	$< (A,1), (B,3), (C,7), (E,10) >$
20	$< (A,1), (B,4), (E,7) >$
30	$< (A,2), (B,6), (B,9) >$
40	$< (A,2), (B,5) >$
50	$< (A,2), (B,7) >$

このとき, SDB のオリジナル SDB O_D は表 2 のようになるため, O_D の最小支持度 $MinSup = 0.4$ における頻出シーケンシャルパターンは, $< A >$,$< B >$,$< E >$,$< A, B >$,$< B, E >$, $< A, B, E >$ となる.

表 2 D のオリジナル SDB O_D

sid	タイムシーケンス
10	$< A, B, C, E >$
20	$< A, B, E >$
30	$< A, B, B >$
40	$< A, B >$
50	$< A, B >$

O_D において要素が一つである頻出シーケンシャルパターンは, D においてそのままタイム頻出シーケンシャルパターンとなるため, $< A >$,$< B >$,$< E >$ は D においてタイム頻出シーケンシャルパターンである. 次に, $< A, B >$ におけるアイテム A とアイテム B の時間間隔を考えた時, D から求められる時間間隔の集合は $\{2, 3, 3, 4, 5\}$ となるため, その最小値, 最頻値, 平均値, 中央値, 最大値を考えると, $< A, (2, 3, 3, 3, 5), B >$ がタイム頻出シーケンシャルパターンとなり, $< B, E >$,$< A, B, E >$ についても同様にタイム頻出シーケンシャルパターンを求めると,$< B, (3, 5, 5, 5, 7), E >$,$< A, (2, 2, 2, 2, 3), B, (3, 5, 5, 5, 7), E >$ となる. よって, 最終的に D の最小支持度 $MinSup = 0.4$ におけるタイム頻出シーケンシャルパターンは $< A >$,$< B >$,$< E >$,$< A, (2, 3, 3, 3, 5), B >$,$< B, (3, 5, 5, 5, 7), E >$, $< A, (2, 2, 2, 2, 3), B, (3, 5, 5, 5, 7), E >$ となる. また, 飽和タイム頻出シーケンシャルパターンは $< A >$,$< B >$, $< A, (2, 3, 3, 3, 5), B >$,$< A, (2, 2, 2, 2, 3), B, (3, 5, 5, 5, 7), E >$ となる.

本研究では PrefixSpan [6] を元にタイム SDB からタイム頻出シーケンシャルパターンを導出する T-PrefixSpan を導入した. T-PrefixSpan のアルゴリズムは以下の **Algorithm 1** の通りである. ただし, タイム SDB D のオリジナル SDB を Original(D), タイムシーケンス S のオリジナルシーケンスを Original(S) と表し, シーケンス A とシーケンス B の連接を AB と表記する. また集合 X の n 番目の要素を X_n, シーケンス S の n 番目の要素を S_n, タイムシーケンス A の n 番目のタイムアイテムにおける時刻ををを $T(A_n)$ とする. 時間間隔集合における外れ値については, Smirnov-Grubbs 検定 [9] を用いて有意水準 $\alpha = 0.05$ で除去した.

医療行為データにおける (大別 Type, 詳しい説明 Explain, 薬効コード Code, 薬剤名称 Name) の 4 つ組をアイテムと見なし, これに医療行為を行った時刻 t を与えることでタイムアイテムとした. その後, ある患者に対して入院から退院まで行ったタイムアイテムを要素としたタイムシーケンスを構成する. 入退院期間が異なれば, 同じ患者であったとしても別タイムシーケンスとした. タイムシーケンスにおいて同時刻に発生したタイムアイテムは, アイテムが同一であればシーケンスからの削除を行い, その後 Type, Explain, Code, Name の順に辞書順にソートを行う. このように構成したタイムシーケンスからタイム SDB を構成し, T-PrefixSpan を適用することで, 薬剤情報とアイテム間の時間間隔情報を含んだパターンを得られる.

4. 実　　験

これまで本研究で用いる薬剤情報の取り扱いについて説明した. 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, 提案手法を適用し, 薬剤情報の取り扱いによって実験結果がどのように変化するのかを観察し, 出力として得られた典型的な流れと医師が経験をもとに作成したクリニカルパスとの比較を行う.

4.1 実験対象データ

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行為デー

Algorithm 1 T-PrefixSpan

Input : タイム SDB D, 最小支持度 $MinSup$

Output : タイム頻出シーケンシャルパターンの集合 P

Call : T-PrefixSpan(<>,D)

Procedure : T-PrefixSpan(α,$D\mid_\alpha$)

1: $D'\mid_\alpha$= Original($D\mid_\alpha$)
2: **if** α ! = $null$ **then**
3: $P \leftarrow$ GetProperTime($\alpha, D\mid_\alpha, D'\mid_\alpha$)
4: **end if**
5: $B \leftarrow \{\beta \mid (s \subseteqq D'\mid_\alpha, \beta \in s) \wedge (Sup(\beta) \geqq Size(\mathrm{D}) \times Minsup)\}$
6: **for** $\beta \in B$ **do**
7: $D\mid_{\alpha\beta} \leftarrow \{ < sid, s > \in D\mid_\alpha \mid \alpha\beta \subseteqq$ Original(s)$\}$
8: Call T-PrefixSpan($\alpha\beta, D\mid_{\alpha\beta}$)
9: **end for**

Subroutine : GetProperTime($\alpha, D\mid_\alpha, D'\mid_\alpha$)

1: **if** $length(\alpha) == 1$ **then**
2: **return** α
3: **end if**
4: $K \leftarrow \{k \mid < sid, s > \in D\mid_\alpha$,Original($s$)$\in D'\mid_\alpha$,$k \subseteqq s$,Original(k) == $\alpha\}$
5: $T = \{\{\}, \{\}, ..., \{\}\}(\mid T \mid = length(\alpha) - 1)$
6: **for** $k \in K$ **do**
7: **for** $i = 0, ..., length(k-1)$ **do**
8: $T_i \leftarrow T(k_{i+1}) - T(k_i)$
9: **end for**
10: **end for**
11: $W = <\alpha_0, \alpha_1, ..., \alpha_{length(\alpha)-1}>$
12: **for** $i = 0, ..., length(\alpha) - 2$ **do**
13: T_i から時間間隔の外れ値を除去
14: $min_i = \min T_i$
15: mod_i = (T_i の最頻値)
16: ave_i = (T_i の平均値)
17: med_i = (T_i の中央値)
18: $max_i = \max T_i$
19: $X_i = (min_i, mod_i, ave_i, med_{[i]}, max_{[i]})$
20: $W = <\alpha_0, ..., \alpha_i, X_i, \alpha_{i+1}..., \alpha_{length(\alpha)-1}>$
21: **end for**
22: **return** W

タを対象とする. この医療データは宮崎大学医学部附属病院で使われている電子カルテシステム WATATUMI [10] によって取得されており, 個人情報保護の観点から患者を一意に特定する情報を含んでいない. ある患者に対して行った医療行為を抽出する際には, 連結不可能な匿名化患者 ID を用いた. 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行為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.

電子カルテシステムに記録された (1) 停留精巣固定術, (2)TUR-Bt という 2 つのクリニカルパスを元に行った医療行為データを対象データセットとして, 3. 章で説明した薬剤の取り扱いを行う. (1) 停留精巣固定術は医療行為の流れが固定化しているクリニカルパスで, それに対し (2)TUR-Bt の術後の医療行為の流れはあまり定まっていないパスであるため, これら 2 つのクリニカルパスを選んだ.

4.2 実験内容

適当に定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方法「薬名分類」と薬剤名を用いずに薬効に着目したマイニング方法「薬効分類」での比較を行う. 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, その後飽和タイム頻出シーケンシャルパターンのみを出力パターンとした.

実験では薬効分類と薬名分類に対して, 出力飽和タイム頻出シーケンシャルパターン数, 平均パターン長, 薬剤が絡んだ医療行為を含むパターンの割合の比較を行う. 「薬剤が絡んだ医療行為を含むパターンの割合」とは全出力に対する薬剤が絡んだ医療行為を含むパターンの割合を表す.

比較実験の後, 抽出により得られた典型的な流れと医師が経験をもとに作成したクリニカルパスとがどの程度一致しているのか確認する.

実行環境は以下の通りである.

- OS : Windows7 Professional 64bit
- CPU : Intel(R) Xeon(R) CPU E3-1241 v3 @ 3.65GHz(8CPUs)
- Memory : 16GB
- Java 1.8.0_45

前節で説明した 2 つのデータセット (1) 停留精巣固定術 , (2)TUR-Bt の薬名分類, 薬効分類におけるシーケンス数, 平均シーケンス長, 最小シーケンス長, 最大シーケンス長は以下表 3 の通りである.

表 3 対象データセット

データセット	停留精巣固定術		TUR-Bt	
分類方法	薬名分類	薬効分類	薬名分類	薬効分類
シーケンス数	265		488	
平均シーケンス長	19.64	19.16	53.21	49.89
最小シーケンス長	10	9	11	11
最大シーケンス長	460	465	655	485

4.3 実験結果と考察

出力パターン数は図 1, 図 2, 平均パターン長は図 3, 図 4, 薬剤が絡んだ医療行為を含むパターンの割合は図 5, 図 6 に示す.

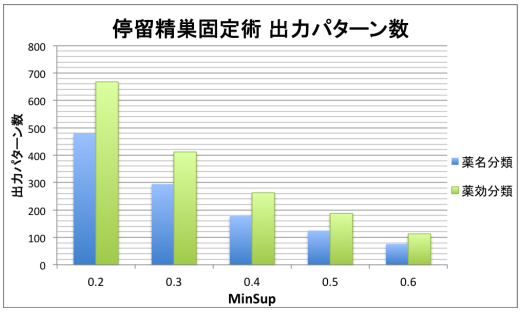

図 1 停留精巣固定術 出力パターン数

上の結果を受けて, 考察を行う. 実験結果より, 停留精巣固定

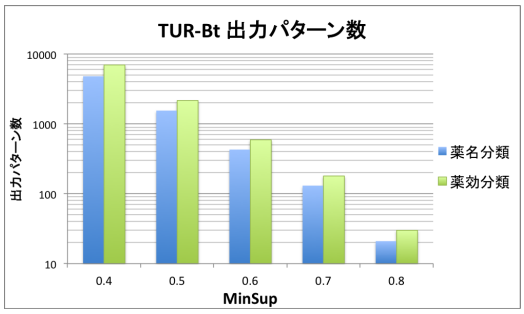

図 2　TUR-Bt 出力パターン数

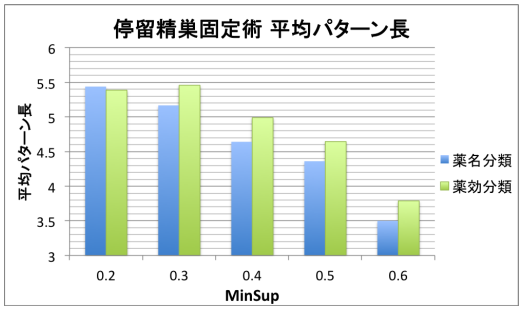

図 3　停留精巣固定術 平均パターン長

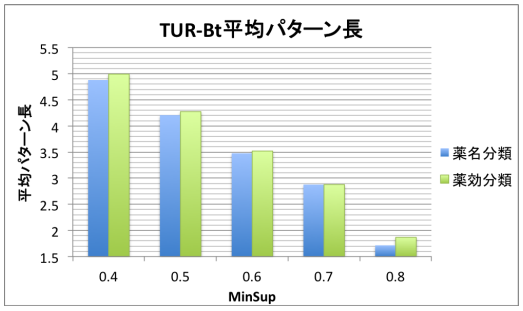

図 4　TUR-Bt 平均パターン長

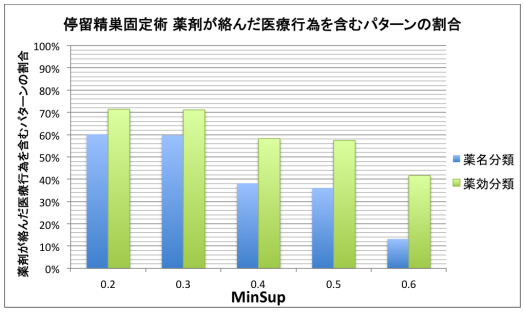

図 5　停留精巣固定術 薬剤が絡んだ医療行為を含むパターンの割合

術と TUR-Bt のどちらでも出力パターン数が薬名分類より薬効分類の方が大きくなっている．本来薬名で区別していたアイテムを薬効が同じであれば同一アイテムと見たために，サポート計算を行う際に最小支持度を超えるアイテムの種類数が増えるために出力数が増大していると言える．また，最小支持度を超えるアイテムの種類数が増えると T-PrefixSpan の再帰回数が増えるため，薬効分類のほうが実行時間も大きくなる．どちらのデータセットについても薬名分類より薬効分類の方が平均パターン長は大きくなる傾向にあるのは，患者毎に同一の薬名，異

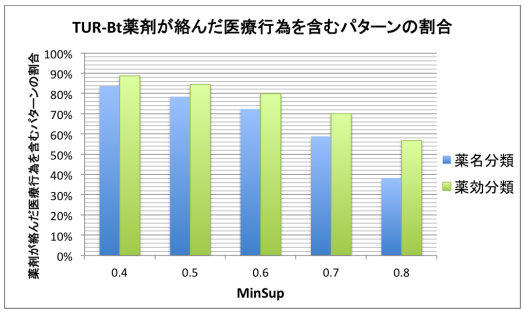

図 6　TUR-Bt 薬剤が絡んだ医療行為を含むパターンの割合

なる薬効の薬剤を投与していることが頻繁に見られることを意味している．薬剤が絡んだ医療行為を含むパターンの割合が大きくなっているのは，薬名分類では抽出することができなかった薬剤が絡んだ医療行為を多く抽出できたためといえる．

実際に最小支持度 0.02 から最小支持度 1.0 までの抽出によって得られたすべての出力に対して，手術を 0 日目として各実施日に起こったアイテムをまとめると下図 7 のような医療行為の流れにほぼすべてのパターンが含まれることがわかった．青枠が手術，赤枠が薬剤が絡んだ医療行為，緑枠が薬剤が絡んでいない医療行為を表す．2 アイテム間の時間間隔の最小値と最大値が一致するとき，その 2 アイテム間の時間間隔が一定となることを用いて”目視”で作成した．図 7 は同日におこったアイテムを日毎の集合として大まかに表しており，実施日が不定の医療行為は除いてある．

医師が経験を元に作成したクリニカルパスを図 8 に示す．クリニカルパスの図においても実施日が不定のアイテムは除いてある．実施日とアイテムが両方一致しているものを赤丸で，実施日は異なるがクリニカルパス中には存在するアイテムを橙色三角で表す．図 7 と図 8 確認してみると，薬剤情報を含まない医療行為についてはアイテムと実施日の両方が一致している比率が高いが，薬剤情報を含むアイテムの場合は実施日が一致していないものが多いことがわかる．これは医療関係者が想定した医療行為の流れと実際に患者に行っている医療行為の流れにはある程度差異がある一方で，薬剤が絡まない医療行為については本手法によって医師が望む結果を抽出できることを意味する．クリニカルパスで用いられている薬剤を本研究の手法では抽出することができなかったのは，最低支持度 $MinSup = 0.02$ と極小さな値によるマイニングによっても得られることが出来なかったため，データセットによるものといえる．図 8 のクリニカルパスに現れず，図 7 にのみ現れた医療行為については，クリニカルパス上で実施日が不定であるかそもそも現れないものである．このため，図 7 においてクリニカルパスと一致していない部分が果たして医学的に有益なのかを医療関係者と議論する必要がある．

5.　まとめと今後の課題

5.1　ま　と　め

本研究では医療行為データから生成される SDB において，従来研究では考慮していなかった薬剤情報に着目した提案手法を

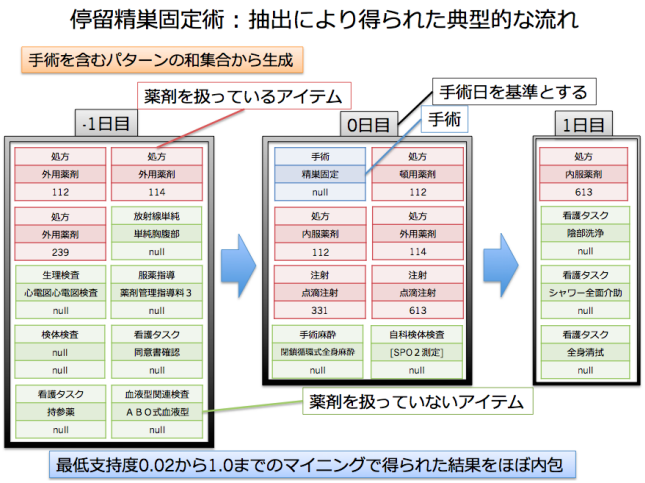

図 7　停留精巣固定術 抽出により得られた典型的な流れ

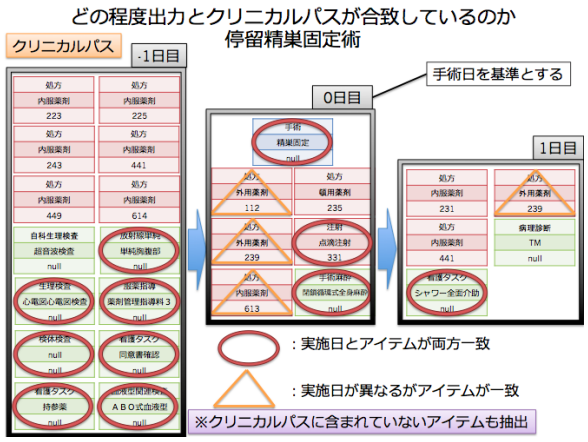

図 8　停留精巣固定術 クリニカルパス

適用した．提案手法を用いた実験の結果，薬効分類の方が薬名分類よりも薬剤情報を含むパターンを多く抽出できた．

また，医療行為間の時間間隔を最小値，最頻値，平均値，中央値，最大値の 5 つの指標によって提示することで，医師がクリニカルパスを作成する場合の支援ができるようになった．

さらに，停留精巣固定術の典型的な流れをおおまかに示すことができ，これは薬剤情報が絡まない医療行為に関して，医師が経験をもとに作成したクリニカルパスと類似した結果であった．

5.2　今後の課題

今回タイム頻出シーケンシャルパターンの概念を PrefixSpan [6] に適用したが，T-PrefixSpan では実時間で計算することができない低い最低支持度でマイニングを行うためには，高速なアルゴリズムを導入する必要があることが挙げられる．飽和頻出シーケンシャルパターンを高速に求めるアルゴリズムとして，CloSpan [4],Clasp [12],CSpan [13] が存在するため，これらのアルゴリズムを拡張することが考えられる．

また，今回の研究では 2 つのクリニカルパス適用患者に行った医療行為に対して手法を適用したが，他のクリニカルパスに対しても同様の手法を適用し，改善を目指していきたい．

本研究では，T-PrefixSpan と TI-SPM の比較を行っていないため，今後行う必要がある．

抽出した典型的な流れは目視による確認の元行ったため，医療行為の分岐「バリアント」を考慮した適切な形で医療関係者に出力を提示する手法を検討する必要がある．今回確認を行ったのは停留精巣固定術のみであるため，TUR-Bt のみならず他のクリニカルパスでも確認を行わなければならない．

最後に，出力がどの程度医学的に有益なのかを評価を行い，医療関係者と議論する必要がある．評価を行う際に最小値及び最大値を用いることが予測されるが，最頻値，平均値，中央値を評価に組み込むことも考えられる．

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 (A) (#25240014) の助成により行われた．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行為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.

[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] Zhengxing Huang, Xudong Lu and Huilong Duan. On mining clinical pathway patterns from medical behaviors. Artificial Intelligence in Medicine 56 (2012) 35-65, 2012.

[8] Aileen P. Wright, Adam T. Wright, Allison B. McCoy and Dean F.Sittig. The use of sequential pattern mining to predict next prescribed medications. Journal of Biomedical Informatics 53(2015) 73-80, 2015.

[9] http://aoki2.si.gunma-u.ac.jp/lecture/Grubbs/Grubbs.html

[10] 電子カルテシステム WATATUMI. http://www.corecreate.com/02_01_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] Antonio Gomariz, Manuel Campos, Roque Marin and Bart Goethals. Clasp: An efficient algorithm for mining frequent closed sequences. PAKDD 2013, LNAI7818, Part I, pp. 50-61, 2013.

[13] V.Purushothama Raju and G.P. Saradhi Varma. MINING CLOSED SEQUENTIAL PATTERNS IN LARGE SEQUENCE DATABASES. International Journal of Database Management Systems (IJDMS) Vol.7, No.1, February 2015.