

動的に医療指示を除外したシーケンス解析による 異なる患者グループ間の治療パターン比較

黒川 健人[†] Le Hieu Hanh[†] 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

^{††} 一般社団法人ライフデータイニシアティブ 〒606-0805 京都市左京区下鴨森本町 15

^{†††} 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: [†]{kurokawa_k, hanhhlh}@de.cs.titech.ac.jp, [†]yamazaki.cp@gmail.com, [†]yokota@cs.titech.ac.jp,
^{††}matsuo@ldi.or.jp,
^{†††}araki6925@gmail.com

あらまし 近年、電子カルテの普及により、蓄積された医療情報の二次利用による有効活用が期待されている。二次利用の例として、先行研究では、入院から退院までの標準的な医療指示の流れを表した頻出医療シーケンスの抽出や、複数病院間のシーケンスの比較などを行ってきた。本研究では、動的に医療指示を除外したシーケンス解析による、異なる患者グループ間の治療パターンを比較する手法を提案する。解析対象の医療指示を変更して解析を行うことで、基礎疾患や既往症の有無などで分類した患者グループ間の治療パターンについて、最長の治療パターンから最短の治療パターンまで、網羅的に治療パターンを比較することができる。患者グループ間の治療パターンを比較して、特徴的な治療パターンを明らかにすることで、各患者に適した医療行為を行えるよう支援することを目的としている。実験では、新型コロナウイルスの電子カルテに対して提案した手法を適用し、基礎疾患の有無で分類した患者グループ間での治療パターンを比較する。

キーワード 電子カルテ, 医療支援, シーケンシャルパターンマイニング, 新型コロナウイルス

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が大規模病院を中心に普及しており、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。特に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]の計画は、電子カルテの普及に加えて、ビッグデータ解析としての医療情報の二次利用が期待されている [2]。二次利用の例として、入院から退院までの標準的な医療行為の流れであるクリニカルパスの生成支援が挙げられる。一般的に、クリニカルパスの作成は医療従事者の医学的経験に基づいて行われることが多いので、高度な専門的知識と多くの労力が必要である。そのため、データ工学の観点から様々な手法を電子カルテに適用して解析することで、医療行為改善の支援を行う研究が行われてきた。本研究では、クリニカルパス生成支援のために、異なる患者グループ間での医療シーケンスの解析により、治療パターンを比較する。基礎疾患や既往症の有無などで分類した患者グループ間における治療パターンを比較することで、特徴的な治療パターンが明らかになり、医療の効率化や安全性の向上が見込まれる。

1.2 本研究の目的

本研究は、基礎疾患や既往症の有無などを基に分類した、二つの患者グループ間での医療シーケンスの解析を行い、治療パターンを比較する。

アプローチとして、電子カルテデータベースから二つの患者グループの医療指示データを対象として、シーケンシャルパターンマイニング (Sequential Pattern Mining 以下 SPM) をそれぞれの患者グループに適用して、頻出治療パターンの抽出する。電子カルテに含まれる医療指示はシーケンスとなるため、医療指示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、頻出治療パターンを抽出できる。次に、それぞれの頻出治療パターンをシーケンスバリエーションとして併合シーケンスバリエーション (Merged Sequence Variant 以下 MSV) の作成を行う。MSV の作成により、頻出治療パターン同士の差異を表すことが可能である。さらに、MSV を基に医療指示を変更して、SPM, MSV の作成を繰り返すことで、MSV ネットワーク (Merged Sequence Variant NetWork, 以下 MSVNW) の作成を行い、MSV を網羅的に求める。最後に網羅的に求めた MSV に対して、SV 間の治療パターンの差異の程度を MSV の特徴量として表現して、MSV のうち、着目すべき MSV を決める。

SPM は最長の頻出治療パターンを抽出するアルゴリズムであるので、異なる患者グループのそれぞれに対して SPM を適

用して差異を表す MSV を作成することができるが、最長の治療パターン同士を比較することに限られていた。本研究ではマイニング対象の医療指示を除外することで、最長でない治療パターン同士の差異を表した MSV を作成する。網羅的に求めた MSV を基に、基礎疾患の有無など異なる患者グループ間の治療パターンの比較を行い、医療行為の改善に貢献することを目指す。

1.3 関連研究

山田ら [5] は、頻出シーケンスに関する情報を、シーケンスの識別子であるシーケンス ID を保持しながらマイニングすることによって、頻出シーケンスごとの安全性や効率性の指標を算出し、SV 評価が正確に行える可視化手法を提案した。坂本ら [6] は、分岐先頻度と実施時間情報を用いたグループ化による適切分岐候補選択を行い、頻出シーケンス分岐先の適性候補選択による併合シーケンスの生成を行った。また、分岐要因である患者情報の拡充をし、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た。李ら [4] は、シーケンス間の共通部分を抽出し、共通部分とシーケンスを併合して可視化することによって、二つの医療機関の治療パターンの違いを明示する手法を提案した。安光ら [7] は、クラスタリングを用いることで三つ以上の医療機関における頻出治療パターンの比較する手法を提案した。趙ら [8] は、頻出治療パターンの時期による差異の可視化と差異の発生時期の推定に関する手法を提案した。

1.4 本稿の構成

本稿は以下の通り構成される。2 節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 節では提案概念である、動的な医療指示の除外の手法と、MSVNW を求めるアルゴリズム、特徴量の算出について述べる。4 節では 3 節の手法を用いて実際の医療指示データを解析し、MSVNW の作成と比較した治療パターンを示す。5 節では、提案手法に関する考察を述べる。最後に 6 節でまとめと今後の課題について述べる。

2 背景知識

本節では、本研究に関連するシーケンシャルパターンマイニング (SPM)、併合シーケンスバリエーション (MSV) について説明する。

2.1 シーケンス

定義 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。

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

$i_j \in I$ がアイテムとなる。

アイテムを解釈するとき、 i をアイテム i の name と呼ぶ。

定義 2. (シーケンス). アイテムセット I に対し、シーケンス S を以下のように定義する。

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

$s_j = (id, i), i \in I$. id は S の中でユニークなインデックスとなる。 \prec_S は S 上の全順序関係である。 $\forall s_i, s_j \in S, s_i \prec_S s_j \vee s_i \succ_S s_j$.

2.2 SV

定義 3. (シーケンスバリエーション). アイテムセット I に対し、束 SV をシーケンスバリエーションと定義し、以下のように表す。

$$SV = (\{sv_1, sv_2, \dots, sv_l\} | \prec_{SV})$$

$sv_j = (id, i), i \in I$. id は SV の中でユニークなインデックスとなる。 \prec_{SV} は SV 上の半順序関係である。また、 SV が上限および下限を持っている。

以下、 A, B, C のような大文字をアイテムとして説明する場合、 A, B, C は id 付きアイテムの name を指す。

シーケンスにおいて分岐が発生した場合、シーケンスバリエーションになる。例として、 A, B, C, D, E をアイテムとする、シーケンス $\langle A, B, C, E \rangle$ と $\langle A, B, D, E \rangle$ に対して、 C と D がバリエーションと呼ばれ、この二つのシーケンスをまとめてシーケンスバリエーションと考える。また、 $A \prec B, B \prec C, B \prec D, C \prec E, D \prec E$ を、 $\langle A, B, (C, D), E \rangle$ と表す。

SV を可視化する方法として、シーケンスバリエーショングラフを定義する。

定義 4. (シーケンスバリエーショングラフ). シーケンスバリエーション SV に対し、有向非巡回グラフ G_1 をシーケンスバリエーショングラフと定義し、以下のように表す。

$$G_1 = (V_1, E_1)$$

ノード集合 V_1 について SV のアイテム集合となる。また、 $\forall e = (sv_{out}, sv_{in}) \in E_1, sv_{out} \prec_{SV} sv_{in}$.

シーケンスバリエーショングラフを可視化することで、簡単にシーケンスバリエーションを示すことができる。例として、 A, B, C, D, E がアイテムで、 $SV = \langle A, B, (C, D), E \rangle$ のシーケンスバリエーショングラフが図 1 のように示されている。シーケンスバリエーション

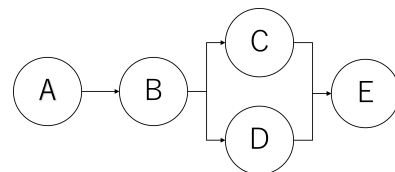


図 1: シーケンスバリエーショングラフ

とシーケンスバリエーショングラフは互いに変換可能であるため、以下両者を同じものと扱う。

2.3 LCSV・MSV

李ら [4] は二つの SV の差異を示すために、SV 間の共通部分を抽出する最長共通サブシーケンスバリエーション (以下 LCSV) と、SV 間の差異を表現する併合シーケンスバリエーション (以下 MSV) を考案した。

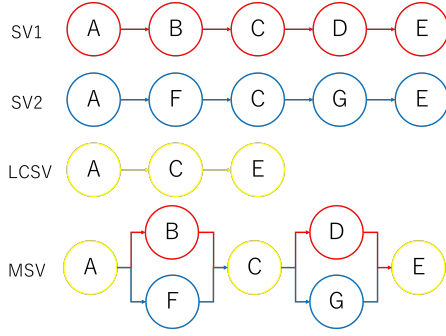


図 2: SV1 と SV2 による LCSV と MSV の例

2.3.1 LCSV

定義 5. (共通サブシーケンスバリエント (CSV)). シーケンスバリエント SV_α と SV_β に対し, 共通サブシーケンスバリエント CSV を下記の条件を満たすシーケンスバリエントと定義する. $\forall csv_i, csv_j \in CSV, csv_i, csv_j \in SV_\alpha \cap SV_\beta$ かつ $csv_i \prec_{CSV} csv_j \iff csv_i \prec_\alpha csv_j \wedge csv_i \prec_\beta csv_j$.

CSV のサブシーケンスバリエントも CSV となるので, CSV は複数存在する. そのため, 最長共通サブシーケンスバリエントを定義する.

定義 6. (最長共通サブシーケンスバリエント (LCSV)). シーケンスバリエント SV_α と SV_β に対し, 最長共通サブシーケンスバリエント $LCSV$ を下記の条件を満たすシーケンスバリエントと定義する. SV_α と SV_β に対する任意の共通サブシーケンスバリエント CSV に対し, $|CSV| \leq |LCSV|$. $|CSV|$ は CSV のノード数を表す.

2.3.2 MSV

定義 7. (併合シーケンスバリエント (MSV)). シーケンスバリエント SV_α, SV_β とその $LCSV$ に対し, 併合シーケンスバリエント MSV を下記の条件を満たすシーケンスバリエントと定義する. $\forall sv_i, sv_j \in SV_\alpha \cup SV_\beta, sv_i, sv_j \in MSV$ かつ $sv_i \prec_{MSV} sv_j \iff sv_i \prec_\alpha sv_j \wedge sv_i \prec_\beta sv_j$. $\forall msv_i, msv_j \in MSV, msv_i \prec_{MSV} msv_j \implies (msv_i, msv_j \in SV_\alpha \wedge msv_i \prec_\alpha msv_j) \vee (msv_i, msv_j \in SV_\beta \wedge msv_i \prec_\beta msv_j)$

以下 SV1 と SV2 による MSV の具体的な例を図??に示す.

2.4 SPM

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 [3] である. 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, シーケンスの識別子であるシーケンス ID を組とする要素からなる. Agrawal は利用者により設定された最小支持度 (以下 minsup) を満足するシーケンスを SDB から抽出する手法として, アプリオリベースの SPM を提案した. 基本的な考え方は, アイテム集合のある要素の支持度が minsup を満たしていなければ, その要素を組み合わせた場合の支持度も minsup を満たすことはないという事象を基にしている [2].

3 提案手法

本節では, 動的な医療指示の除外アルゴリズム, 網羅的に求めた MSV を表現するネットワーク構造とその作成アルゴリズム, 治療パターンの差異に基づいた特徴量の算出手法, 特徴量を基に着目すべき MSV の決定方法提案する. 前提としてマイニング対象とする医療指示を全て含む SDB から, SPM によって頻出シーケンスを抽出し, SV が生成可能であるものとする.

3.1 動的な医療指示の除外アルゴリズム

本研究ではマイニング対象の医療指示について最小限の除外により, MSV を網羅的に求めるための手法として, 動的に医療指示を除外する手法を提案する.

3.1.1 除外する医療指示の候補

MSV に含まれる医療指示が除外する候補である. すなわち以下の条件を満たすアイテム集合 I_α を除外する医療指示の候補とする.

$$\forall I_\alpha = (\{i_{\alpha 1}, \dots, i_{\alpha n}\} | \exists i_{\alpha j} \in MSV) \in SDB$$

ここで $i_{\alpha j}$ は全て同じ医療行為を表す医療指示であるものとする.

ここで除外する医療指示の候補として, MSV を構成する医療指示を考慮すれば十分である理由を述べる. 今回, 用いたマイニングアルゴリズムはアプリオリベースの SPM であり, 最長の頻出治療パターンのみが抽出される. ここで, 最長頻出治療パターンを構成しない医療指示を除外して SPM を行った場合は, 抽出される最長治療パターンに影響を及ぼさないので, 頻出治療パターン同士を併合した MSV も除外する前後で同様の MSV が作成される. したがって, MSV を網羅的に求めるためには, MSV を構成する医療指示をマイニング対象から除外すれば十分である.

3.1.2 医療指示除外アルゴリズム

上記の除外する医療指示の候補に基づいて, MSV を網羅的に求めるアルゴリズムを以下に示す.

1. 電子カルテにおける二つの患者グループを対象としてそれぞれ SPM を適用して, 頻出シーケンスを基に MSV の作成をする.
2. MSV に対して, 除外する医療指示の候補を抽出する.
3. 2. の候補からいずれか一つの医療指示を除外した上で, SPM を適用して, MSV の作成をする.
4. 以下, 任意の MSV に対して, 2. を行い, 候補となる医療指示のそれぞれに対して, 3. を行う. 新たに MSV を作成できなくなった時点でアルゴリズムを終了する.

. 医療指示除外アルゴリズムの例を示す.

$MSV_0 = \langle OT_{day0}, INS_{day0}, (INS_{day1}, DT_{day0}), OT_{day1} \rangle$ とする. $OT = \{OT_{day0}, OT_{day1}\}$, $INS = \{INS_{day0}, INS_{day1}\}$, $DT = \{DT_{day0}, DT_{day1}\}$, $MT = \{MT_{day0}, MT_{day1}\}$ として, OT, INS, DT, MT

が医療指示であり、日付を考慮した同一の医療行為の集合を示すものとする (表 3 参照)。この場合、 MSV_0 に含まれる医療指示は、 $I_\alpha = OT \vee INS \vee DT$ であるので、除外する医療指示の候補も I_α である。

ここで、 MSV_0 から OT を除外したとき、 $MSV_1 = \langle MT_{day0}, INS_{day0}, (INS_{day1}, DT_{day0}), MT_{day1} \rangle$ が作成されたとする。この場合、除外する医療指示の候補は、 $I_\alpha = MT \vee INS \vee DT$ である。以下、任意の MSV に対して、候補となる医療指示をそれぞれ除外して、 MSV を網羅的に求めることができる。

3.2 MSV ネットワーク (MSVNW)

上記に示した医療指示除外アルゴリズムを適用して、網羅的に MSV 同士の関係を表現するため、 MSV ネットワーク (以下 $MSVNW$) を導入する。

定義 8. MSV ネットワーク ($MSVNW$). 網羅的に求めた MSV を表現するネットワーク構造として、有向非巡回グラフ G_2 を定義し、以下のように表す。

$$G_2 = (V_2, E_2)$$

ノード集合 V_2 について、以下の条件を満たすものとする。

$\forall v = id \in V_2$. ここで、 id はある MSV の識別子を表すものとする。また、エッジ集合 E_2 について、以下の条件を満たすものとする。 $\forall e = (\{v_i, v_j\}, I_\alpha) | \exists v_i, v_j \in V_2 | \exists I_\alpha \in SDB \in E_2$ 。

次に、 $MSVNW$ を作成するアルゴリズムについて述べる。以下、 v について MSV の識別子 $(0, 1, 2, \dots)$ を用いて、 $v = 0, 1, 2, \dots$ と表す。また、 e についても同様に MSV の識別子のみを用いて、 $e = \{0, 1\}, \{0, 2\}, \{1, 3\}, \dots$ と表す。

1. $V_2 = \phi, E_2 = \phi$ とする。医療指示を一切除外せずに、SPM で抽出された治療パターンの差異を表す MSV に識別子として $id = 0$ を与える。
2. $id = 0$ に対応するノード $v = 0$ を生成する。
3. 医療指示を除外したことで作成される MSV に対して、 $id = 1$ を与える。
4. $id = 1$ に対応する $v = 1$ を生成して、 $e = \{0, 1\}$ となるエッジを生成する。また、 e に除外した医療指示を付与する。
5. $v = i$ である任意の MSV から、医療指示を除外したことで作成される MSV に対して、生成済みのノードの最大値を $v = j$ として、 $id = j + 1$ を与える。新たに MSV が作成されなくなった時点でアルゴリズムを終了する。
6. MSV に対応する $v = j + 1$ を生成して、 $e = \{i, j + 1\}$ となるエッジを生成する。また、 e に除外した医療指示を付与する。
7. $v = j + 1$ について、 $v = 0$ から $v = j + 1$ までを結んだ経路に付与された医療指示について、 $v = 0, \dots, j$ のいずれかのノードで $v = 0$ から結んだ経路に付与された医療指示と組合せが一致している場合について 8. に進む。上記を満たさない場合には、5. に戻る。

8. $v = j + 1$ と組合せが一致したノードを $v = v'$ に対して、 $e = \{i, v'\}$ を生成して、 e に除外した医療指示を付与する。また、 $e = \{i, j + 1\}$ と $v = j + 1$ を削除する。5. に戻る。

7. , 8. の結果、あるノードについて $id=0$ から辿った経路が複数あるとき、除外した医療指示の組合せが同じで、除外した順序が異なる MSV であることを表している。

$MSVNW$ をアルゴリズムに基づいて作成した例を示す。

$id = 0$ の MSV について、 OT または DT が除外する医療指示であるとき、 $MSVNW$ は図 3 のように表現される。

また、図 3 の $id = 1$ は、 INS または DT が除外する医療指示であり、 $id = 2$ は、 OT が除外する医療指示であるとき、 $MSVNW$ は図 4 のように表現される。

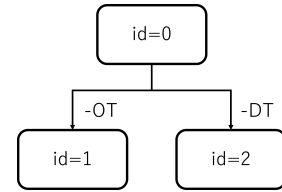


図 3: $MSVNW$ の作成例 1

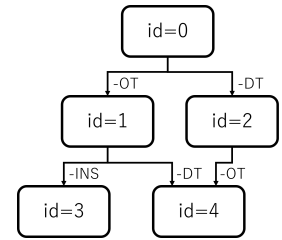


図 4: $MSVNW$ の作成例 2

3.3 特徴量の算出

ネットワーク構造で網羅的に表現された MSV に対して、治療パターンの差異に基づいて着目すべき MSV を決定する。ここで、治療パターンの差異の程度を表す指標が存在しないため、 MSV のノードに基づいた特徴量を差異の程度を表す指標として定義する。 MSV の表す治療パターンの差異の程度として二つの観点で見たときに、SV 間差分 (以下 $DiffSV$), SV 間共通度 (以下 $CmnSV$), SV 間類似度 (以下 $SimSV$) を MSV の特徴量と定義することができる。算出された特徴量は $MSVNW$ のノードに付与されるものとする。

まずは治療パターンの差異の程度について二つの観点を説明する。

1. 数量ベース

異なる患者グループ間の治療パターンの差異について SV 間のノードの数量に基づいて考える。

2. 割合ベース

異なる患者グループ間の治療パターンの差異について SV 間のノードの割合に基づいて考える。

それぞれの観点での特徴量を考える。以下、シーケンスバリエーション SV_α と SV_β における, $LCSV$ が既知であるとする。また, $|SV|$ は SV のノード数を表す。

3.3.1 数量ベースでの特徴量 DiffSV・CmnSV

SV 間のノードの数量に基づいた治療パターンの差異を表す指標として DiffSV と CmnSV を定義する。

定義 9. (SV 間差分 (DiffSV)). SV_α と SV_β 間の治療パターンの差異が大きいことを表す指標を, SV 間差分 $DiffSV$ として以下のように定義する。

$$DiffSV = |SV_\alpha| + |SV_\beta| - 2|LCSV|$$

ここで, DiffSV の値が大きいほど, SV_α と SV_β 間の治療パターンの差異が大きいと考える。

定義 10. (SV 間共通度 (CmnSV)). SV_α と SV_β 間の治療パターンの差異が小さいことを表す指標として SV 間共通度 $CmnSV$ として, 以下のように定義する。

$$CmnSV = 2|LCSV|$$

ここで, CmnSV の値が大きいほど, SV_α と SV_β 間の治療パターンの差異が小さいと考える。

3.3.2 割合ベースでの特徴量 SimSV

SV 間のノードの割合に基づいた治療パターンの差異を表す指標として SimSV を定義する。

定義 11. (SV 間類似度 (SimSV)). SV_α と SV_β 間の治療パターンの差異を表す SV 間類似度 $SimSV$ を指標として, 以下のように定義する。

$$SimSV = \frac{2|LCSV|}{|SV_\alpha| + |SV_\beta|}$$

ここで, SimSV の値が小さいほど, SV_α と SV_β 間の治療パターンの差異が大きいと考える。一方で, SimSV の値が大きいほど, SV_α と SV_β 間の治療パターンの差異が小さいと考える。

以下の図 5 に基づいて, 着目すべき MSV についての例を示す。

まず, 数量ベースでの抽出では, DiffSV が最大であるノードは id=0 であり, CmnSV が最大であるノードは id=2 であるので, それぞれ二つの患者グループ間で最も差異が大きい治療パターンと最も差異が小さい治療パターンである。

次に, 割合ベースでの抽出では, SimSV が最小であるノードは id=1 であり, 二つの患者グループ間で最も差異が大きい治療パターンである。また, SimSV が最大であるノードは id=2 と id=3 であり, 二つの患者グループ間で最も差異が小さい治療パターンである。

3.4 提案手法の利点

提案手法を用いる利点は, 網羅的に MSV を作成することで, 隠れている治療パターンについても比較できる点である。マイニング対象の医療指示を除外しない場合や, 医学的な経験則に基づいて医療指示を除外する場合と比べると, 満遍なく MSV の作成が可能であり, 医学的に一般的な治療パターンに加えて,

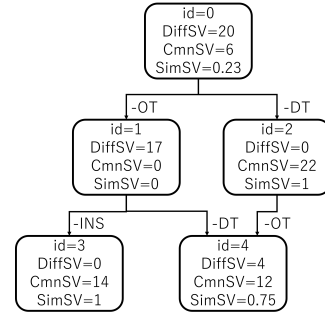


図 5: 特徴量を付与した MSVNW の例

隠れていた治療パターンを比較できる可能性がある。隠れていた治療パターンを含む特徴的な治療パターンに着目することで, クリニカルパスの改善などの医療支援に貢献することができる。

他の利点としては, ネットワーク構造で MSV の変化が推移的に表現されたので, マイニング対象の医療指示の変化と治療パターンの変化の関係が捉えやすい点である。特定の MSV と, そこから医療指示を除外することで得られた MSV を比較した場合について, MSV の特徴量が大きく変化していれば治療パターンの差異も大きく変化していると考えられるので, 除外した医療指示が治療パターンの差異に大きく影響を及ぼしていると考えられる。医療指示の除外による MSV の変化を知ること, 特徴的な治療パターンを明らかにして, クリニカルパスの改善などの医療支援に貢献することができる。

4 実験

本節では, 実際の電子カルテデータに対して提案手法を適用し, 基礎疾患の有無による治療パターンの比較を行う。

4.1 実験環境

個人情報保護のために, 利用者を限定した実験環境として, 宮崎大学医学部附属病院が管轄するクラウド環境を用いる。解析を行うメンバーは, 匿名化された電子カルテデータをクラウドから取り出して解析するのではなく, 解析プログラムの Docker Image を作成して, その Docker Image をクラウド上で実行する。個別の患者情報を含まない頻出医療指示パターンのみ含む結果を取り出すことで, 情報セキュリティとプライバシー保護を確保している。

4.2 パラメータの設定

電子カルテデータに SPM を適用したとき, MSV の構成はデータの中身に加えて SPM のパラメータにも依存する。用いる SPM はアプリアベースであり, minsup の設定と, 最長の頻出シーケンシャルパターン以外のパターンを生成するべきかの設定が可能である。minsup については, 全体の二割以上のシーケンスに含まれるアイテム要素は頻出であると仮定し, minsup=0.2 を設定する。minsup を極端に低くすることで頻出でないアイテム要素を出現させることも可能であるが, 頻出の治療パターンの抽出という目的には不適切である。minsup の変更については一考の余地があり, 複数の minsup で実験を

行ったり、医療従事者との相談により決定するのが望ましいと考えられる。また、今回は最長の頻出シーケンシャルパターンのみを生成する。最長以外のパターンの生成により、多くの医療指示で MSV を構成することも可能であるが、最長と最長以外のパターンが混在して、差異を適切に求められない欠点がある。ただし、今回の提案手法は、最長以外のパターンでも適用できるので、設定を変えて実験することも可能である。

4.3 データセット

本研究のデータは「COVID-19 感染の診療への影響調査と予測モデル開発」における、23 の研究協力機関の 1 つである、医療機関 A で実際に使用されているクリニカルパスの医療指示データを対象とする。この医療指示データは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。なお、本研究は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、COVID-19 の第 4 波 (2021 年 3 月 1 日から 2021 年 6 月 30 日) を対象として、この期間に入院から退院を完了している医療指示データを用いた。医療指示データには医療指示行為の情報、日付に関する情報が含まれており、これらの情報を基に SPM を適用した。また、本実験は基礎疾患を持つ患者と基礎疾患を持たない患者との間での治療パターンの特徴の抽出が目的であり、COVID-19 の患者データを基礎疾患の有無で二つの患者グループに分けた。基礎疾患について、医療従事者のアドバイスに基づいて表 1 を選択した。また、二つの患者グループのシーケンス数は表 2 に示した。

表 1: 基礎疾患名称

高血圧症	慢性心不全
糖尿病・糖尿病性合併症なし	うっ血性心不全
2 型糖尿病・糖尿病性合併症あり	非弁膜症性心房細動
2 型糖尿病・糖尿病性合併症なし	特発性拡張型心筋症
耐糖能異常	労作性狭心症
慢性腎不全	アルツハイマー型初老期認知症
末期腎不全	高尿酸血症
急性間質性肺炎の疑い	高尿酸血症

表 2: 基礎疾患の有無によるシーケンス数

患者グループ	シーケンス数
基礎疾患あり	88
基礎疾患なし	134

4.4 前処理

本実験では、基礎疾患の有無による治療パターンの比較を行うため、治療に寄与しない医療指示は前処理の段階で除外した。表 3 の 11 種の医療指示をマイニング対象とした。実験結果では表 3 の略号を用いて表すこととする。また、医療指示について名称の異なる、同一の医療指示である場合がある。そこで表 4 のように医療指示を同一視した。なお、マイニング対象とする

医療指示の決定と医療指示の同一視については、医療従事者のアドバイスに基づいて行った。

表 3: マイニング対象とする医療指示

医療指示種類	略号	説明
インスリン	INS	血糖値の上昇を防ぐ
アセトアミノフェン	AA	解熱鎮痛剤
デキサメタゾン	DT	PR と比べて強いステロイド剤
プレドニゾロン	PR	DT と比べて弱いステロイド剤
デキストロメトルファン	MT	咳止め剤
リン酸コデイン散	CP	咳止め剤
ヘパリン Ca	HP	血液凝固阻害作用
人工呼吸	AR	HF と比べて重症患者を対象
ハイフローセラピー	HF	AR と比べて軽症患者を対象
中心静脈注射	IVH	心臓付近に薬・栄養を補給
酸素吸入	OT	低酸素状態の改善

表 4: 同一視した医療指示

略号	同一視した医療指示
INS	インスリン, ノボリン, ノボラビッド, ヒューマログ, ヒューマリン
AA	アセトアミノフェン錠, カロナール
MT	デキストロメトルファン臭化水素酸塩錠, メジコン錠
DT	デキサメタゾン, デキサート注射液, デカドロン錠
PR	プレドニゾロン錠, プレドニン

4.5 実験結果

4.5.1 MSVNW の作成

前処理を行った医療指示データに対して MSVNW を作成する。結果は図 6 に示す。

4.5.2 治療パターンの比較

MSVNW に基づいて患者グループ間の治療パターンの比較を行う。図 6 で丸で囲ったノードが着目すべき MSV であり、特徴的な治療パターンである。図 7 から図 13 に示す。

まず、数量ベースで比較を行う場合、最も差異の大きい治療パターンの二つを、図 8、図 10 に示す。基礎疾患ありのグループでは IVH や AR を含んだ治療パターンが抽出された。基礎疾患なしのグループでは、図 8 で INS、DT、HP を含んだ治療パターンが抽出されて、図 10 では、図 8 から DT を除外した治療パターンが抽出された。また、最も共通している治療パターンの二つを、図 7、図 9 に示す。どちらのパターンにおいても OT が共通の治療パターンとして抽出された。

次に、割合ベースで比較を行う場合、最も差異の大きい治療パターンのうち、三つを、図 8、図 10、図 11 に示す。図 8 と図 10 は前述の通りである。図 11 の基礎疾患ありのグループでは、AR を含んだ治療パターンが抽出されて、基礎疾患なしのグループでは AA を含んだ治療パターンが抽出された。また、最も共通している治療パターンの二つを、図 12、図 13 に示す。図 12 では DT が、図 13 では HF が基礎疾患ありと基礎疾患なしのグループ間に共通する治療パターンとして完全に一致している。

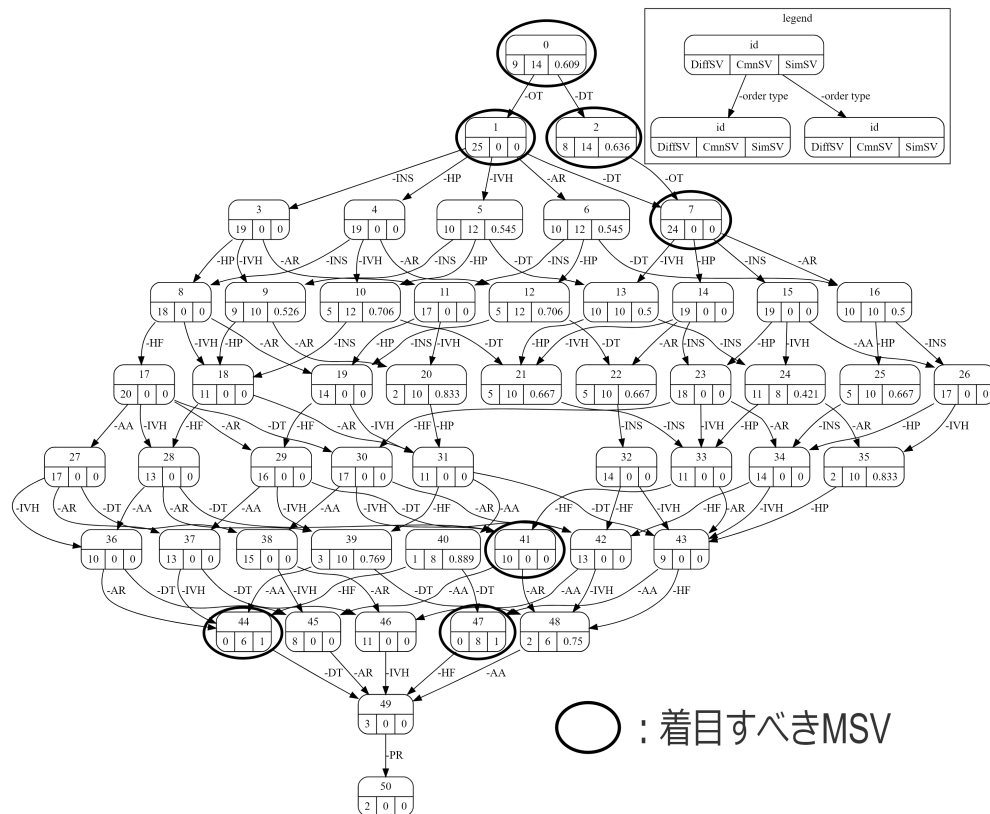


図 6: 医療指示データを基に作成した MSVNW

5 考察・評価

本手法では、最長の治療パターン同士の比較だけでなく、最長でない治療パターンを比較することも可能であり、網羅的に治療パターンを比較することが可能である。実際に本実験により、最長の治療パターンに含まれない、インスリンや人工呼吸などで構成された治療パターンを比較することができた。また、特徴量の算出について、二つの観点で特徴量の算出を行ったが、数量ベースでは MSV のノード数が多い治療パターンが抽出されて、割合ベースではノード数が少ない治療パターンが抽出される傾向があった。そのため、医療従事者が数量による差異と割合による差異のどちらの観点を重視するかにより、MSV の特徴量を選択してもらうことにする。

さらに、本手法の結果について医療従事者にも確認してもらったところ、網羅的に治療パターンを比較するのは科学的手法として適切であり、医学的に一般的でない治療パターンが得られる可能性がある点が興味深いとのコメントを頂いている。

6 結 論

6.1 ま と め

本研究は、シーケンス解析を用いて患者グループ間の治療パターンを比較して、特徴的な治療パターンが明らかになることで医療の改善支援に貢献することを目的とした。マイニング対象とする医療指示を除外する手法、医療指示を変更したときの MSV をネットワーク構造で網羅的に表現する MSVNW、

MSVNW の作成アルゴリズム、SV 間の治療パターンの差異の定量化を提案した。実際の電子カルテデータに対して提案手法を適用して、MSVNW の作成を行い、基礎疾患の有無による治療パターンの比較まで行った。

6.2 今後の課題

今後の課題として、まず適切な minsup を決定する指針に対して検討を進める必要がある。minsup の値の大小と、多様な治療パターンの抽出はトレードオフの関係にあり、医療従事者によるアドバイスや、複数の minsup での実験が重要である。また、特徴量の算出について検討を進める必要がある。本研究の DiffSV, CmnSV, SimSV は、医療シーケンスのノード数のみに基づいた算出であり、治療パターンの差異の程度をより適切に表した指標を決める必要がある。

謝 辞

本研究の一部は、日本学術振興会科学研究費 (#20H04192, #21K17746) の助成及び「COVID-19 感染の診療への影響調査と予測モデル開発」からの支援によって行われた。データを提供頂いた研究協力機関に感謝の意を表する。

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] 横田治夫. 電子カルテデータ解析-医療支援のためのエビデンス・ベースド・アプローチ, 共立出版, 2022 年 3 月.
- [3] R. Agrawal, R. Srikant. Fast algorithms for mining association rules in large databases. Proceeding of the 20th Interna-

tional Conference on Very Large Data Bases, pp. 487-499, 1994.

- [4] 李玉清, Le Hieu Hanh, 松尾亮輔, 山崎友義, 荒木賢二, 横田治夫. シーケンシャルパターンマイニングに基づく多病院間の頻出治療パターンの比較. In DEIM Forum 2022, pp. A21-4, 2021.
- [5] 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 D1-1, 2019.
- [6] 坂本任駿, 小林莉華, Le Hieu Hanh, 松尾亮輔, 山崎友義, 荒木賢二, 横田治夫. 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. DEIM Forum C25-1, 2021.
- [7] 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. DEIM Forum 5b-6-3, 2023.
- [8] 趙子泰, Le Hieu Hanh, 松尾亮輔, 山崎友義, 荒木賢二, 横田治夫. COVID-19 に関する頻出医療指示パターンの時期による差異と差異発生時期の可視化. DEIM Forum 5a-2-3, 2023.

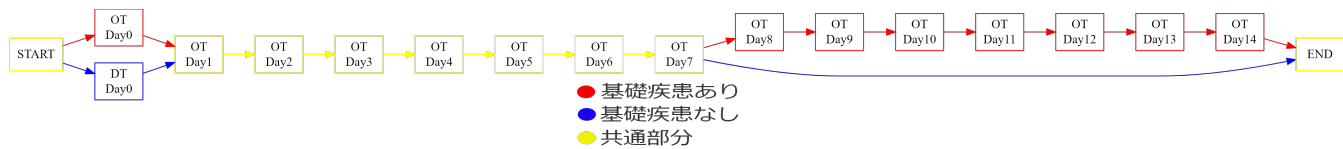


図 7: CmnSV 最大の治療パターン 1 (id=0)

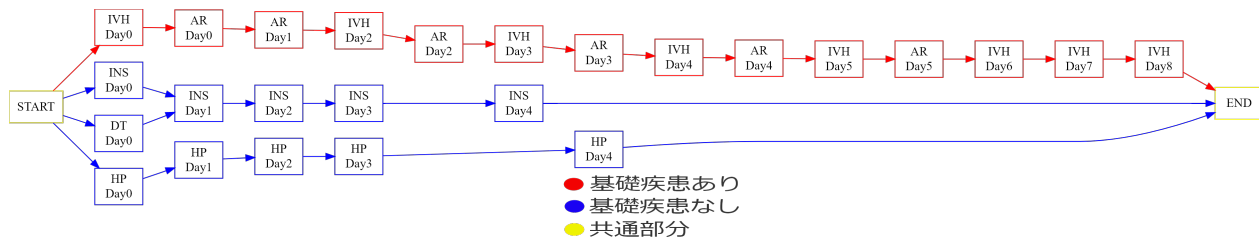


図 8: DiffSV 最大の治療パターン 1・SimSV 最小の治療パターン 1 (id=1)

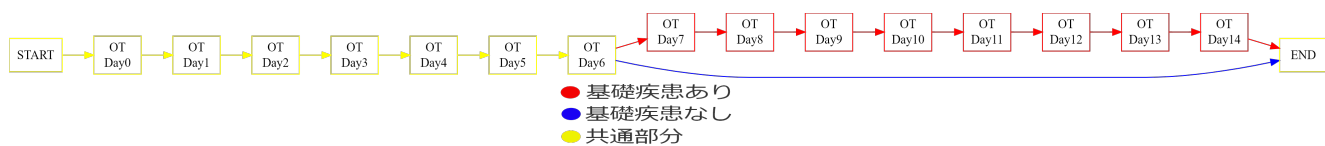


図 9: CmnSV 最大の治療パターン 2 (id=2)

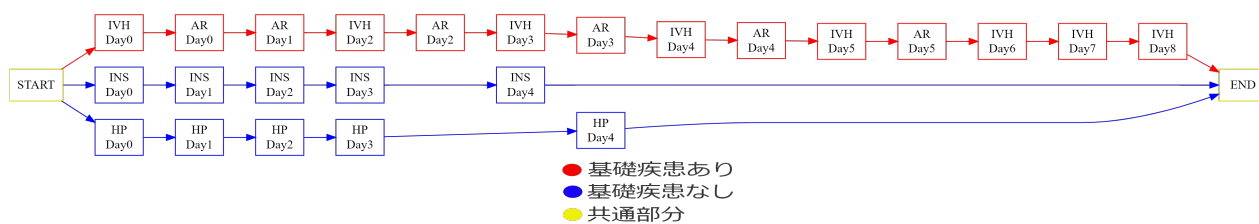


図 10: DiffSV 最大の治療パターン 2・SimSV 最小の治療パターン 2 (id=7)

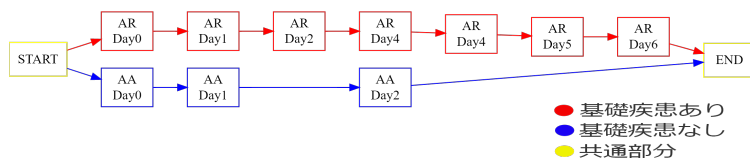


図 11: SimSV 最小の治療パターン 3 (id=41)

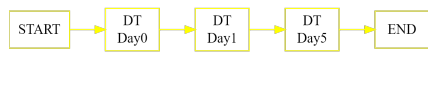


図 12: SimSV 最大の治療パターン 1 (id=44)

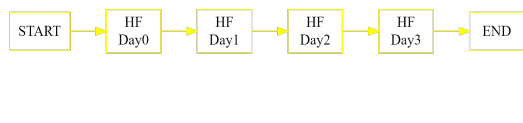


図 13: SimSV 最大の治療パターン 2 (id=47)