

甲状腺疾患の薬物療法における臨床判断支援システムの構築に向けて

森田 祐司[†] 吉川 正俊^{††} 濱崎 暁洋^{†††} 岡本 和也^{††††} 黒田 知宏^{††††}

[†] 京都大学大学院情報学研究科 〒606-8501 京都市左京区吉田本町

^{††} 京都大学大学院情報学研究科 〒606-8501 京都市左京区吉田本町

^{†††} 北野病院 糖尿病内分泌センター 〒530-8480 大阪市北区扇町 2-4-20

^{††††} 京都大学医学部附属病院 医療情報企画部 〒606-8507 京都市左京区吉田本町 36-1

E-mail: [†]morita@db.soc.i.kyoto-u.ac.jp, ^{††}yoshikawa@i.kyoto-u.ac.jp,

^{†††}{hamasaki,kazuya,tomo}@kuhp.kyoto-u.ac.jp

あらまし 近年、電子カルテの普及が進み膨大な医療データが蓄積され、データの二次利用が期待されている。膨大なデータから有用な知識を見つけ出す技術としてデータマイニングがあり、医療データを対象としたデータマイニングの適応研究は過去に多く行われている。医師は多くの患者の診療に際し、投薬間の関係性を把握した上で、処方薬の種類や用量を適切に選択するなどの臨床判断を行う必要があることから、計算機によって投薬間の関係性を提示する臨床判断支援システム (CDSS; Clinical Decision Support System) は有用であると考えられる。本研究では、データマイニング技術の一つである系列パターンマイニングを用いて電子カルテ上のデータから薬物療法行為の実施手順を分析し、投薬に関する頻出シーケンスを抽出することで、特定の患者に対する次の処方薬を予測・推薦し、薬物療法に関する医師の臨床判断支援を行うことを目的とする。

キーワード 系列パターンマイニング, 電子カルテ, 臨床判断支援

1. はじめに

近年、電子カルテの普及が進み、多くの医療機関で膨大な医療データが蓄積されている。蓄積された膨大な医療データから有用な知識を見つけ出すなどの電子カルテのさらなる活用が求められている。

膨大なデータから有用な知識を見つけ出す技術として、データマイニングがある。医療データへのデータマイニング技術の適応には大きな関心が寄せられており、多くの研究が行なわれている [1]。薬物療法において、医師は特定の疾患の患者の固有の情報や、投薬間の関係性に基いて、処方薬の種類や用量を適切に選択し、次の処方薬を決定する。このとき、薬物療法における頻出投薬パターンを提示することで臨床判断を支援できると考えられる。しかし、特定の疾患の薬物療法における頻出投薬パターンをマイニングする研究は少ない。処方薬の決定という臨床判断の際に、頻出投薬パターンを医師に提示する臨床判断支援システム (CDSS; Clinical Decision Support System) を用いることで、薬物療法における臨床判断の支援を行うことが可能であると考えられる。

本研究では、医師が全患者集合の中から、検体検査結果などの医学的な基準から、ある特定の患者集合を選択し、選択された患者に対して適切であると医師が判断した処方薬が投与されていると考え、投薬歴や検体検査結果を基準にグルーピングされた患者集合に対して、系列パターンマイニングを適用することによって、特定の患者に対する次の処方薬の決定に関する医師の臨床判断を再現することを最終目的とする。得られたマイニングの結果を医師に提示することで、薬物療法の判断支援を

行うことを目指す。本研究では甲状腺疾患の薬物療法における臨床判断支援を行うことを目的とする。

2. 背景

我が国における甲状腺疾患の頻度は高く、日本内科学会の調査によると、成人男性の 14.4% (7 人に 1 人)、成人女性で 24.7% (4 人に 1 人) が甲状腺疾患の患者と推定されており、年々増加傾向にある [2]。また、一般外来を受診する患者の約 13% に甲状腺疾患が見つかるなど、一般外来においても甲状腺疾患の占める割合は高い [3]。しかし、甲状腺疾患は非常に狭い専門領域であるため、甲状腺疾患の専門医は全国的に見ても非常に少ない。そのため、一般外来や健康診断などのプライマリ・ケアの場面において、甲状腺疾患を専門としていない医師が甲状腺疾患の患者に対して薬物療法を行う機会は少なくない。このような現状を踏まえると、本研究が目的とする次の処方薬の予測結果を医師に提供し、甲状腺疾患の薬物療法における臨床判断支援を行うことは有益であると考えられる。

技術的な観点からも甲状腺疾患には、注目すべき特徴が二つある。一つ目の特徴としては、薬物療法の際に使用する処方薬の種類が非常に少ないことが挙げられる。我が国で甲状腺疾患と並んで頻度の高い慢性疾患として糖尿病があるが、糖尿病における薬物療法で使用される処方薬の数は非常に多いため、膨大な系列パターンが生成されることが予想される。一方、甲状腺疾患を対象にした場合、処方薬の種類が非常に少ないため、処方薬の用量まで考慮したマイニングを行うことが可能となる。二つ目の特徴としては、処方薬の決定に用いられる要因について、少数の特定の検査値に基づく比重が高いことが挙げられ

る。糖尿病における薬物療法の場合、対象とする患者の検査値の他、年齢、性別、身長、体重などの数多くの医学的基準に基づいて処方薬を判断する。したがって、系列パターンマイニングによって医師の臨床意思決定を再現することを考えた際に、特定の患者集合を選択するために、どの医学的基準を用いるのかを決めることが非常に困難である。さらに、臨床では食習慣をはじめとした生活状況の把握に応じた判断が処方に加味されるが、その定式化が非常に難しく、電子カルテから必要な情報を収集するための手法の確立があらたに必要となる。一方、甲状腺疾患の薬物療法の場合、患者の来院時の症状や体格などが少なからず処方薬の判断に加味されるが、次の処方箋は主として限られた項目の検体検査結果の値に基づいて決定している。したがって、マイニングの際に考慮すべき患者固有の情報を検体検査結果の情報に絞ることが可能となる。

3. 関連研究

電子カルテデータの二次利用を扱う研究は盛んに行われている。Massoud らの研究 [4] では、糖尿病の診療ガイドラインで提供されていない状況における診療指針を提供する補完的情報を抽出するために、電子カルテ上の医師の処方に関するデータに、データマイニング技術の一つである C5.0 決定木の学習アルゴリズムを適用した。結果、実際の医師の処方から抽出される決定規則のいくつかは、後に公表された新しい診療ガイドラインに加えられていた情報と類似していることを示し、診療ガイドラインの不足を補うためにデータマイニング技術を用いる有効性を示した。この研究は、電子カルテ上のデータから決定規則を抽出することに着目し、診療ガイドラインの記述内容の情報を用いた分析を行っている。IBM の人工知能 Watson は、自然言語文処理能力を活用し、与えられた電子カルテの内容をもとに、医療分野での診断支援を行っている。将来的には、医師に医学論文などの医学的根拠に併せて治療方法の推薦を行うことを試みている [5]。しかし、この研究では、電子カルテ上のデータの中でも、特に自然言語によって記述された部分に着目しており、本研究が対象としているデータとは異なる。また、時刻情報が付与された医療データから、頻出パターンをマイニングを行う研究は Temporal Pattern Mining などと呼ばれ、多くの研究が行われている [6, 7]。Wright ら [8] の研究では、患者の処方歴から投薬間の時系列的な関係性を抽出するために系列パターンマイニングを用いた。この研究は、患者の投薬に関するデータから頻出パターンを抽出し、次に投与される処方薬を予測するという目的が本研究と類似しているが、頻出パターン抽出の際に考慮する患者の属性として、過去の処方歴以外の情報を考慮していない。Michele らの研究 [9] では、検査値からイベントを抽出し、各時刻とイベントの集合を対応付け、各イベント集合ごとの間隔に幅を持たせた頻出パターンの抽出を行っている。この研究では、各イベント集合ごとの間隔に幅を持たせた頻出パターンの定式化を主眼においている。また、平野らの研究 [10] では、医療行為を支援するために、電子カルテに蓄積されたオーダログデータを解析し、全体を通じて典型的と考えられる診療プロセスを抽出する手法を提案した。平野

表 1 系列データベース

シーケンス ID	トランザクションタイム	アイテム
患者 1	10	チアマゾール 5mg
患者 1	15	チアマゾール 10mg
患者 1	20	プロピルチオウラシル 50mg
患者 2	15	チアマゾール 10mg
患者 2	20	チアマゾール 5mg
患者 3	15	レボチロキシンナトリウム 20mg
患者 3	20	レボチロキシンナトリウム 30mg
患者 3	25	チアマゾール 5mg

表 2 表 1 から生成されるシーケンス ID とトランザクション時刻をまとめたリストの一部

チアマゾール 5mg		チアマゾール 10mg	
シーケンス ID	トランザクションタイム	シーケンス ID	トランザクションタイム		
患者 1	10	患者 1	15		
患者 2	20	患者 2	15		
患者 3	25				

らの手法では、オーダ列をある特定の期間毎にフェーズに分割し、すべての事例の情報をもとに各フェーズにおいて各オーダ項目がどの程度の頻度で適用されているのか、また、隣接するフェーズ間で、どのオーダ組が接続関係として多用されているのかを頻度ベースで集計し、クリニカルパスの候補を生成した。この研究は、電子カルテデータを解析することでクリニカルパスの作成を支援することを目的としており、電子カルテデータの中でもアクセスログデータを分析対象にしている。

4. 準備

4.1 系列パターンマイニング

本研究では、系列パターンマイニングのアルゴリズムの一つである Zaki の考案した SPADE (Sequential PAttern Discovery using Equivalence classes) アルゴリズムを使用する [11]。SPADE は入力として最小支持度 (最小サポート) と表 1 のような系列データベースを受け取る。系列データベースは複数の系列データから構成されており、各系列データはシーケンス ID、トランザクション時刻、アイテムで構成される。すべてのアイテムを $\Sigma = \{i_1, i_2, \dots, i_n\}$ と表し、各アイテムは処方薬を表す。このとき、アイテムの時系列 S をシーケンスと呼び、 $S = \langle a_1, a_2, \dots, a_m \rangle \mid a_i \in \Sigma$ と表す。例えば、 $S = \langle \text{チアマゾール}, \text{プロピルチオウラシル} \rangle$ と表す。シーケンスのデータベースをトランザクションと呼び、 $T = \{S_1, S_2, \dots, S_n\}$ と表す。また、シーケンス S 中のすべてのアイテムが、別のシーケンス S' 中に存在し、さらにその順序が保持されている場合、シーケンス S をシーケンス S' の部分シーケンスであると定義し、 $S \subseteq S'$ と表す。例えば、シーケンス $S = \langle \text{チアマゾール}, \text{プロピルチオウラシル} \rangle$ と、シーケンス $S' = \langle \text{チアマゾール}, \text{プロピルチオウラシル}, \text{レボチロキシンナトリウム} \rangle$ を考えると、シーケンス S は、シーケンス S' のサブシーケンスと定義される。ここで、トランザクション T を構成するすべてのシーケ

表 3 甲状腺疾患の処方薬の一般名と商品名

商品名	一般名(成分名)
メルカゾール	チアマゾール
プロバゾール, チウラジール	プロピルチオウラシル
チラーヂン S, チラーヂン S 散	レボチロキシンナトリウム

ンスのうち、部分シーケンス の出現回数を、 の支持度 (サポート値) と定義し、支持度は系列パターンの出現頻度を表す値である。

SPADE は系列データベースから表 2 のようにアイテムごとに、シーケンス ID とトランザクション時刻をまとめたリストを作成し、そのリスト同士を結合させ、最小支持度未満の系列パターン除外しながら、新たな系列パターンを検出していく。例えば、表 1 において、<チアマゾール 10mg> というシーケンスはシーケンス ID = 患者 1, シーケンス ID = 患者 2 で見られるため、このシーケンスの支持度は $2/3 = 0.66$ となる。また、ここでは、一回の処方行為をアイテムとして扱った。

4.2 処方薬に関する定義

処方薬の名称には、薬の有効成分を表す一般名 (成分名) と、製薬会社が独自に付けた商品名が存在する。我々が扱ったデータセットには、処方薬は商品名で記載されていた。しかし、商品名レベルの粒度で系列パターンマイニングを行った場合、処方薬の組み合わせの数が非常に多くなり、膨大な数のパターンが生成されてしまう問題がある。また、処方薬の用量も含めてマイニングすることを考えると、さらに多くパターンが生成されてしまう。そこで、処方薬の粒度としては一般名レベルとし、これに用量を付与したデータをマイニングの対象とした。処方薬の粒度は、厚生労働省の医薬品名称調査会が定めた医薬品一般的名称 (JAN: Japanese Accepted Name) [12] に基づいて決定した。表 3 に本研究で扱う甲状腺疾患の処方薬の商品名と一般名の対応関係表を示す。

4.3 分析対象データ

我々が分析対象としたデータは、京都大学医学部附属病院で、甲状腺疾患の処方薬を投与された患者の投薬データおよび検体検査結果データである。

投薬データに関して、本研究では甲状腺疾患の非専門医に対して、プライマリ・ケアの場面での薬物療法における臨床判断支援を行うことを目的としているため、以下の三つの条件を全て満たす投薬データを分析対象とした。

- (1) 甲状腺疾患の専門医が行った投薬データ
- (2) 外来患者に対する投薬データ
- (3) 妊娠経験のない患者に対する投薬データ

(1) に関して、分析対象とするデータを甲状腺疾患の専門医が行った処方限定することで、分析結果は専門医の薬物療法における知識を表したものと考えられる、この分析結果を非専門医に提供することで臨床判断の支援が可能であると考えられる。ここでの専門医の提議として、京都大学医学部附属病院における甲状腺疾患を専門とする診療科である内分泌・代謝科、内分泌・代謝内科、糖尿病・内分泌・栄養内科のいずれかの診療科

に所属しており、かつ、甲状腺疾患の処方を 100 回以上行った経験のある医師を専門医と定義した。(2) に関して、プライマリ・ケアの場面での臨床判断支援を想定しているため、過去に 1 回以上甲状腺疾患で入院経験のある患者に対する処方分析対象から除き、外来患者に対する処方のみを分析対象とした。(3) に関して、甲状腺疾患は患者の妊娠経験によって処方の仕方が異なるという特徴がある。妊娠患者に対する薬物療法はプライマリ・ケアの範囲外であるため、過去に 1 回以上妊娠経験のある患者に対する処方分析対象から除き、妊娠経験がない患者に対する処方のみを分析対象とした。

また、検体検査結果データに関して、甲状腺疾患における薬物療法では限られた項目の検体検査結果を考慮して次の処方薬を選択するが、本研究では考慮すべき検体検査として以下の三つの検体検査の結果 (数値) を用いた。

- ・ FT3: 血液中に存在する甲状腺ホルモンの遊離トリヨードサイロニンの量を調べる検査
- ・ FT4: 血液中に存在する甲状腺ホルモンの遊離サイロキシンの量を調べる検査
- ・ TSH: 血液中の甲状腺刺激ホルモンの量を調べる検査

データマイニングの際には、上記の検体検査を患者固有の情報として用い、各検査値を以下の 3 値に離散化した。

- 正常値
- 正常値を下回る異常値
- 正常値を上回る異常値

各患者を検体検査結果に基づいてグルーピングし、グルーピング後の各患者集合に対して系列パターンマイニングを適用し、頻出パターンを抽出する。

5. 提案手法

5.1 医師の多様性を考慮した頻出パターンの抽出

我々は、多くの医師が偏りなく使用している処方箋は信頼性の高い処方である、という仮定のもと、医師の多様性という概念を考慮した頻出パターンの抽出手法を提案する。具体的には、系列パターンマイニングによって抽出された各頻出パターンの支持度と医師の多様性の値から、以下の式に基づいて算出される新たなスコアである DF-score に基づいて各頻出パターンをランキングし、頻出パターンの抽出を行う。

$$DF\text{-score}(\) = \text{Support}(\) \times D\text{-rank}(\)$$

ここで、 をシーケンスとすると、 $\text{Support}(\)$ は の支持度の値を示し、 $D\text{-rank}(\)$ は についての医師の多様性の値を示す。すなわち、 $DF\text{-score}(\)$ は患者ベースの頻度に基づく情報と、医師ベースの多様性に基づく情報を組み合わせた指標である。

5.2 医師の多様性

ここでの医師の多様性とは、

- ・ ある投薬パターンを行った医師の割合: D_{rate}
 - ・ 各医師の投薬パターンの使用頻度の偏り: D_{disp}
- を表した指標であり、以下の式に基づいて算出される。

$$D\text{-rank}(\) = D_{rate}(\) \times D_{disp}(\)$$

$D_{rate}(\)$ は、シーケンス を行った医師の割合を表すもので、次のように定義される。

$$D_{rate}(\) = \frac{\text{count}(D(\))}{|D|}$$

$\text{count}(D(\))$ は、シーケンス を行った医師の数を示し、 $|D|$ は全医師の総数を示す。すなわち、 D_{rate} は、すべての医師に対する、シーケンス を行ったことがある医師の割合を表す。

また、 $D_{disp}(\)$ は、各医師のシーケンス の使用頻度の偏りを表すもので、次のように定義される。

$$D_{disp}(\) = \frac{1}{V_{UB}(\)}$$

$V_{UB}(\)$ は、各医師のシーケンス の使用頻度の偏りを表す不偏分散として、次のように定義される。

$$V_{UB}(\) = \frac{1}{n-1} \sum_{i=1}^n (d_i - \bar{d})^2$$

すべての医師を $\sum = \{D_1, D_2, \dots, D_n\}$ と表し、医師 D_i のシーケンス を処方した患者数を d_i とし、 \bar{d} は、シーケンス を使用したすべての医師が処方した患者数の平均値を示す。すなわち、 D_{disp} は、各医師の投薬パターンの使用頻度の偏りを表す不偏分散の逆数をとった値である。

5.3 分析対象データ取得から多様性を考慮した頻出パターン抽出までの全体プロセス

図 1 に多様性を考慮した頻出パターンの抽出手順の概略図を示す。手順は以下の通りである。

(1) 甲状腺疾患の処方薬の商品名と一般名の対応関係を示した対応表を作成する。

(2) 作成した対応表をもとに、電子カルテから分析対象となる甲状腺疾患の処方薬を投与された患者の投薬データおよび検体検査結果データを抽出する。

(3) 分析の前処理として、検体検査結果に基づいて各患者をグルーピングした後、電子カルテから抽出したデータを系列パターンマイニングを行うためにトランザクション形式に整形する。

(4) 整形したデータセットについて系列パターンマイニングを実行し、頻出パターンを抽出する。このとき、最小サポート値は $1.0e-10$ に設定した。

(5) 抽出した各頻出パターンの医師の多様性 ($D\text{-rank}$) および、新たなスコア ($DF\text{-score}$) を計算し、 $DF\text{-score}$ に基づいて多様性を考慮した頻出パターンを抽出する。

6. 実 験

6.1 データセット

本研究では、京都大学医学部附属病院の電子カルテシステムに蓄積された 2000 年 9 月 20 日から 2015 年 12 月 22 日までの匿名化されたオーダデータセットを使用した。このデータは、京都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 KING (Kyoto university hospital INFORMATION GALAXY) によって取得されたものである。また、個人情報保護の点から、

表 4 $Suppot$ に基づいてランキングした頻出パターン

2つのアイテムから構成される頻出パターン上位10個 シーケンス_用量[mg]	Support	DF-score
チアマゾール_10, チアマゾール_5	0.2442	51.6
チアマゾール_5, チアマゾール_10	0.1257	36.7
チアマゾール_15, チアマゾール_10	0.1039	74.4
チアマゾール_20, チアマゾール_15	0.0550	40.3
プロピルチオウラシル_100, プロピルチオウラシル_50	0.0540	69.6
チアマゾール_10, チアマゾール_15	0.0467	85.5
プロピルチオウラシル_150, プロピルチオウラシル_100	0.0384	59.4
チアマゾール_20, チアマゾール_10	0.0322	9.8
チアマゾール_15, プロピルチオウラシル_5	0.0280	51.7
チアマゾール_30, チアマゾール_20	0.0280	38.1

表 5 $DF\text{-score}$ に基づいてランキングした頻出パターン

2つのアイテムから構成される頻出パターン上位10個 シーケンス_用量[mg]	Support	DF-score
チアマゾール_30, チアマゾール_10	0.0051	732.2
プロピルチオウラシル_250, プロピルチオウラシル_200	0.0031	279.3
プロピルチオウラシル_150, プロピルチオウラシル_300	0.0072	154.5
プロピルチオウラシル_300, プロピルチオウラシル_250	0.0062	115.0
チアマゾール_5, プロピルチオウラシル_50	0.0218	109.0
チアマゾール_5, チアマゾール_15	0.0218	97.7
チアマゾール_10, チアマゾール_15	0.0467	85.5
チアマゾール_15, チアマゾール_20	0.0249	85.4
チアマゾール_5, プロピルチオウラシル_100	0.0031	80.8
プロピルチオウラシル_200, プロピルチオウラシル_300	0.0093	77.7

このデータは患者の名前を含んでいない。電子カルテシステムで扱われる各々の電子カルテが記載する内容は、患者の名前や年齢といった基本的な情報から、経過記録、各種検体検査結果、医療画像など様々であるが、本研究では少なくとも一つの甲状腺疾患の処方薬を投与された患者 1420 人の投薬データと検体検査結果データを用いて分析を行った。

6.2 結 果

各検体検査結果が全て正常値の患者集合に対して実行したマイニングの結果として抽出された頻出パターンを表 4 と表 5 に示す。表 4 にシーケンスの支持度 ($Suppot$) の値に基づいてランキングした 2 つのアイテムから構成されるシーケンス上位 10 個を示し、表 5 にシーケンスの支持度 ($Suppot$) の値と、医師の多様性 ($D\text{-rank}$) の値から算出されるスコア ($DF\text{-score}$) の値に基づいてランキングした 2 つのアイテムから構成されるシーケンス上位 10 個を示す。

すなわち、表 4 に示すシーケンスは患者ベースの頻度のみを考慮した場合にランキング上位に出現する頻出パターンであり、一方の、表 5 に示すシーケンスは患者ベースの頻度に加えて医師の多様性を考慮した場合にランキング上位に出現する頻出パターンである。表 4 と表 5 において、一列目がシーケンス名、二列目がシーケンスの支持度の値、三列目がシーケンスの多様性を考慮したスコアの値を示す。

7. 考 察

我々は、多くの医師が偏りなく行っている処方方は信頼性の高い処方である、という仮定のもと、医師の多様性を考慮した頻

①対応表作成 (Microsoft Excel)

商品名	一般名
メルカゾール錠 5mg	チアマゾール
プロパゾール錠 50mg	プロピルチオウラシル
チウラゾール錠 50mg	プロピルチオウラシル
チラーヂン S 錠 12.5 μg	レボチロキシナトリウム
チラーヂン S 錠 25 μg	レボチロキシナトリウム
⋮	⋮

③患者のグルーピングおよびデータ整形 (R 言語)

シーケンス ID	イベント ID	アイテム数	アイテム_用量[mg]
1	1	1	チアマゾール_5
1	2	1	チアマゾール_10
1	3	1	プロピルチオウラシル_50
2	1	1	チアマゾール_10
2	2	1	チアマゾール_5
⋮	⋮	⋮	⋮

④D-rank および DF-score を計算

シーケンス	Support	D-rank	DF-score
プロピルチオウラシル_50	0.1216	35.1	4.2
プロピルチオウラシル_100	0.0919	90.8	8.3
チアマゾール_20	0.1288	109.7	14.1
レボチロキシナトリウム_60,レボチロキシナトリウム_50	0.0017	83.2	0.1
チアマゾール_5, チアマゾール_7.5	0.0031	112.9	0.3
⋮	⋮	⋮	⋮

②データ抽出

患者 ID	タイムスタンプ	アイテム_用量[mg]
123	2008/08/16	チアマゾール_5
123	2008/11/11	チアマゾール_10
123	2009/01/10	プロピルチオウラシル_50
456	2004/05/21	チアマゾール_10
456	2004/06/16	チアマゾール_5
⋮	⋮	⋮

④系列パターンマイニングを実行

(R 言語の arulesSequences パッケージ上の SPADE アルゴリズム)

シーケンス	Support
プロピルチオウラシル_50	0.1097
プロピルチオウラシル_100	0.0979
チアマゾール_20	0.1288
レボチロキシナトリウム_60,レボチロキシナトリウム_50	0.0007
チアマゾール_5, チアマゾール_7.5	0.0031
⋮	⋮

図 1 多様性を考慮した頻出パターンの抽出手順

出パターンの抽出手法を提案した。実験結果として、表 4 と表 5 を比較すると、Support が最大値であるパターンと、DF-score が最大値であるパターンは、異なるパターンとなっており、全体として見てもランキングが変動していることが分かる。パターンの支持度に加えて医師の多様性を考慮することで、抽出される頻度パターンが変化することが示唆された。

今回は、実際の電子カルテ上のデータから、頻出パターンを抽出する実験のみに留まり、抽出されたパターンの客観的な有効性に関する評価は行っていない。今後の課題として、多様性を考慮することによって、マイニングの結果が具体的な応用として、次の処方薬の予測においてどれほどの有意差を生じるか評価を行い、より適切な臨床判断支援を行うために医師の多様性を考慮することの有効性を検証する必要がある。具体的な評価方法としては、実データを用いた交差検証による実験や、日本臨床検査医学会の甲状腺疾患に関する診療ガイドライン [13] との比較、医療関係者からの反応をもとにした評価などを考えている。

また、医師の多様性に加えて、根拠という概念も臨床判断支援に有用であると考えられる。ここでの根拠とは、次の処方薬の候補としてマイニングされた推薦結果について、なぜこの処方薬を投与するのか、つまり、推薦結果の臨床的理由に関する情報である。近年、根拠に基づく医療 (EBM; Evidence Based Medicine) に大きな関心が集まっているように、臨床判断にお

いて根拠は非常に重要なものである [14]。次の処方薬の候補と併せて、根拠を医師に提示することで、より良い臨床判断支援を行うことができると考えている。根拠に関して、具体的には検体検査結果の情報をを用いる計画である。推薦された処方薬を投与することで、その患者の検体検査結果の値がどのように変動するのかという情報を根拠として医師に提示することを考えている。

8. ま と め

本研究では、電子カルテ上のデータから、プライマリ・ケアの場面での甲状腺疾患の薬物療法行為を系列パターンマイニングを用いて分析し、投薬に関する頻出シーケンスを抽出することで、特定の患者に対する次の処方薬をマイニングする。マイニングの結果を医師に提示することで、甲状腺疾患における薬物療法の臨床判断支援を行うことを目標としている。その基礎的検討として、患者の投薬歴データおよび検体検査結果データに着目し、医師の多様性という概念を導入した頻出シーケンスの抽出を行った。実験では、京都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上の匿名化された実データを分析対象とした。2000年9月20日から2015年12月22日までの甲状腺疾患の患者1420人を対象として、提案手法を用いて頻出シーケンスの抽出を行ったところ、医師の多様性を考慮した場合と、考慮しない場合とでは、異なるパターンが抽出されることが分かった。今回の結果から、多様性を考慮して抽出された

頻出パターンを用いて、次の処方薬の予測することで、より適切な臨床判断支援が可能となることが示唆された。

文 献

- [1] Jonathan C. Prather, David F. Lobach, Linda K. Goodwin, Joseph W. Hales, Marvin L. Hage, and W. Edward Hammond. Medical data mining: Knowledge discovery in a clinical data warehouse. In *In 1997 Annual Conference of the American Medical Informatics Association*, pp. 101–105, 1997.
- [2] 森昌朋. 甲状腺疾患の診断と治療 現状と展望 . 日本内科学会雑誌, Vol. 99, No. 4, pp. 683–685, 2010.
- [3] 浜田昇. 一般外来で見逃してはいけない甲状腺疾患の頻度. 日本医事新報, Vol. 3740, pp. 22–26, 1995.
- [4] Massoud Toussi, Jean-Baptiste Lamy, Philippe Le Toumelin, and Alain Venot. Using data mining techniques to explore physicians' therapeutic decisions when clinical guidelines do not provide recommendations: methods and example for type 2 diabetes. *BMC medical informatics and decision making*, Vol. 9, No. 1, p. 28, 2009.
- [5] David A Ferrucci, Anthony Levas, Sugato Bagchi, David Gondek, and Erik T Mueller. Watson: beyond jeopardy! *Artif. Intell.*, Vol. 199, pp. 93–105, 2013.
- [6] Robert Moskovitch and Yuval Shahar. Medical temporal-knowledge discovery via temporal abstraction. In *AMIA annual symposium proceedings*, Vol. 2009, p. 452. American Medical Informatics Association, 2009.
- [7] David Gotz, Fei Wang, and Adam Perer. A methodology for interactive mining and visual analysis of clinical event patterns using electronic health record data. *Journal of biomedical informatics*, Vol. 48, pp. 148–159, 2014.
- [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*, Vol. 53, pp. 73–80, 2015.
- [9] Michele Berlingerio, Francesco Bonchi, Fosca Giannotti, and Franco Turini. Mining clinical data with a temporal dimension: a case study. In *Bioinformatics and Biomedicine, 2007. BIBM 2007. IEEE International Conference on*, pp. 429–436. IEEE, 2007.
- [10] Shoji Hirano and Shusaku Tsumoto. Clustering of order sequences based on the typicalness index for finding clinical pathway candidates. *2013 IEEE 13th International Conference on Data Mining Workshops*, Vol. 0, pp. 206–210, 2013.
- [11] Mohammed J Zaki. Spade: An efficient algorithm for mining frequent sequences. *Machine learning*, Vol. 42, No. 1-2, pp. 31–60, 2001.
- [12] 日本医薬品一般名称データベース. <http://jpdb.nihs.go.jp/jan/>.
- [13] 池田斉. 甲状腺機能亢進症, 甲状腺機能低下症. 診断群別臨床検査のガイドライン 医療の標準化に向けて, pp. 107–110, 2003.
- [14] 福井次矢. Evidence-based medicine の手順と意義. 日本内科学会雑誌, Vol. 87, No. 10, pp. 2122–2134, 1998.