3O1-4in

診療所電子カルテデータを用いた診療プロセス可視化の試み

Investigation of clinical process visualization using EMR data in clinics

中嶋 航大 *1 Kodai Nakajima　田村 哲嗣 *1 Satoshi Tamura　速水 悟 *1 Satoru Hayamizu　山本 けい子 *2 Keiko Yamamoto　一宮 尚志 *3 Takashi Ichinomiya
紀ノ定 保臣 *3 Yasutomi Kinosada

*1岐阜大学 工学部 応用情報学科
Department of Information Science, Faculty of Engineering, Gifu University

*2函館工業高等専門学校
National Institute of Technology, Hakodate College

*3岐阜大学大学院 医学系研究科 医療情報学分野
Biomedical Informatics, Gifu University Graduate School of Medicine

Recently, as electronic medical record systems have been widely spread, medical data records have been accumulated so much and attracted researchers in the data mining fields. In this paper, we propose a mining method to extract a clinical process from electronic medical records in a clinic. Data records are classified into several clusters according to term information, subsequently these clusters are associated with each other based on time-series information. We conducted subjective evaluations by investigating keywords obtained from extracted processes, as well as objective evaluations by using additional information which is not employed in the process extraction.

1. 序論

近年，電子カルテシステムが広く普及し，大きな病院だけでなく中小規模の診療所でも電子カルテを用いた診療記録が一般的になりつつある．また，電子カルテシステムの普及とともに，医療記録が電子的に記録されることにより，日々膨大な量の医療データが蓄積されるようになった．現在，こうした医療データをデータマイニング技術を用いて医療の質の向上につながるような知見を発見しようという試みが広がっている．

[三浦 2010] は退院時サマリの自由記述欄に記入されている副作用の記述に着目し，機械学習手法によって副作用情報を自動的に発見・抽出する手法を提案した．[荒牧 2009] は退院時サマリデータを対象に，患者の情報を抽出して表形式に可視化するシステムを提案した．しかしこれらの研究は医療テキストからの情報抽出や，退院時サマリの情報を表形式に構造化するだけにとどまっている．また，退院時サマリを元にした研究は多く存在するが，診療所の電子カルテを用いたり，時系列情報を用いたりする研究は少ない．こうした背景を元に本研究では，診療所電子カルテのテキストデータを用いて対象データを状態ごとに分類し，時系列情報を元に患者の状態遷移を可視化するプロセスマイニングを行った．

2. 診療所電子カルテデータ

本研究では 1 つの診療所の電子カルテデータを用いた．電子カルテに記載されていた項目は「診療日・患者 ID・生年月日・年齢・性別・受診形態（時間内，時間外など）・初再診（初診，再診，その他）・病名・既往歴・所見」の 9 項目である．表 1 に診療所から用いた電子カルテデータの詳細を示す．本稿で用いたデータは個人情報保護の観点から，従前開発した匿名化処理を施し，固有名詞を匿名化している．電子カルテデータは一回の診療につき一つ作成され，ある患者が複数回にわたって受診を行った場合，同一の患者 ID をもって電子カルテが作成される．

表 1: 診療所電子カルテデータ詳細

診療データ数	3273 件
収集期間	2002/8/17　～　2009/11/2
受診人数	270 名
性別	男性 124 名，女性 146 名

この診療所では SOAP 形式での電子カルテ記録を行っており，医療記録が項目ごとに分けて記録されている．SOAP 形式の電子カルテとは診療記録を Subject(主観情報)，Objective(客観的情報) Assessment(評価)，Plan(計画) の 4 つの項目に分類しながら記録する電子カルテ記入手法の一つである．Subject(主観情報) は医師が患者からヒアリングした主訴情報が記述され，Objective(客観的情報) には診察，検査から得られた情報が記載されている．Assessment(評価) には患者がどんな病気であるのか，また進行中の治療が順調であるかどうかなどが医師の裁量で記入されており，Plan(計画) は受診の際行った治療や説明，指導，今後患者に施す予定の治療内容を記述する項目である．診療データの一例を図 1 に示す．なお個人情報保護のため患者 ID と年齢は除外した．

診療日	2009/2/26	生年月日	1997/4/1
性別	女	受診形態	時間内
初再診	再診	病名・既往歴	アレルギー性鼻炎 溶連菌性扁桃炎の疑い
所見	【S】主訴(全身) 発熱 2日後解熱した 主訴(胸部) 湿性咳嗽 主訴(鼻) 鼻汁 水様透明から白色で粘っこい 水様透明でぽたぽた落ちる 【O】所見(肺) 肺胞呼吸音 【A】インフルエンザ 経過は良好 アレルギー性疾患 鼻炎 鼻汁型 疑い。 【P】(検査、治療) 投薬追加 右記投薬 経過観察 【P】(指導) 年月日21/02/28より 登校可能。 【P】(指導) 5分間以上診察した。 (院外・変更不可)処方,アレグラ錠30mg 2錠, 分2朝・夕食後服用 5日分 再診,外来管理加算		

図 1: 診療データの例

3. 医療用語辞書

本研究ではプロセスマイニングを行う際，医療用語を元に単語ベクトルを作成し，その単語ベクトルをもってクラスタリ

連絡先: 〒 501-1193，国立大学法人岐阜大学工学部　速水・田村研究室

ング処理を行う．そこで，使用する医療用語を JAPIC データベース及び ICD10 より薬剤名リスト，病名リストを作成した．さらに診療データに頻出する用語のうち薬剤名．病名に含まれない「Hb1Ac」など 42 語を重要語リストとして追加した．薬剤名リスト，病名リスト，重要語リスト内の一例，詳細を表 2 に示す．なお，医療用語リストには単語だけでなく「胃炎の疑い」などの句単位での登録もされている．

表 2: 医療用語辞書の詳細

リスト名	リスト内用語の一例	総用語数
薬剤名リスト	アイケア, アクトス, アザニン	11244 語
病名リスト	胃炎の疑い, 胃けいれん, 胃痛	62546 語
重要語リスト	インスリン, コレステロール, レントゲン,HbA1c,ECG.BMI	42 語

4. プロセスマイニング手法

本章では本研究で用いたプロセスマイニング手法に関して，段階的に説明する．本手法の概要図を図 2 に示す．

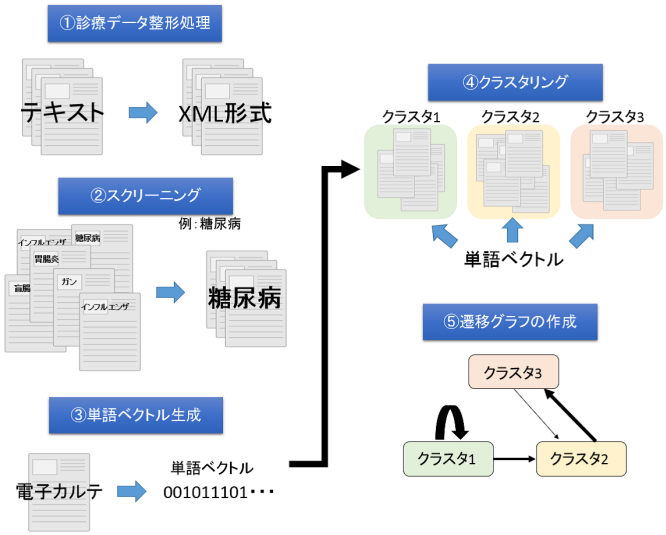

図 2: プロセスマイニング手法概要図

4.1 診療データの整形処理

本研究で使用した診療所の電子カルテデータは平文で構成されていた．そこで扱いを簡単にするために XML 形式のデータへ自動的に変換し，各項目ごとに構造化を行う．

4.2 スクリーニング

続いてデータのスクリーニングを行う．本研究では症例ごとにプロセスマイニングを行うため，診療データから特定の症例の診療データのみ抽出する．対象とする症例の単語を含む診療データを抽出する．これ以降の処理は，スクリーニングによって抽出された診療データのみを対象として行う．

4.3 単語ベクトルの抽出

次の処理は，診療データから単語ベクトルを生成する処理である．単語ベクトルは診療データ毎に生成する．各診療データの病名・既往歴欄・所見欄のテキストから医療用語を元に，単語ベクトルを生成する．この際，医療用語の出現回数は考慮せず，出現したか否かの 0,1 をもって 2 値ベクトルを生成した．前項のスクリーニングによって抽出された対象データ群に，一回以上出現する総単語数を N とすると，診療データ E_i の単語ベクトル v_i は j 次元目を $v_i(j)$ とする式 (1) で表される．なお w_j は j 番目の医療用語を表現している $(1 \leq j \leq N)$.

$$v_i(j) = \begin{cases} 1 & (\text{if } w_j \text{ in } E_i) \\ 0 & (otherwise) \end{cases} \quad (1)$$

4.4 クラスタリング

各診療データから抽出した単語ベクトルを元に，クラスタリングを行う．まず単語ベクトル v_i のみを要素とする要素数 1 の初期クラスタを作成する．ここで i 番目のクラスタの重心ベクトルを g_i とすると，式 (2) を満たすようなクラスタ C_p 及び C_q を求め，それらを統合して新しいクラスタとする．クラスタ間の距離尺度にはハミング距離を用いる．

$$(p, q) = argmin_{(p,q)}||g_p - g_q|| \quad (2)$$

この操作をクラスタ数が 1 になるまで繰り返す．その際，クラスタリング過程を木構造にすることで，生成した木構造を用いて，任意のクラスタを生成することができる．例えば L 個のクラスタに分割するときは，葉の数が L となるまで距離に応じて木構造を根からたどっていけばよい．

4.5 クラスタ間遷移の抽出

最後に，診療所を複数回受診した患者データを用いてクラスタ間遷移を計算し，遷移グラフを作成する．具体的には同じ患者 ID の複数の診療データを用いて 4.4 節のクラスタリングで得られたクラスタ間の遷移回数を求め，クラスタ間の遷移を矢印で表すことにする．この時矢印線の太さは，そのクラスタからの遷移頻度の総量に対する相対的な大きさを示している．例えばクラスタ C_x からの総遷移回数が 10 回，C_y からの総遷移回数が 50 回存在したとする．その際 C_x から C_y への遷移が 2 回，C_y から C_x への遷移が 10 回だったとすると，どちらも総遷移回数における割合が 20%となるので，2 つの遷移を表す矢印線は同じ太さで表される．遷移頻度統計に対する割合が 10%を下回る遷移は図に示していない．

5. 評価指標

プロセスマイニングを行った結果得られた遷移グラフまたは各クラスタの評価を，以下の 4 つの評価指標と所見欄の内容を用いて行った．

- 初再診率
- 患者状態ラベル (インフルエンザ・風邪のみ)
- 状態キーワード
- 遷移に基づくキーワード

5.1 患者状態ラベル

今回実験を行った 2 症例のうちインフルエンザ・風邪の症例に関して，対象データ 243 件に対して人手で 3 種類の患者状態ラベルを振り，各クラスタの評価に用いた．各ラベルの詳細と件数を表 3 に示す．括弧内の数字はデータ数全体に対する割合を示している．またインフルエンザ・風邪において罹患中と見られる状態のうち，インフルエンザと診断されたデータは 34 件存在した．

表 3: 患者状態ラベル詳細

状態	件数
罹患中とみられるもの	98 件 (40%)
症状が回復したとみられるもの	13 件 (6%)
別の症例である，また判断がつかない	132 件 (54%)

5.2 初再診率

プロセスマイニングに用いなかった客観的データとして，各診療データに含まれる初再診情報を用いた．電子カルテには診察を受けた患者が初診であるか再診であるかが，初診，再診，初再診なしの 3 種類に分けて記録されている．各クラスタの初再診率を調査することで，適切に診療データの時系列情報を扱えているか評価を行う．各クラスタの 3 種類の初再診情報を集計し評価に用いた．

5.3 電子カルテの構造に基づくキーワード抽出

本研究では，各クラスタがどんな患者の状態を表しているのかを調べるために，SOAP 形式で入力された電子カルテの構造に着目し，Subject(主観情報)・Assessment(評価)，Plan(計画) の項目からキーワード抽出を行った．抽出されたキーワードを元に各クラスタの状態を主観評価した．

SOAP 形式電子カルテの 4 項目の中でも，Subject(主観情報)，Assessment(評価) の項目に着目して，キーワード抽出を行ったのはこの 2 項は電子カルテ内の情報の中でも患者の病態をよく表していると考えたためである．また診療プロセスを評価するには各クラスタの評価だけではなく，状態遷移とともにどういった診療を行っているかを明らかにする必要がある．そこで，状態キーワード抽出とは異なり，クラスタ間遷移が起こった際の Plan(計画) からも文章を収集し，遷移に基づくキーワードの抽出を行った．

まず状態キーワードは電子カルテの Subject(主観情報)，Assessment(評価) の項目からクラスタごとに文章を抽出し，遷移に基づくキーワードはクラスタ間で同じ遷移している診療データの Plan(計画) から文章を抽出する．次に TF*IDF 法を用い，抽出した文章から 2 文字以上の単語に関してキーワード抽出を行い，最後に抽出したキーワードの TF*IDF 値を元に上位 10 個を各クラスタ，各遷移のキーワードととし評価に用いた．IDF は全診療データの所見欄からテキストを抽出し，計算を行った．

6. 実験と考察

6.1 結果

急性疾患からインフルエンザ・風邪，慢性疾患から糖尿病の 2 症例についてプロセスマイニングを行い，得られた各クラスタや遷移グラフを第 5 章の指標を用いて評価を行った．各症例の実験条件を表 4 に示す．クラスタ数に関しては，クラスタ数を増やすほどクラスタ間の遷移回数が分散し，遷移グラフを作成した際，独立したクラスタが生成されてしまう傾向にあった．本研究では診療の大きな流れを知ることを目的とするため，クラスタ間遷移が分散しすぎないよう，経験的にクラスタ数を 4 とした．また，遷移グラフの楕円の大きさはそのクラスタに含まれる診療データ数を反映している．得られた遷移グラフを図 3，図 4 に示す．

表 4: 実験条件

対象症例	インフルエンザ・風邪	糖尿病
データ件数	243 件	306 件
単語ベクトル次元数	113 次元	160 次元
クラスタ数	4	4

6.2 初再診率の考察

各クラスタの初再診率を図 5 に示す．初再診率を見ると糖尿病に関しては No.1 の約 98%が初診患者であり，No.1 以外のクラスタはほぼ再診患者が占めている．同様にインフルエンザ・風邪に関しても初診患者のクラスタが No.2 と No.4，再診患者のクラスタが No.1 と No.3 であることが読み取れる．また遷移グラフと比較してみると，初診患者クラスタには自身への遷移が存在せず，初再診率と遷移グラフの形状が一致する．これらのことから本手法を用いて時間的情報を反映した遷移グラフが作成できていると考えられる．

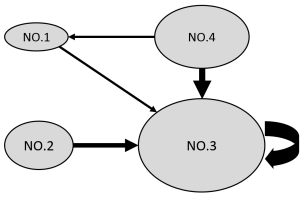

図 3: プロセスマイニング結果「インフルエンザ・風邪」

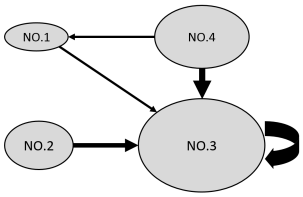

図 4: プロセスマイニング結果「糖尿病」

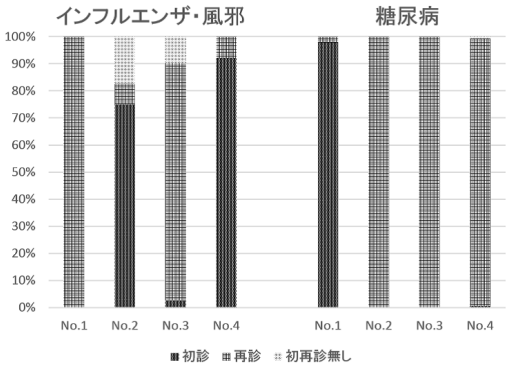

図 5: 各クラスタの初再診率

6.3 インフルエンザ・風邪の考察

6.3.1 患者状態ラベル・状態キーワード

各クラスタの患者状態ラベル分布を図 6 に，インフルエンザ・風邪の状態キーワードを表 5 に示す．No.1 の状態キーワードには黄色，白色，透明などの鼻汁の色を表現する単語や，咳嗽や湿性などの咳の症状に関連したキーワードが抽出された．一方でインフルエンザの症状の特徴である発熱に関するキーワードは抽出されていない．このことやインフルエンザのクラスタである No.4(詳細は後述) から遷移があることを考え，No.1 は，インフルエンザ患者の完治一歩手前の状態を表したクラスタである と推測できる．

No.2 は状態ラベル分布を見ると罹患中である割合が増加し，状態キーワードからは寒気，感冒，かぜなどの風邪の諸症状に関するキーワードが抽出されている．加えて「学校でインフルエンザが流行」や「インフルエンザの疑い」等の記述がみられるものの，インフルエンザと診断している診療データは存在しないことから No.2 は風邪のクラスタである と推定できる．

No.3 は状態ラベル分布の罹患中の割合が 5%へ大きく減少している．また，改善したとみられる割合が微増し，その他の割合が 85%を占めている．そのため No.3 は症状回復，もしくは異なる病例へ遷移した患者のクラスタである と考えられる．そのためキーワードには改善，良好といった症状改善に関するキーワードの他に，前立腺や血圧などインフルエンザ・風邪とは関連が低いキーワードが抽出された．

No.4 の状態ラベル分布は罹患中の割合が 92%を占めている．またインフルエンザと断定されたデータ 34 件のうち約 80%がこのクラスタに存在することから，No.4 はインフルエンザ患者のクラスタである ことがわかる．キーワードにはインフルエンザのほかに高熱，発熱など体温上昇に関する記述が多くみ

られ，罹患時には体温が著しく上昇するインフルエンザの典型的な症状と一致する．

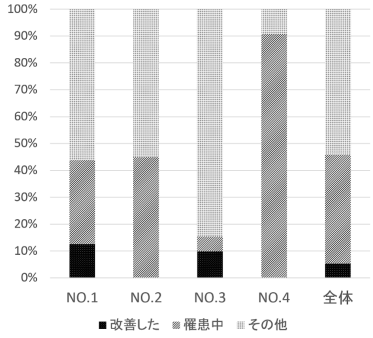

図 6: 各クラスタの患者状態ラベル分布

表 5: 「インフルエンザ・風邪」状態キーワード

クラスタ No	S・A から抽出したキーワード
No.1	鼻汁, 咳嗽, 胸部, 湿性, 鼻炎, 黄色, アレルギー性, 白色, 透明, インフルエンザ
No.2	全身, 咽喉, 寒気, 疑い, 感冒, 感染, 発熱, かぜ, ウイルス, 高熱
No.3	全身, 前立腺, 血圧, 説明, 改善, 検査, 腹部, 低下, 良好, 疑い
No.4	咽喉, 胸部, 全身, 寒気, 高熱, 発熱, 微熱, 急性, インフルエンザ, 感冒

6.3.2 遷移に基づくキーワード

「No.1 → No.3」,「No.2 → No.3」,「No.4 → No.1」,「No.4 → No.3」の 4 つの遷移の Plan(計画) から得られたキーワードを比較するとほぼ同様のキーワードが得られる結果となった．所見欄の内容は「再検査・精密検査が必要です.」や「仕事を休み充分な安静，睡眠時間を取る」など検査の必要を伝える記述や休養をすすめるなどの記述がなされていた．キーワードが似通っていることから，どの状態においても同じような診療行為がなされたと考察できる．

6.4 糖尿病の考察

6.4.1 状態キーワード

No.1 クラスタは他院からの転院事例や，検診により異常を指摘されたとの指摘が多く見られた．合わせて，月日のキーワードが抽出され，いつからどういった症状が起こったかという患者の主訴情報を記録した文章が多く見られた. 患者の主訴には，腹部や胸部は痛みや不快感を訴えるものが多く存在し，腹部，胸部のキーワードが抽出された．これらの症状は，糖尿病の症状の一つである．

No.2 は糖尿病と診断している例や，何型の糖尿病であるかと言った診断結果の表記が多く，糖尿，疑いというキーワードが抽出されている．また血圧、血糖のコントロールを評価している記述、血液検査の説明などが存在した．

No.3 の所見欄にはリピトールや高脂血症など脂質異常症に関する記述が多く見られた．状態キーワードからは血圧や血糖コントロールを評価する表記や，自宅で行った血圧測定の結果を記入していることもあり、血圧、血糖、コントロール，自宅のキーワードが抽出された．

No.4 は 2 名の患者の定期健診クラスタとなった. 投薬情報がやや他と異なるため独立したと思われる．

6.4.2 遷移に基づくキーワード

糖尿病の遷移に基づくキーワードを表 6 に示す．インフルエンザ・風邪の場合とは異なり，糖尿病の場合には抽出されたキーワードに違いが見られた.「No.3 → No.2」には PLT, HbA(1c)，BUN などの検査項目に関するキーワードが抽出されている．また抽出されたキーワードのうち，末梢，検査，血液，一般から末梢一般血液検査，CRP から CRP 検査を行っていることがわかり，「No.3 → No.2」の遷移では検査が行われていることが判明した.

「No.2 → No.3」と「No.2 → No.2」の遷移に関してはよく似た文章が抽出された．内容は 検査を促す記述や，現在の治療を続行，経過観察という記述が多く見られた. 例として「歩行 1 日 8000 歩程度、合計 30 分〜60 分程度。週 3 回以上」などの運動療法を指示する内容が存在し，程度，療法などのキーワードが抽出された.「No.2 → No.2」の遷移に関しては,「No.2 → No.3」より療法が上位のキーワードとなっていることや，運動，食事などのキーワードが抽出されていることもあり，運動療法，食事療法に関する記述が多く見られた．

表 6: 「糖尿病」遷移に基づくキーワード

クラスタ間遷移	Plan から抽出したキーワード
No.1 → No.2	検査, 必要,ABI,ECG, 血圧,mmHg,PWV, 次回, 精密, 来院, 初診
No.3 → No.2	PLT, 末梢, 検査, 血液,CRP,HbA, 定量, グルコース,BUN, 一般
No.2 → No.3	程度,ABI, 検査, 以上, 療法, 血圧, 現在, 続行, 治療,mmHg
No.2 → No.2	検査, 療法, 程度, 運動,IRI, 食事,ABI,gOGTT, 血糖, 次回

7. まとめ

本研究では診療所電子カルテデータを元にした初診から治癒までのプロセスを導く，診療プロセスマイニングを行った．プロセスマイニング結果に対して初再診情報，所見欄の内容，電子カルテの特定の項目から抽出したキーワード，状態ラベル分布と照合し，主観的に評価を行った．初再診率からは診療データの時系列情報を遷移グラフに反映できていることが判明し，状態ラベル分布からは，各クラスタの病態の変化を読み取ることができた．また電子カルテの構造に基づいて抽出したキーワードに関して，インフルエンザ・風邪の状態キーワードからは風邪のクラスタ，インフルエンザのクラスタなど各クラスタを特徴付けるようなキーワードが抽出された．一方糖尿病は遷移に基づくキーワードから，診療内容の違いを明らかにした．

以上を踏まえ，本研究で提案するプロセスマイニング手法により診療プロセスを抽出できることを確認した．

8. 今後の課題

今回の実験では，2 症例についてのみプロセスマイニングを行った．しかし，診療所や病院で扱う症例は多岐にわたり，その診療プロセスも多様であることが予想される．そのため本手法の有効性を検証するためには，他の症例に関しても実験を行う必要がある．また，医療機関によっても診療方法に違いがあるため，一つだけではなく複数の診療所や病院のデータを用いたプロセスの比較，検証を行う必要がある．

参考文献

[三浦 2010] 三浦康秀，荒牧英治，大熊智子，外池昌嗣，杉原大悟，増市博，大江和彦: 電子カルテからの副作用関係の自動抽出，言語処理学会第 16 回年次大会，pp78-81，2010

[荒牧 2009] 荒牧英治，三浦康秀，外池昌嗣，大熊智子，増市博，大江和彦: 退院サマリ文章可視化システムの構築，言語処理学会第 15 回年次大会，pp.348-351，2009