様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９　（共通）

科学研究費助成事業　　研究成果報告書

令和　元 年　 ６ 月　 ９ 日現在

機関番号：１６４０１
研究種目：基盤研究(C)（一般）
研究期間：2016～2018
課題番号：１６Ｋ０９１７２

研究課題名（和文）大規模解析のためのサマリ及びオーダ情報連携による患者背景推定手法の構築

研究課題名（英文）Classification process for medical data combined with order information and text information in hospital information system for estimation of patient background

研究代表者
畠山　豊（HATAKEYAMA, Yutaka）

高知大学・教育研究部医療学系連携医学部門・准教授

研究者番号：００３７６９５６

交付決定額（研究期間全体）：（直接経費）　　3,100,000円

研究成果の概要（和文）：病院情報システムに蓄積されている退院サマリや初診時記録などのテキストデータに対して潜在トピックモデルを適用して特徴量を抽出し、病名登録情報、処方オーダ情報などの構造化されたオーダ情報の解析結果と組み合わせて評価することで、各対象患者の背景を同一の群を抽出するなどの、より精度の良い分類結果を抽出することが可能になり、また、テキストデータ特徴を加えることデータ分類結果の説明をすることより容易な結果が得られた。

研究成果の学術的意義や社会的意義
本研究により、病院情報システムに蓄積されているテキストデータを適切に利用し、検査オーダ情報などの構造化データと組み合わせることで、これまでより精度よく解析が行うことが可能となった。また、医師の意図を反映させた解析を行うことが可能となった。このことにより、病院情報システムデータに対する解析がより進捗し、知見も得られやすくなり、患者への医学的な知見を反映させやすくなったと考える。

研究成果の概要（英文）：We constructed the classification methods for patient data in hospital information system combined with order information and text information which is processed by latent topic model in order to obtain the characteristics for patient in each cluster. The classification results show that it is easy to explain the background of each patient based on the feature of input text information.

研究分野：医療情報学

キーワード：医療情報学

様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９、ＣＫ－１９（共通）

１．研究開始当初の背景

近年、病院情報システムに蓄積されているデータを対象として大規模データ解析が積極的に行われている。データソースから複数の解析対象群を抽出することで解析を行うが、不均質なデータなため適切な条件で抽出することが重要である。このようなデータに対する手法として、propensity score や disease risk score などが提案されている。これらの手法は各患者の背景因子(共変量)を用い調整スコアを算出し、結果の補正を行っている。これらスコアの算出には、各種検査結果、登録病名情報、処方オーダ情報などの構造化されたデータが用いられている。検査結果情報は客観データとなるが、病名やオーダ情報などは医師の判断が反映された結果データなため、不均質なデータになる可能性がある。

また、病院情報システムに蓄積されているデータにおいて検査の結果や処方などのオーダ情報がよく使われる。これらのデータは構造化されたデータのため、計算機上で処理が容易なためである。一方、構造化されていない自由記載文章も電子カルテの導入以降膨大に蓄積されているが、テキスト処理を行ったうえで解析を行う必要があるため、解析研究で利用されることはまれである。そこで、非構造化データも加えて背景因子が類似する患者データの精度の良い抽出が必要となる。

そのため、構造化データであるオーダ情報を医師の判断を考慮して評価・解析を行うためには、医師の意図が反映されたカルテ記載文章のデータ特徴を考慮する必要がある。

２．研究の目的

病院情報システムに蓄積されているオーダ情報とテキスト情報を組み合わせることでオーダ情報のみ患者状態を把握する手法よりも精度よく患者状態を評価することが可能かどうかの評価行う。２つのアプローチにより評価を行った。

まず、オーダ情報と退院サマリの記載情報を連携して敗血症によって引き起こされる急性腎障害（septic AKI）データ抽出および、その背景ごとに分類する手法を構築する。退院サマリに対して潜在トピックモデルを適用し、各サマリの特徴モデルから類似性を定義する。敗血症を含む病名登録履歴から患者データを抽出し、サマリの類似関係を評価することで、背景情報の評価・分類および病名履歴のない患者データ抽出を行う。

さらに、類似した症状の記載及びオーダがある同一の登録病名がある患者群を、類似背景患者群として考え、診療記録に患者背景がまとめて記載されている初診時の患者データを対象として、初診時記録及びオーダ情報に基づき患者群を抽出し、その群における類似背景患者群の比率を評価する。つまり、抽出した患者群における同一病名登録率の評価を行う。

３．研究の方法

高知大学医学部附属病院の病院情報システムにおいて蓄積されているデータを対象として解析研究を行う。septic AKI に対する分類する手法において、AKI の判定は KDIGO の血清クレアチニン(SCr)の変動に基づいて行う。入院期間中に AKI イベントが起こったサマリ情報を抽出する。病名履歴について、「敗血症」を含む履歴が AKI イベント前 2 週間、後 1 週間に存在している患者を対象とする。対象退院サマリデータに対して、潜在トピックモデル(LDA)解析を実施する。トピック数を 12 として実施する。実際の解析処理は R の topicmodels パッケージを利用し行う。LDA によって得られた各サマリにおけるトピックの確率値に対し、類似性を Jensen-Shannon divergence(JSD)によって計算する。敗血症の病名登録履歴がある患者における類似サマリを JSD<10^-4 として定義し抽出する。類似しているサマリを連結した群を類似群として扱う。その群において最低次数（連結数）5 以上の類似群に対し、それぞれのサマリを手動評価することで、背景情報の評価を行う。

初診時記録及びオーダ情報に基づく同一病名登録率の評価においても、同様に入力テキスト情報に対し形態素解析を実施し、得られた単語データの潜在トピックモデル解析結果を各入力テキストデータの特徴量として利用する対象とするテキストデータは、総合診療部の初診時記録データとする。算出されたテキストデータの特徴量（確率分布）の Jensen-Shannon ダイバージェンスを各テキスト間の非類似度として定義し、Ward 法による階層化クラスタリングを実施する。一方、初診時において出された処方オーダ及び画像検査オーダを対象として、オーダ情報によるクラスタリング結果に基づき分類を行う。処方オーダに関して、日本標準商品分類番号における 3 桁の薬効分類コードごとに集計しオーダの「あり」「なし」の 2 値で表現を行い、画像検査オーダに関して一種類のオーダとして「あり」「なし」の 2 値で表現する。オーダ情報の特徴をビット列として用い、各患者のオーダ特徴間の違いをマンハッタン距離によって表現する。Ward 法による階層化クラスタリングを実施することで、分割を行う。各初診時患者データ対して独立にクラスタリングを行った結果を組み合わせ新規のクラスタを生成し、各クラスタに割り当てられた患者における各病名登録率を評価する。各患者に対してテキストデータで割り当てられたクラスタ情報(1)及びオーダ情報で割り当てられたクラスタ情報(2)を特徴として、同一クラスタ結果で分割を行う。つまり、(1)で割り当てられたクラスタ情報と(2)で割り当たれたクラスタ情報が一致する患者群を一つのクラスタとして扱う。この対象群ごと

に 3 桁の ICD10 コード情報を集計し、対象群の患者数に対する各 ICD10 コード登録患者数の比を病名登録率として評価する。また、、全対象患者を ICD10 コード情報毎に分割し、分割した群が 2 特徴で生成されたクラスタで、どの程度被覆されるか評価する。

４．研究成果

septic AKI に対する分類において、対象病名履歴が 58 件存在した。LDA に基づく類似群の関係をネットワークグラフで表現した結果を図 1 に示す。次数が 5 以上の群が 3 つ生成され、直線で囲まれた群が DIC の発症、全体状態の悪化について記述され、二重線の群が主として意識障害について記述され、点線の群が白血病について記述されていた。履歴がないが類似サマリと判定されたデータが 31 件存在した。

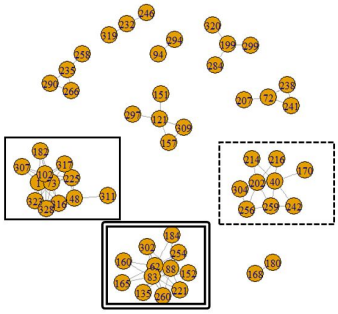

図1　LDAに基づく類似群のネットワークグラフ

LDA によって得られたトピック毎の単語特徴では確率が高い単語は「施行」や「錠」などが共通して出ていたが、以降の単語では心臓に関連する単語や泌尿器に関する単語がトピック毎に異なって表れてきており、適切にトピックを抽出していると考える。その結果、図 1 で示されているように、同じような症例群が類似サマリと判定されている。

確実に septic AKI と判定できる条件として、病名登録イベントを用いた。入院中では必ずしも病名登録が行われないため、敗血症データの漏れが存在する。直線で囲まれた類似群における 1 患者は、敗血症についての病名登録が行われていない。しかし、履歴のあるサマリに対して類似サマリと判定されていた。細菌検査の履歴を確認した結果、AKI 判定以降 2 週間以内に血液培養で菌検出結果が存在していた。これらの結果から確実に septic AKI と判定できるデータとサマリの類似性から septic AKI 可能性患者の抽出が可能である結果が示せた。サマリ中に腎機能について記述されていないと適切な類似判定が期待できないという限界が存在するが、septic AKI 患者の抽出、および背景による分類は十分可能と考える。

初診時記録及びオーダ情報に基づく同一病名登録率の評価においては、初診時記録に記載のある患者 548 人のデータに対し形態素解析を実施、単語抽出条件を満たす 543 人のデータが解析対象となった。初診時に対象患者に対する登録病名は、3 桁の ICD10 コードに関し、158 コード存在した。オーダ情報に関しては、画像検査オーダが存在する患者が 16 名であり、処方オーダは薬効コードで 37 コード存在した。

各クラスタリング結果を組み合わせたクラスタを生成した結果、患者が存在するクラスタ数は 94 となった。5 人以上の患者数をもつクラスタ数が 28 で、患者数の 78%を占める。患者数が多く存在するクラスタ及び最大 ICD10 コード登録患者数の結果を表 1 に示す。漢方製剤以外を主に処方する患者数 5 人以上のクラスタでの病名登録率の最大値は、初診時クラスタ及びオーラクラスタが(1,1)の時、77%となった。10 名以上登録されている ICD10 コードを対象として、

表1　テキスト及びオーダ情報による分類結果

初診時クラスタ	初診時クラスタ患者数	トピック番号	オーダクラスタ	薬効コード	薬効分類名	患者数	ICD10	ICD10名称	ICD10登録患者数
1	20	12	1	231	止しゃ剤、整腸剤	13	A09	感染症と推定される下痢及び胃腸炎	10
1	20	12	6	114	解熱鎮痛消炎剤	4	A09	感染症と推定される下痢及び胃腸炎	3
2	46	11	2	「なし」		40	I10	心臓併発症の記載のないリウマチ熱	3
3	64	5	3	520	漢方製剤	18	F48	その他の神経症性障害	4
3	64	5	2	「なし」		17	F41	その他の不安障害	5
4	75	13	3	520	漢方製剤	57	R68	その他の全身症状及び徴候	11
5	36	10	2	「なし」		18	M54	背部痛	6
5	36	10	3	520	漢方製剤	9	M54	背部痛	3
6	64	3	3	520	漢方製剤	15	J00	急性鼻咽頭炎[かぜ]<感冒>	6
6	64	3	8	222	鎮咳剤	14	J06	多部位及び部位不明の急性上気道感染症	6
7	21	18	4	117	精神神経用剤	9	F32	うつ病エピソード	9
7	21	18	2	「なし」		5	F32	うつ病エピソード	2
8	56	6	2	「なし」		30	I10	心臓併発症の記載のないリウマチ熱	3
9	31	19	2	「なし」		13	R42	めまい<眩暈>感及びよろめき感	8
9	31	19	3	520	漢方製剤	8	R51	頭痛	4
10	18	15	2	「なし」		10	R63	食物及び水分摂取に関する症状及び徴候	2
11	37	14	2	「なし」		19	M47	脊椎症	2
12	18	4	2	「なし」		9	T78	有害作用，他に分類されないもの	5
12	18	4	7	614	主としてグラム陽性菌、マイコプラズマに作用するもの	7	A37	百日咳	7
13	28	20	2	「なし」		20	R10	腹痛及び骨盤痛	3
14	29	2	2	「なし」		23	M62	その他の筋障害	2

表2　ICD10コード登録結果に対する分類結果

ICD10コード	ICD10名称	登録患者数	初診時クラスタ	オーダクラスタ	クラスタ患者数	記載医師数
R68	その他の全身症状及び徴候	16	4	3	11	3
A09	感染症と推定される下痢及び胃腸炎	22	1	1	10	4
J02	急性咽頭炎	11	6	8	4	3
K59	便秘	11	4	3	4	2
R63	食物及び水分摂取に関する症状及び徴候	14	4	3	5	3
J06	多部位及び部位不明の急性上気道感染症	17	6	8	6	6
J00	急性鼻咽頭炎[かぜ]<感冒>	22	6	3	6	3
M54	背部痛	22	5	2	6	4
R42	めまい<眩暈>感及びよろめき感	31	9	2	8	6
R53	倦怠（感）及び疲労	39	4	3	10	3

ICD10 コードに基づき分割し 21 クラスタを生成した。ICD10 によるクラスタを組み合わせることによって生成されたクラスタ結果に基づいて細分化した結果を表 2 に示す。

各患者に対して初診時クラスタ情報及びオーダクラスタ情報に基づき細分化したクラスタを構築するため、多くのクラスタが生成される可能性がある。しかし、ほとんどのクラスタに患者は割り当てられていない結果となった。これは初診時記録の特徴とオーダ情報の特徴が関連していることを示していると考える。つまり、同じ患者の症状に対して、同じ処方などが行われていることが病院情報システム上に記録されていることが示されている。

表 1 で示されているように、2 つのクラスタリング結果により細分化することで、症状が同じでもオーダを出していない群とそれ以外の群に分割することが可能となる。つまり、オーダ情報をテキスト情報の解析結果に組み合わせることで対象群の絞り込みが可能な結果示している。そのため、処方オーダの有り無しによる登録病名の絞り込みにも寄与していると考える。また、症状だけで病名がほとんど決定される患者群も存在している。これまでの解析とは逆に同じ ICD10 コードを登録している患者群であっても、症状やオーダが異なっていることを表 2 が示している。実際の登録病名から ICD10 コード 3 桁に変換している影響もあるが、同じ病名登録をしていても記載されている患者背景やオーダも異なっている。そのため、病名だけで解析用患者データ抽出を行っても異なる患者背景群が複数存在している可能性を示している。

初診時記録を詳細に記載している総合診療部データを入力として用いることでテキストデータによる分割が良好な結果が得られ、高頻出の単語がトピック毎に異なる結果になったと考える。他の診療科でも扱う疾患ごとに患者の症状を記載する単語が異なっている場合は本手法でも十分類似背景患者群を抽出可能だと考える。

本手法の限界は、処方情報が限定できない患者に対しては適切な患者背景情報が得られないことである。つまり、処方されていない患者群では十分な特徴情報が得られないため、病名登録率が低い傾向を示す。逆に漢方製剤を処方されている患者群では薬剤の適用症例範囲が広いため、処方情報が患者状態識別の情報量としては低くなる。そのため、処方情報がない群と同様に病名登録率が低い結果となっている。

初診時記録及びオーダ情報の特徴量に基づき分類したそれぞれの結果を組み合わせて生成した患者群での病名登録率を評価した結果、特徴的な記述及び処方の履歴がある集団では同一の病名を登録している傾向を示し、類似した背景を持つ患者群が抽出可能なことを示した。

これらの解析結果から、非構造化データであるテキストデータを適切に解析し、オーダ情報と連携することにより、より精度の良く分類などのデータ解析を行うことが可能なことを確認した。この研究成果により、病院情報システムに蓄積されているデータを医師の意図を把握したうえで解釈・評価することが可能になり、幅広い応用が可能になると考える。

５．主な発表論文等

〔雑誌論文〕（計 2 件）

1. Hatakeyama Y, Horino T, Nagata K, Kataoka H, Matsumoto T, Terada Y, Okuhara Y. Evaluation of the accuracy of estimated baseline serum creatinine for acute kidney injury diagnosis. Clin Exp Nephrol. 2018 Apr;22(2):405-412. doi: 10.1007/s10157-017-1481-y.

2. Hatakeyama Y, Horino T, Nagata K, Kataoka H, Matsumoto T, Terada Y, Okuhara Y. Transition from acute kidney injury to chronic kidney disease: a single-centre cohort study. Clin Exp Nephrol. 2018 Dec;22(6):1281-1293. doi: 10.1007/s10157-018-1571-5

〔学会発表〕（計　2 件）

1. Yutaka Hatakeyama, Hiromi Kataoka, Noriaki Nakajima, Teruaki Watabe and Yoshiyasu Okuhara. Baseline estimation for Serum Creatinine for definition of Acute Kidney Injury. 15th IEEE/ACIS International Conference on Computer and Information Science 7th International Workshop on Intelligent Computational Science

2. Yutaka Hatakeyama, Noriaki Nakajima, Teruaki Watabe, and Yoshiyasu Okuhara. Association Analysis between Nutrition Condition and Acute Kidney Injury for Early Detection of Onset. 7th International Symposium on Computational Intelligence and Industrial Applications

〔図書〕（計 0 件）

〔産業財産権〕
○出願状況（計 0 件）

○取得状況（計 0 件）

〔その他〕

６．研究組織

(1)研究分担者
研究分担者氏名：奥原　義保
ローマ字氏名： OKUHARA, Yoshiyasu
所属研究機関名：高知大学
部局名：教育研究部医療学系連携医学部門
職名：教授
研究者番号（8 桁）：40233473

研究分担者氏名：片岡　浩巳
ローマ字氏名： KATAOKA, Hiromi
所属研究機関名：川崎医療福祉大学
部局名：医療技術学部
職名：教授
研究者番号（8 桁）：80398049

研究分担者氏名：渡部　輝明
ローマ字氏名： WATABE, Teruaki
所属研究機関名：高知大学
部局名：教育研究部医療学系連携医学部門
職名：講師
研究者番号（8 桁）：90325415

研究分担者氏名：中島　典昭
ローマ字氏名： NAKAJIMA, Noriaki
所属研究機関名：国立研究開発法人国立がん研究センター
部局名：情報統括センター
職名：研究員
研究者番号（8 桁）：00335928

(2)研究協力者
研究協力者氏名：
ローマ字氏名：