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　２ 年　 ６ 月　 ５ 日現在

機関番号：１２５０１
研究種目：基盤研究(C)（一般）
研究期間：2016～2019
課題番号：１６Ｋ００４５８

研究課題名（和文）多施設統合退院サマリーデータベースの臨床応用

研究課題名（英文）Clinical application of multi-institution integrated discharge summary database

研究代表者
鈴木　隆弘（Suzuki, Takahiro）

千葉大学・医学部附属病院・准教授

研究者番号：４０３２３４２２

交付決定額（研究期間全体）：（直接経費）　　3,500,000円

研究成果の概要（和文）：本研究「多施設統合退院サマリーデータベースの臨床応用」は、連携研究者の所属する千葉大学病院、聖路加国際病院、高知大学病院、長崎大学病院、大阪大学病院、香川大学病院,群馬大学病院の７施設から約３５万例の退院サマリーを収集して、千葉大学で開発したテキストマイニング技法を用いて解析することで、共通の文書ベクトル空間を持つ大規模なテキストデータベースの構築を行い、施設や疾患による属性の違いを検証した。また、構築したデータベースを用いて、自動疾患判定や類似症例の検索などの応用アプリケーションを開発し、千葉大学において試用を行った。

研究成果の学術的意義や社会的意義
近年の日本は電子カルテシステムの普及期を迎え、医療情報は電子化された形で蓄積されつつある。しかし、それらの多くは数値データや画像データが対象であり、医師のカルテ記録や退院時サマリーなどの診療文書類は先送りされてきた。また、個人情報保護法の制約のため、複数施設間の情報を集約した研究は進んでいなかった。本研究では複数の医療機関から退院サマリーを電子的に抽出し、テキストマイニング技術によって共通の文書ベクトル空間を構築した大型文書データベースを作成し、類似症例の検索やＤＰＣの自動判定を初めとした様々な応用を試行することで、将来的なテキスト利用の基礎として、臨床医学に貢献できることが期待できる

研究成果の概要（英文）：This study, "Clinical application of multicenter integrated discharge summary database" describes the collaboration researcher's affiliated Chiba University Hospital, St. Luke's International Hospital, Kochi University Hospital, Nagasaki University Hospital, Osaka University Hospital, Kagawa University Hospital, Gunma University Hospital. A large-scale text database with a common document vector space was constructed by collecting about 350,000 discharge summaries from 7 facilities and analyzing them using the text mining technology developed at Chiba University. We verified the differences in attributes due to facilities and diseases. Using the constructed database, we developed applications such as automatic disease determination and retrieval of similar cases, and conducted trials at Chiba University.

研究分野：医療情報学

キーワード：テキストマイニング　退院サマリー　多施設研究　データベース

様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９（共通）

１．研究開始当初の背景

電子カルテシステムが一般化するのに伴って医療情報は電子化された形で蓄積されつつあり、これを解析する試みが行われている。しかし、医師のカルテ記録などの診療文書類の分析は遅れていた。また、個人情報保護法の制約のため、複数施設間の情報を集約した研究は進んでいなかった。我々は早くから診療文書の解析に取り組んでおり、複数の施設の文書を統合した巨大なテキストデータベースを構築してテキストマイニングが可能となれば、精度の向上だけでなく応用範囲も広がると期待された。（図１）

２．研究の目的

我々は、複数の医療機関から退院サマリーを電子的に抽出し、テキストマイニング技術によって共通の文書ベクトル空間を構築して様々な応用を可能とする統合大型データベースの構築を目指して、千葉大学医学部附属病院、群馬大学医学部附属病院、香川大学医学部附属病院、高知大学医学部附属病院、長崎大学病院、大阪大学医学部附属病院、聖路加国際病院の7施設から約35万例の退院時サマリーを収集して解析を行うと共に、DPC コードの自動判定や類似症例の検索など、得られたデータベースを用いた応用アプリケーションの開発を行った。対象となった退院時サマリー数は千葉大学 67266 例、大阪大学 58726 例、香川大学 20836 例、高知大学 34994 例、長崎大学 61582 例、聖路加国際病院 73065 例、群馬大学 31432 例、総計として 347901 例であった。

３．研究の方法

3.1.形態素解析とベクトル化

抽出したサマリーから重要語を抽出してテキストデータベースを作成した。形態素解析に MeCab を使用し、解析用辞書としては、我々がこれまでに作成した２３万語の医療用語辞書とパラメディカル用医療辞書の ComeJisyoV5 を併用した。得られた索引語が病態の特徴を表す上でどれだけの重要度を持っているか示す為に、索引語の重み付けの処理が必要となる。本研究では重み付けにシンプルで処理速度が速いことから、索引語の文書集合内での出現頻度と文書集合間の出現頻度を用いる Tf×Idf 法を用いた。（図２）

得られたデータからベクトル空間を構築し、検索対象のサマリーベクトル空間と入力されたサマリーを比較して、最も類似したベクトルを検索する。類似度の算出はベクトルの内積によって求める。これらの手法により、ＤＰＣの自動判定や類似症例検索などが可能となる。

3.2.退院時サマリー属性の施設間比較

得られたテキストデータベースから、単語数や索引語の傾向など、退院時サマリーの各種属性について施設別、疾患別などの様々な観点からの比較を行った。

3.3.DPC 自動判定と病院間クロスマッチ

各施設から症例数が 20 以上の DPC コードを抽出し、7 対 3 の比率で無作為にモデル作成用と検証用のサマリーに分け、検索対象のサマリーベクトル空間と入力されたサマリーを比較して最も類似したベクトルを検索し、DPC を判定する。参加 7 病院各々の検証用データとモデルデータおよび、全病院のデータを統合したモデルデータとの間で 7 対 8 のクロスマッチ自動判定を、DPC コードの 14 桁全てと上位 6 桁のそれぞれで実行した。（図 3）

3.4.Web アプリケーションの開発

構築したテキストデータベースを利用して、退院時サマリーから類似度の高い DPC コードを検索して一覧表示することが出来る Web アプリケーションを開発した。（図４）。

４．研究成果

4.1.退院時サマリーの施設間比較

索引語の総出現回数は 63,587,539 回で、単語数は 357,017 語であった。各施設別では千葉大学 83,546 語、群馬大学 53,038 語、香川大学 91,658 語、高知大学 98,556 語、長崎大学は 62,457 語、大阪大学 175,478 語、聖路加国際病院 117,422 語であった。1 サマリー内の平均単語数は千葉大 150 語、群馬大 123 語、香川大 242 語、高知大 241 語、長崎大 68 語、大阪大 295 語、聖路加国際病院 210 語で、大阪大、高知大は長めで千葉大、群馬大、長崎大では短かった。(表1)

索引語数別のサマリー数を施設ごとに表示したグラフを図5に示す。施設による差は大きいが、概ね 200 語台にピークがあり、聖路加国際病院はバラツキが少なかった。

出現頻度の高い索引語を施設別に表2に示す。語には違いが見られるが、いずれの施設でも見出しに類する索引語が上位を占めている。全施設の索引語を MDC 別にまとめると、見出し語が上位を占めるのは施設別と同様であるが、15 位以降になると疾患に特徴的な語が増えてくる。

疾患とサマリーの長さとの関係では、全ての病院で眼科疾患、耳鼻科疾患は短く、精神科疾患では長い傾向が認められた。(図6)

4.2.DPC クロスマッチ判定

DPC コードは 6 桁までで疾患を指定する。表3に DPC6 桁までを用いたクロスマッチ照合の結果を示す。いずれの病院でも自施設のモデルデータと検証データの組み合わせでは、70～84%が一致と高い

判定率を示した。異施設との間では 32～76%の判定率に低下するものの、全体としても 65%と高い判定率を示した。全施設を統合したモデルデータとの検証では 65～82%と自施設と同等の判定率を示した。

表4に DPC14 桁全て一致での判定結果を示す。判定率は自施設同士で 55～72%、異施設との間でも 33～60%と 6 桁よりは下がるものの全体としての判定率は 54%で、統合モデルデータとの照合では、6 桁の場合と同様に自施設と同等の判定率を示した。

疾患別の傾向として、図7に全病院の統合データに対する判定結果を MDC 別に示す。判定率の高い疾患としては眼科疾患、乳房の疾患、新生児疾患などが 100%近い判定率を示した。一方で判定率の低い疾患としては血液･免疫系疾患、その他の疾患などが挙げられた。

4.3.Web アプリケーションの機能

開発した Web アプリケーションは退院日を指定して全病棟の退院サマリーから DPC コードを推定する機能と、患者を直接指定して類似度の高い DPC コードを提示する機能を備えていて、どちらも照合の対象として各施設別と全国共通のベクトルを切り替えて DPC を付与することが出来る。解析はサマリー単位で行われ、類似度の高い順に 10 例の DPC コードおよび最も医療資源を投入した病名、ICD10 病名、類似度及びキーワードを表示する。

4.5.まとめ

本研究によって複数施設を統合し、施設間の違いを包含した大規模テキストデータベースを構築し、疾患による違いを反映していることを示すことができた。近年は BERT（Bidirectional Encoder Representations from Transformers）をはじめとした新たなテキスト解析手法が提案されているが、医療への応用はまだ少ない。本研究で得られたデータはこれらの新技術を適用するための基礎データとしても利用可能で価値の有るものである。

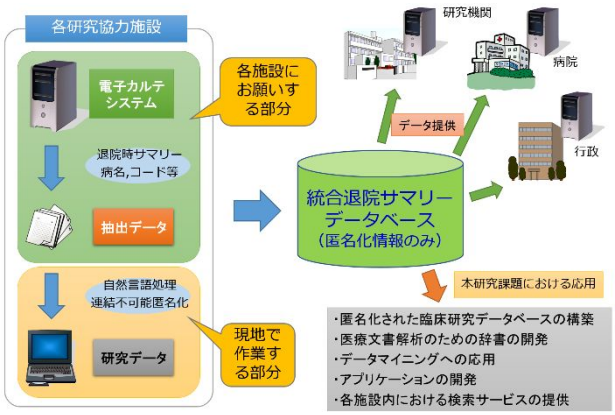

図１．　退院サマリーデータベース概念図

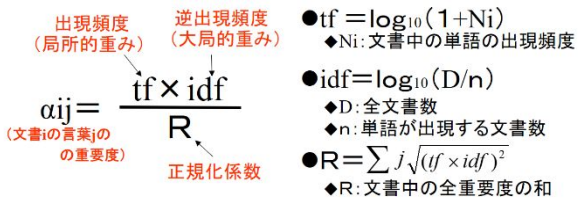

図２．Tf×Idf 法

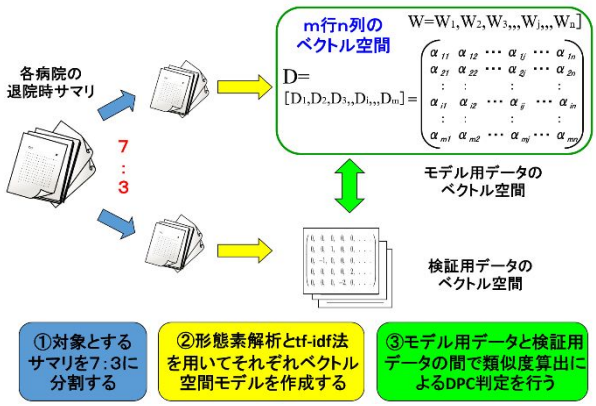

図３．　病院間クロスマッチテスト

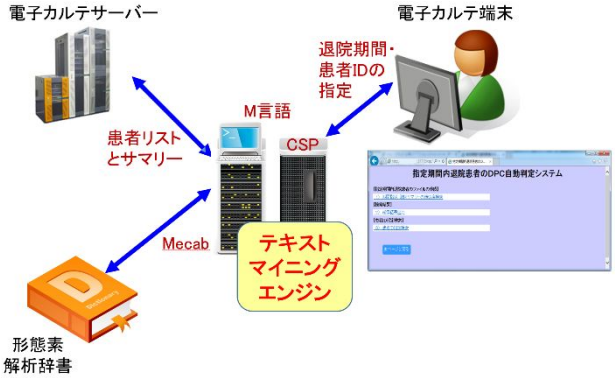

図４．　DPC 判定アプリケーション概要

表１．施設毎のサマリー内索引語数

病院名	平均	中央値	標準偏差	最大	最小
千葉大学	148.7	126	101.2	1097	4
群馬大学	122.5	109	81.4	640	7
香川大学	242.4	215	103.7	1111	61
高知大学	241.2	220	137.2	1166	2
長崎大学	67.8	120	49.9	408	1
大阪大学	295.5	258	148.0	1442	86
聖路加	209.9	208	85.6	519	2
全体	187.7	167	128.6	1442	1

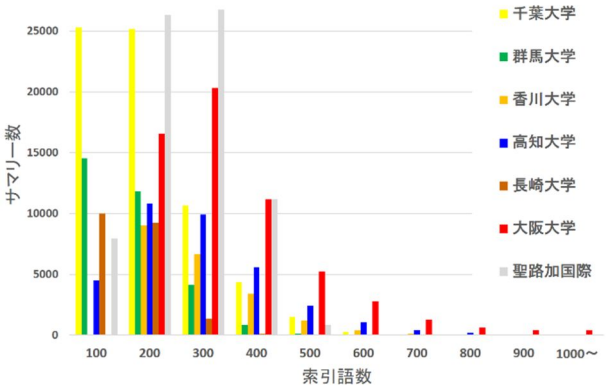

図５．索引語数別のサマリー件数

表２．出現頻度の高い索引語

順位	千葉大学	群馬大学	香川大学	高知大学	長崎大学	大阪大学	聖路加
1	入院	入院	ID	入院	退院	氏名	現病歴
2	退院	手術	コード	mg	施行	住所	検査所見
3	経過	既往歴	内容	退院	入院	大阪大学	mg
4	所見	経過	患者	当科	月	生年月日	軽快
5	既往歴	治療方針	病歴	施行	経過	患者	主訴
6	治療	概要	種別	月	開始	番号	所見
7	KG	退院時処方	項目	診断	mg	病院	既往歴
8	来院	方針	順序	病院	予定	紹介	退院
9	CM	退院	月	紹介	外来	アレルギ	入院
10	職業	手術所見	基本	+	方針	病棟	身体

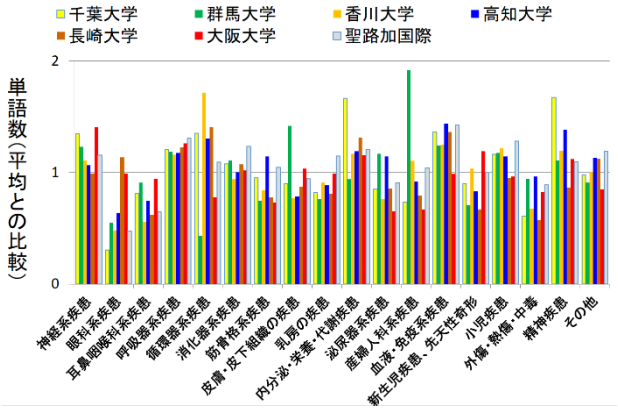

図６．　ＭＤＣ別索引語数比較

表３．　DPC6 桁によるクロスマッチ結果

モデル／検証	千葉大学	群馬大学	香川大学	高知大学	長崎大学	大阪大学	聖路加	統合データ
千葉大学	79%	55%	58%	68%	59%	68%	61%	76%
群馬大学	58%	79%	63%	70%	60%	72%	64%	77%
香川大学	43%	32%	70%	59%	40%	59%	43%	65%
高知大学	61%	63%	68%	82%	71%	76%	70%	82%
長崎大学	52%	49%	56%	66%	76%	68%	60%	77%
大阪大学	67%	63%	67%	75%	63%	84%	71%	81%
聖路加	64%	58%	56%	66%	61%	69%	77%	75%

表４．　DPC14 桁によるクロスマッチ結果

モデル／検証	千葉大学	群馬大学	香川大学	高知大学	長崎大学	大阪大学	聖路加	統合データ
千葉大学	69%	43%	42%	54%	46%	54%	49%	67%
群馬大学	50%	72%	51%	60%	52	61%	54%	71%
香川大学	37%	26%	60%	42%	27%	44%	30%	55%
高知大学	50%	50%	49%	73%	57%	62%	57%	72%
長崎大学	44%	39%	43%	56%	69%	57%	50%	71%
大阪大学	52%	47%	47%	60%	52%	73%	57%	71%
聖路加	53%	48%	42%	56%	51%	57%	69%	68%

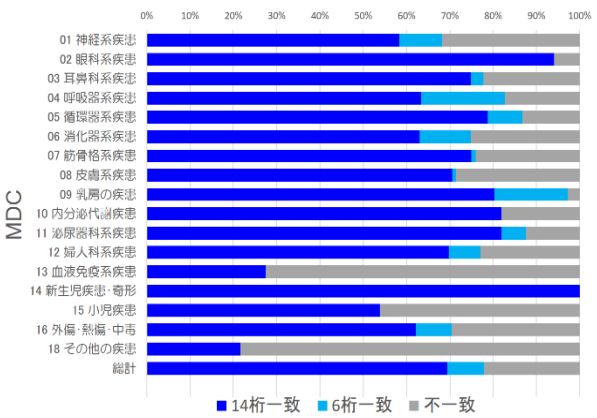

図７．　MDC 別の判定結果(統合データ)

５．主な発表論文等

〔雑誌論文〕　計1件（うち査読付論文　1件／うち国際共著　0件／うちオープンアクセス　0件）

１．著者名 鈴木隆弘、土井俊祐、木村隆、島井健一郎、嶋田元、野口怜、 畠山豊、松村泰志、松本武浩、横井英人、本多正幸	４．巻 29
２．論文標題 多施設統合退院サマリーデータベースの構築と臨床応用	５．発行年 2020年
３．雑誌名 MUMPS	６．最初と最後の頁 -
掲載論文のDOI（デジタルオブジェクト識別子） なし	査読の有無 有
オープンアクセス オープンアクセスではない、又はオープンアクセスが困難	国際共著 －

〔学会発表〕　計2件（うち招待講演　0件／うち国際学会　0件）

１．発表者名 鈴木隆弘、土井俊祐、近本圭祐 、川上英良、木村隆、島井健一郎、嶋田元、野口怜、畠山豊、本多正幸、松村泰志、横井英人
２．発表標題 多施設テキストデータベースを用いた退院時サマリー作成・監査支援の試み
３．学会等名 第39回医療情報学連合大会
４．発表年 2019年

１．発表者名 鈴木隆弘、土井俊祐、木村隆、嶋田元、畠山豊、本多正幸、松村泰志、横井英人、島井健一郎
２．発表標題 退院サマリー監査を支援するDPC判定アプリケーション
３．学会等名 第38回医療情報学連合大会
４．発表年 2018年

〔図書〕　計0件

〔産業財産権〕

〔その他〕

－

６．研究組織

	氏名 （ローマ字氏名） （研究者番号）	所属研究機関・部局・職 （機関番号）	備考
研究協力者	嶋田　元 (Shimada Gen)	聖路加国際病院	

6．研究組織（つづき）

	氏名 （研究者番号）	所属研究機関・部局・職 （機関番号）	備考
連携研究者	土井　俊祐 (Doi Shunsuke) (40323422)	東京大学・医学部附属病院・特任講師 (12501)	
連携研究者	高崎　光浩 (Takasaki Mitsuhiro) (70236206)	佐賀大学・医学部附属病院・准教授 (17201)	
連携研究者	津本　周作 (Tsumoto Shusaku) (10251555)	島根大学・医学部附属病院・教授 (15201)	
連携研究者	野口　怜 (Noguchi Rei) (50828861)	群馬大学・医学部附属病院・助教 (12301)	
連携研究者	畠山　豊 (Hatakeyama Yutaka) (00376956)	高知大学・医歯学系・准教授 (16401)	
連携研究者	松本　武浩 (Matsumoto Takehiro) (20372237)	長崎大学・医歯（薬）学総合研究科・准教授 (17301)	
連携研究者	松村　泰志 (Matsumura Yasushi) (90252642)	大阪大学・医学部附属病院・教授 (14401)	
連携研究者	横井　英人 (Yokoi Hideto) (50403788)	香川大学・医学部附属病院・教授 (16201)	

６．研究組織（つづき）

	氏名 （研究者番号）	所属研究機関・部局・職 （機関番号）	備考
連携研究者	本多　正幸 (Honda Masayuki) (10143306)	長崎大学・医歯（薬）学総合研究科・教授 (17301)	