平成２９年度厚生労働科学研究費補助金
（政策科学総合研究事業（臨床研究等ICT基盤構・人工知能実装築研究事業））総括研究報告書

「電子カルテ情報をセマンティクス（意味・内容）の標準化により分析可能なデータに変換するための研究」

研究代表者	宮本　恵宏	国立循環器病研究センター・循環器病統合センター・センター長
分担研究者	竹村　匡正	兵庫県立大学大学院・応用情報科学研究科・教授
	中村　文明	国立循環器病研究センター・循環器病統合センター・室長
	竹上　未紗	国立循環器病研究センター・研究開発基盤センター　予防医学・疫学情報部　室長
	興梠　貴英	自治医科大学・医療情報部・教授
	中山　雅晴	東北大学大学院・医学系研究科・教授
	的場　哲哉	九州大学病院・循環器内科・講師
	小室　一成	東京大学大学院医学系研究科・循環器内科・教授
	斎藤　能彦	奈良県立医科大学・循環器内科・教授
	安田　聡	国立循環器病研究センター・副院長・心臓血管内科部門長
	宍戸　稔聡	国立循環器病研究センター・研究推進支援部・部長
	西村　邦宏	国立循環器病研究センター・循環器病統合情報センター・室長
	平松　治彦	国立循環器病研究センター・情報統括部・部長
	上村　幸司	国立循環器病研究センター・研究推進支援部・室長
	辻田　賢一	熊本大学大学院・生命科学研究部・教授
	宇宿　功市郎	熊本大学医学部附属病院・医療情報経営企画部・教授
研究協力者	住田　陽子	国立循環器病研究センター・循環器病統合センター・専門職
	都島　健介	東京大学医学部付属病院・循環器内科・助教
	中村　太志	熊本大学医学部附属病院・医療情報経営企画部・副部長

研究要旨

本研究では、電子カルテの記事情報から自然言語処理を活用して自動的に MACE であると判断するためのシステムを開発し、電子カルテ情報を用いた MACE のビッグデータ分析を行うためのシステムを開発する。日本語で記述される電子カルテからの臨床データベースにおいては初めての試みである。昨年度までに、電子カルテの標準フォーマットである SS-MIX2 の整備を行った。日本循環器学会標準フォーマット（SEAMAT）に基づき、心電図、心エコー、心臓カテーテル検査の結果を SS-MIX2 拡張ストレージに格納する作業を行った。また、国立循環器病研究センターにおいて電子カルテ記事の抽出を行い、電子カルテ記事の自然言語処理を行う準備を行った。本年度（平成 29 年度）は、電子カルテシステム上のデータの確認と、機械学習を用いた症状記載の自動抽出に関する実験、自然言語処理を行う準備である医療用語辞書の作成、SS-MIX2 データからのデータベースを構築と電子カルテ記事の抽出を行った。

A．研究目的

高齢化社会の中にある我が国をはじめとする先進諸国において、循環器疾患が急増している。循環器疾患は再発を繰り返し徐々に進行していくという臨床経過をたどることが多い。例えば、虚血性心疾患では再発・入院を繰り返して終末像として心不全を呈することがしばしばある。そのため循環器疾患においては、Major Adverse Cardiac Event（MACE）とよばれる主要有害心血管イベントを発生させないための再発予防が大事である。循環器疾患

の新規治療法の開発目標として、MACEの発生減少を目標としたものを開発することも考えられるが、MACEを判断するためには担当した臨床医の判断か診療録を読み返し判断するしかない。そのため、レセプト/DPCなどの診療報酬請求情報を使用した分析、または電子カルテ情報を用いてビッグデータの分析においては、MACEなどのイベントをアウトカムにした研究をすることができないという限界がある。本研究では、電子カルテの記事情報から自然言語処理を活用して自動的にMACEであると判断するためのシステムを開発し、電子カルテ情報を用いたMACEのビッグデータ分析を行うためのシステムを開発する。日本語で記述される電子カルテからの臨床データベースにおいては初めての試みである。
日本循環器学会の事業で実施している医療コストがかかる疾患・治療（心筋梗塞・狭心症とその病態に対するステント治療、重症心不全とそれに対する再同期療法（CRT））と循環器領域で特にその重要性が指摘されている疾患（急性心不全など）を抽出し、医療の質とその妥当性を検証するため時間軸を念頭においたデータベースである「臨床効果データベース」を用いて自然言語解析による診断判定システムの構築をおこなうことを目的としている。

B. 研究方法

疾患コホート研究であり、虚血性心疾患、急性心不全の患者を対象とする。対象施設は、国立循環器病研究センター、東京大学、自治医科大学、自治医科大学さいたま医療センター、東北大学、九州大学にてデータの収集を行う。臨床効果データベースから、患者基本情報、診断名、入退院情報、経時的な内服薬、経時的な臨床検査情報、経時的な生理検査情報、経時的な心臓カテーテル検査情報を取得する。別途、電子カルテの記事情報を、臨床効果データベースと同じ匿名化番号にて匿名化したIDにて連結可能匿名化して受け取り、臨床効果データベースのデータと連結を行うことにより、電子カルテの記事情報と臨床データの結合を行う。さらに、入退院情報、検査結果等のデータからMACEによる入院かどうかの判断を行う。これらのMACEのそれぞれのイベントに対して、ICD10を用いて病名のコーディングを行う。次に、電子カルテの記事の医療用語を傷病名、愁訴、身体所見、検査、治療のそれぞれに分類を行う。これらの医療用語の出現とMACEの関係を機械学習（サポートベクターマシンやディープラーニング等）の手法を用いて学習を行い、電子カルテ記事からMACEかどうかを判定する予測モデルを作成する。

（倫理面への配慮）

協力病院からデータを収集する際に、連結可能匿名化とし、個人識別情報および対応表は施設管理者の保管元、施設外に持ち出さないように厳格に管理する。

C. 研究結果

本年度（平成29年度）は、電子カルテシステム上のデータの確認と、機械学習を用いた症状記載の自動抽出に関する実験、自然言語処理を行う準備である医療用語辞書の作成、SS-MIX2データからのデータベースを構築と電子カルテ記事の抽出を行った。

（1）電子カルテシステム上のデータの確認

リアルワールドにおける電子カルテ記事の分析を行うための基盤構築にとっては、電子カルテシステムを中心とする病院情報システムにどのような情報（データ）が保存されているかの把握が必須である。そのため、現行電子カルテシステムで処理されているデータについて、その所在、データ形式について調査した。
国立循環器病研究センター電子カルテ（NEC MegaOak HR）における記述情報（SOAP、退院サマリ）の抽出を行うために、データベースのテーブル構造の把握と実際の抽出作業を実施した。その結果、文字コードの問題などがあったが、必要な記述情報を抽出できることを確認した。この際、記述情報の匿名化についても複数の課題があることがわかり、新たな課題として対応方法を検討することとした。

（2）機械学習を用いた症状記載の自動抽出に関する実験

電子カルテシステム内に蓄積された所見・報告書・サマリなどのテキスト情報から、自然言語処理および機械学習を用いて、カルテ記載内における「症状記載」について、判別・予測する方法論の検討を行った。
具体的には、臨床研究業務担当者が実際に必要とする症状記載データについて、カルテ記載情報から手動で抽出を行った。これらを用いて教師データを作成し、カルテ記載における「症状記載」と「その他の記載」についての自動判別器を作成した。自動判別器は、文章内に出現した各形態素を１次元とした線形サポートベクターマシンを用いて作成した。10分割交差検定を行い評価した結果、本判別器の感度・特異度はともに70〜80%の性能を有していることがわかった。（図1）

出現形態素:13856種類　　10分割交差検定の平均値

		予測値	
		正	負
実際値	正	816	1436
	負	296	7452

Accuracy（正答率）：0.8268±0.0233
False Negative：63.8%（1436/2252）
False Positive：3.8%（296/7748）

図1. 全単語を用いた自動判定結果

（3）自然言語処理を行う準備である医療用語辞書の作成

国立循環器病研究センターにおいて電子カルテ記事の抽出を行い、電子カルテ記事の自然言語処理を行う準備である医療用語辞書の準備を行った。平成 29 年度は、専門医 2 名、統計学者 2 名が国立循環器病研究センターにおいて、電子カルテデータの自然言語処理を行い、医学用語の意味体系（オントロジー）の構築とそれを利用した単語間の相関の度合い（距離等）の利用、形態素解析（名詞、助詞、動詞等の分かち書き）、係り受け解析（主語、述語等の単語間の関係）など文法の解析精度の向上を試みた。約 60 万行のカルテ記事を読み込み、症候の出現頻度を患者ごとに集積し、文章単位での解析が可能なため症候の出現時期、時間が同定可能であった。

さらに自然言語処理技術に関して先進的な IBM ワトソンにより Major Cardiac event をとらえることを目的に辞書チューニングを行った。心筋梗塞レジストリ MIDAS 研究を中心とした約 2000 人の国立循環器病センター入院患者に関して、最も記述が的確と考えられる退院時サマリの記述をもとに虚血性心疾患、心不全、脳卒中、心臓死、全死亡に関して IBM ワトソンエクスプローラーにより抽出を行った。死亡イベントに関しては、電子カルテ上の死亡退院により 100%の把握が可能であった。初回の入院に関しては、入院契機が虚血性心疾患、心不全、脳卒中である場合もほぼ捕捉可能であった。死亡と入院契機の虚血性心疾患、心不全、不整脈項目により心臓死の確認が可能であった。辞書チューニング前は accuracy として 65%前後であるが、、チューニング後は 95%以上の精度達成が可能であった。

（4）SS-MIX2 データからのデータベースを構築と電子カルテ記事の抽出

東北大学、自治医科大学、九州大学では、csv 形式で出力された心電図、心臓超音波検査、心臓カテーテル検査結果を日本循環器学会標準規格である SEAMAT に変換するためのプログラムの実装を行った。（図２）

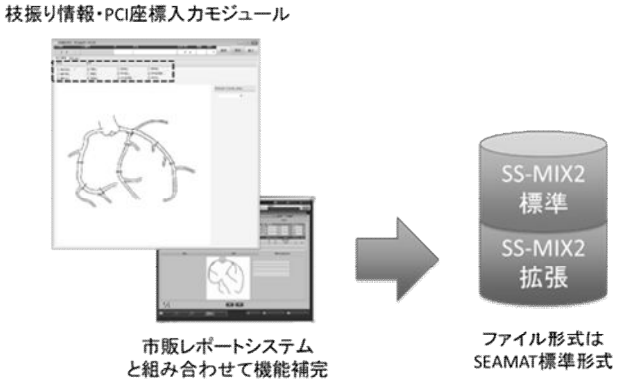

図2. 今後のCAIRS-PCIのデータの流れ

また、関連する学会との意見調整を行うため、SEAMAT 研究会を発足させ、項目の見直しを行った。とりわけ、心臓超音波検査項目に対して心エコー図学会から、より実践的かつ網羅的な提案がなされ、改訂に取り組んでいる。さらに、ISO 取得に向け活動の幅を広げている。東大病院の循環器系生理機能検査データ（心電図、心エコーなど）に関しては日本循環器病学会標準出力フォーマット（SEAMAT 形式）への変換表を作成し、2017 年 11 月より心エコーデータは SS-MIX2 拡張ストレージへ出力が開始されている。また、その他の生理検査は 2018 年 2 月より SS-MIX2 拡張ストレージへ出力開始予定である。熊本大学では、電子カルテからは、データウェアハウス DWH に情報連携、蓄積がなされているが、この DWH から患者基本情報、病名情報、外来受診情報、入退院情報、処方オーダ、注射オーダ、検体検査オーダ、放射線オーダ、検体検査結果、心電図数値データ、心エコー数値データ、心カテ記録、退院サマリ、経過記録に関して SS-MIX2 標準ストレージ、拡張ストレージにデータ出力ならびに提供ができる状況を整えることが出来た。

D. 考察

（1）本研究の目的である非構造化データ（テキスト情報）の自然言語処理や機械学習をするためには、対象とする所見・報告書・サマリなどのテキスト情報の所在・保管形式・データ形式などを把握する必要がある。今回の結果から、様々なシステムで作成・保管され、形式も多様なテキスト情報が、本研究で利用できる形式で抽出・収集可能か検証することができた。また、これらの結果は、他施設における情報抽出・収集においてもフィードバック可能である。そのため、多施設間で大規模なデータを収集する際には有用な知見となりうる。

（2）本研究の最終的な目的は、電子カルテシステム内に蓄積された所見・報告書・サマリなどのテキスト情報から、自然言語処理および機械学習を用いて、Major Adverse Cardiac Event（MACE）とよばれる主要有害心血管イベントを予測するモデルを構築することである。今回行った結果から、自然言語処理を用いた機械学習が症状記載などのイベント判別・予測に有用であることが示せた。

（3）辞書チューニングの過程で抽出された構文からは、看護師、医学部生、研修医程度の精度の症候抽出は可能であり、今後登録研究における省力化、入力の正確性向上に有用と考えられた。

（4）本研究成果により各施設に散在する諸検査結果の収集が可能となり、全国レベルで循環器領域における必須なデータが蓄積しうる。さらに、項目間の違いや表記ぶれ、単位の統一など、データクレンジングに必要な決まりごとを日本における循環器専門医の合意を得て行うため、大規模データを扱う上で大変重要な意義がある。また、現在医療情報分野で課題となっている SS-MIX2 拡張ストレージの充実という点でも、他の学会に先駆けて取り組んでいることは注目に値し、実際問い合わせも増えている。複数病院が参加する共同研究においては標準化した情報の連携を行い、確実な情報の収集が必要であるので今回の成果は大変意義がある。

E. 結論

本研究により、病院情報システムから、SOAP や退院サマリ、種々の検査報告書など、必要な情報を簡便に抽出できる仕組みとして、基幹システムや部門システムのデータを集約・管理できる統合DBの開発が可能となると考えられる。MACE に関連するイベントを精査し、そのイベントの判別に必要な教師データの精度の向上を行えば、機械学習手法によるより最適な予測手法が可能となる。

F. 健康危険情報　　なし

G. 研究発表

1. 論文発表

（1） 平松 治彦:医療情報システムのデータ利用における課題, Jpn Pharmacol Ther (薬理と治療), Vol.45 suppl.2, s76-s78, 2017

（2） 平松 治彦：【改正個人情報保護法】 医学研究編　国際共同研究など外国にある第三者へのデータ提供について注意すること, 医療情報学, 37（5）, 253-5, 2017.

（3） 櫻井理紗、竹村匡正、山口雅和、中井隆史、宍戸稔聡、平松治彦、山本剛、奈良崎大士、上村幸司:ICF を用いた健康情報基盤構築のためのデータ集積手法の検討, 第 37 回医療情報学連合大会論文集, 788-789, 2017

（4） 櫻井理紗、竹村匡正、粂直人、岡本和也、黒田知宏: 我が国における openEHR/アーキタイプを用いた診療データベースの構築可能性の検証, Mumps, vol.28, 15-23, 2017

（5） 山田ひとみ、竹村匡正、桑田成規: 電子カルテの質向上のための診療録監査支援システムの試験的構築, Mumps, vol.28, 3-13, 2017

（6） 山田ひとみ、竹村匡正、岡本和也、黒田知宏、桑田成規: インフォームド・コンセント記載を対象とした診療録監査システムの検討, 日本診療情報管理学会誌 29(1), 53-61, 2017

2. 学会発表

（7） 医療情報連合大会　2017 年 11 月 22 日
日本循環器学会合同シンポジウム
人工知能応用による自然言語処理の活用
電子カルテ情報のセマンティク登録と全国登録事業への将来展望

（8） 第 82 回日本循環器学会学術集会シンポジウム 11（2018 年 3 月 24 日；大阪市）
「わが国の循環器医療提供体制の課題と展望」
The Current Status of Cardiovascular Medicine in Japan; Insights from JROAD and JROAD-DPC Database

（9） Informatics for Health 2017(2017 年 4 月)Poster『Release of the Standard Export Data Format by the Japanese Circulation Society for Standardized Structured Medical Information eXchange Extended Storage』Masaharu Nakayama.

（10） 第 53 回日本小児循環器学会総会・学術集会（2017 年 7 月）教育シンポジウム
『循環器領域におけるビックデータ活用の道標：SS-MIX や日本循環器学会出力標準フォーマット（SEAMAT）について』中山雅晴

（11） 第 37 回日本医療情報学連合大会(2017 年 11 月)共同企画シンポジウム『循環器領域におけるビッグデータ活用の現在』中山雅晴

（12） 第 37 回日本医療情報学連合大会(2017 年

11 月) 一般口演『SS-MIX2 拡張ストレージの充実に向けた取り組み - 日本循環器学会出力標準フォーマット（SEAMAT）について -』中山雅晴、竹花一哉、興梠貴英、IHE-J 循環器

（１３） 日本循環器学会総会（平成 30 年 3 月 25 日、大阪）「臨床効果データベース事業・ImPACT 研究におけるデータ収集の現状」

（１４） 平松治彦:シンポジウム 1「pragmatic clinical trial への誘い」医療情報システムのデータ利用における課題, 日本臨床試験学会第 8 回学術総会

（１５） 櫻井理紗、竹村匡正、山口雅和、松本佳久、本谷崇之、今津貴史、上村幸司、平松治彦、山本剛、奈良崎大士、宍戸稔聡: ICF を用いた個人健康管理システムの構築, 第 44 回日本 M テクノロジー学会大会

（１６） 櫻井理紗、竹村匡正、山口雅和、中井隆史、宍戸稔聡、平松治彦、山本剛、奈良崎大士、上村幸司: ICF を用いた健康情報基盤構築のためのデータ集積手法の検討, 第 37 回医療情報学連合大会

H. 知的財産権の出願・登録状況

1. 特許取得　　なし
2. 実用新案登録　　なし
3. その他　　なし