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））総合研究報告書

「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
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・予防医学・疫学情報部・部長
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
山ノ内　祥訓　熊本大学医学部付属病院・総合臨床研究部・特任教授

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。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。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を行った。さらに、日本循環器学会が事業として行っている「臨床効果データベース」の基盤を用いてデータの収集システムを構築した。「臨床効果データベース」は、個票データを基に精度を重視しており前向きにデータを収集し追跡を行うものである。具体的には、厚生労働省電子的診療情報交換推進事業による標準化ストレージである SS-MIX2 を使用すること、およびレセプト/DPC データを使用することである。SS-MIX2 は標準ストレージに格納された患者基本情報、処方、臨床検査結果のみならず、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も拡張ストレージに

格納することができた。
本研究により、SS-MIX2 データなどの電子カルテ情報を用いてビッグデータの分析において、MACE などのイベントをアウトカムにした分析により、高齢化社会の中で急増する循環器疾患の予後を改善させ医療費を適正化するための医療分析が可能となる。また、この研究成果が電子カルテに実装されるようになれば、診療録の記載や各種検査・投薬オーダを行う際の警告（アラート）の強化や、記載漏れや監査に対応した病名推薦システムを構築することができる。

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。

また、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」を構築している。

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。

B．研究方法

本研究では、日本循環器学会が事業として行っている「臨床効果データベース」の基盤を用いてデータの収集を行う。「臨床効果データベース」は、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、重症心不全とその病態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を策定することを目的としている。その特徴としては、個票データを基に精度を重視しており前向きにデータを収集し追跡を行うこと、また、厚生労働省電子的診療情報交換推進事業による標準化ストレージである SS-MIX2 を使用すること、およびレセプト/DPC データを使用することである。SS-MIX2 は標準ストレージに格納された患者基本情報、処方、臨床検査結果のみならず、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も拡張ストレージに格納する。データの収集は、東京大学が共同開発した多目的臨床データ登録システム（Multi-purpose Clinical Data Repository System: MCDRS）を用いて行う。

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。平成 28 年度は、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である医療用語辞書の準備を行った。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る。さらに、SS-MIX2 データを用いてビッグデータにおける機械学習（サポートベクターマシンやディ

ーブラーニング等)・ベイズ統計学を利用し、MACE の自動判定システムの構築をおこなった。

（倫理面への配慮）
全ての情報は、匿名化し、個人識別情報を消去して解析をおこなう。

C. 研究結果
（1）機械学習を用いた症状記載の自動抽出に関する検討
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

電子カルテシステムにおける SOAP 記載が、病態の特徴を現しているという仮説のもとに、自然言語処理を用いて SOAP 記載内容と医師が付与した病名の関連を学習し、これら機会学習によって病名予測を試みた。病名としては DPC/PDPS における様式４を用いて主病名、ICD-10 の予測を試みた。対象データは心疾患とした。結果、主病名 23 病名付与（23 クラス分類）ICD-10 14 付与（14 クラス分類）について、それぞれ正答率 32.5%, 44,5%であった。しかし、「心房細動」と「発作性心房細動」など付与された病名自体が排他的な分類と言えないこともあり、総じて正確に病態を判定できることが明らかとなった。

（2）人口知能（AI）を活用した循環器疾患の登録システムの整備に関する研究
自然言語処理技術に関して先進的な IBM Watson（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。最終的に、ピナクルレジストリの項目中、完全自由記載の 5%を除く 95%の項目を抽出することに成功した。

（３）自然言語処理を含む機械学習に供するための標準データを電子カルテから抽出するための研究
心臓カテーテル検査を受けたことがある患者約 3000 名をデータ抽出対象とした。電子カルテから抽出した処方データについては当初想定した通りのデータが抽出できた。血液検査値データについては LDL-C 等のデータに一部欠測が認められたため、SS-MIX2 抽出システムを見直し、改めてデータを抽出した。
心エコーデータについては CSV データから SEAMAT 形式に変換して SS-MIX2 拡張ストレージに出力できた。心臓カテーテル検査レポートデータについても CAIRS-DB から CAIRS フォーマットで出力したデータを SEAMAT 形式に変換できた。カルテテキストデータについても電子カルテデータベースから抽出できた。
さらに SS-MIX2 ストレージに格納された各種データを SS-MIX2 agent を用いて抽出することにも成功した。

（４）データ転送プログラムによるデータ収集に関する研究
日本循環器学会標準出力フォーマット（Standard Export data forMAT：SEAMAT）を用いて、厚労省標準保存形式である SS-MIX2 の拡張ストレージに循環器特有の検査結果を転送し、データを 2 次活用するための基盤システムを

整備した。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）

日本循環器学会の他、日本医療情報学会、日本心不全学会、日本不整脈心電学会、心エコー図学会、日本心血管インターベンション治療学会、日本心臓核医学会、心臓リハビリテーション学会が参加する SEAMAT 研究会により項目の改訂や対象検査範囲の拡大を検討した。また、SS-MIX2 agent を設置した施設を増やし、データ収集の規模を拡大している。

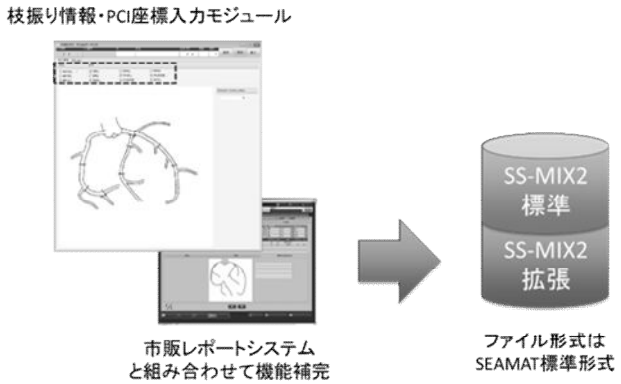

図2. 今後のCAIRS-PCIのデータの流れ

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。ワトソンなどの自然言語処理による自動入力システムの構築は、登録コストの引き下げにつながる可能性がある。疾患レジストリーの構築には通常数千万から臨床治験など数億円が入力、データ管理に必要となりつつあり、基本的な臨床情報を抽出し、さらに臨床試験への対象に合致するか否かの case finding などにも応用が可能な技術と考えられられる。

（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
7. Architecture of the Japan Ischemic Heart Disease Multimodal Prospective Data Acquisition for Precision Treatment (J-IMPACT) System. Tetsuya Matoba, Takahide Kohro, Hideo Fujita, Masaharu Nakayama, Arihiro Kiyosue, Yoshihiro Miyamoto, Kunihiro Nishimura, Hideki Hashimoto, Yasuaki Antoku, Naoki Nakashima, Kazuhiko Ohe, Hisao Ogawa, Hiroyuki Tsutsui, Ryozo Nagai. International Heart Journal. 2019; 60(2): 264-270.

2. 学会発表

1. 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
2. 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
3. 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.
4. 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
5. 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
6. 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
7. 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」
8. 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
9. 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
10. 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
11. Medical informatics Europe 2018 (Apr.24-26, 2018, Gothernburg Sweden)Poster「Five-Year Experience of a Medical Information Network System. 」Oral「Implementation and effect of a novel Electronic Medical Record format for patient allergy information.」Masaharu Nakayama.
12. Healthcare Leadership Conference at InterSystems Global Summit 2018（Sept.30-Oct.3, 2018, San Antonio, USA）Panelist「How Regional HIEs Connect the Health Ecosystem 」Masaharu Nakayama.
13. AMIA 2018 Annual Symposium（Nov. 2-7 San Francisco, USA ）Poster「Development of a Standardized Data Format in Cardiology throughCollaborations between Medical Informaticiansand Cardiologists. 」

Masaharu Nakayama.

14. 第 93 回日本医療機器学会大会（5 月 31 日-6 月 2 日、2018 年、横浜）シンポジウム招待「日本循環器学会データ出力標準フォーマット（SEAMAT）について」中山雅晴
15. 第 66 回日本心臓病学会学術集会（9 月 7-9 日、2018 年、大阪）シンポジウム「電子カルテ情報の活用･･･SS-MIX2 ストレージおよび MIDNET と SEAMAT について」中山雅晴
16. 第 38 回医療情報学連合大会（第 19 回日本医療情報学会学術大会）（11 月 22-25 日、2018 年、福岡）日本循環器学会共同企画「SEAMAT でできること、導入のためにすべきこと」中山雅晴
17. 第 83 回日本循環器学会学術集会（3 月 29-31 日、2019 年、横浜）会長特別企画 20「Secondary Use of Clinical Data from Hospital Information Systems 」Masaharu Nakayama.
18. 的場哲哉、興梠貴英、藤田英雄、中山雅晴、清末有宏、橋本英樹、大江和彦、宮本恵宏、西村邦宏、小川久雄、安徳恭彰、中島直樹、筒井裕之、永井良三.「心臓カテーテルを中心とした多モダリティ循環器診療情報を収集する J-IMPACT システム」第 38 回医療情報学連合大会（第 19 回日本医療情報学会学術大会、平成 30 年 11 月 25 日、福岡）

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