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　６ 年　 ９ 月 １４ 日現在

機関番号：１２３０１
研究種目：若手研究
研究期間：2020～2023
課題番号：２０Ｋ１８８７４

研究課題名（和文）新しい自然言語処理手法を用いた電子カルテデータ構造化と疾患分類AIモデルの構築

研究課題名（英文）Data structuring of electronic medical records and development of artificial intelligence-based model for disease diagnostic support using a novel natural language processing technology

研究代表者

野口　怜（Noguchi, Rei）

群馬大学・医学部附属病院・助教

研究者番号：５０８２８８６１

交付決定額（研究期間全体）：（直接経費）　　3,200,000円

研究成果の概要（和文）：本研究は、電子カルテのテキストデータを活用し、診療録の記述に基づいて疾患名の診断支援を行うAI構築を目指すものである。
本研究により、自然言語処理技術を活用して電子カルテのテキストデータから疾患名・症状名を抽出し、症例ごとの疾患・症状の構造化データ（症例マトリクス）を自動生成する方法論を確立することができた。また、症例マトリクスを学習データとして、特定の循環器疾患を最大で再現率87%で検出可能な疾患分類の機械学習モデル（疾患分類AIモデル）を構築することができた。

研究成果の学術的意義や社会的意義
真の診断支援AI の構築には、電子カルテのテキストデータの活用が不可欠であるが、非構造化データのために扱いが難しくまだ十分に活用されていない。本研究は既存の電子カルテデータの活用可能性を広げるとともに、将来的な診断支援AI の実現に向けたコア技術となるものであり、医療の質向上や均てん化、医師の負担軽減に大きく貢献できると考えられる。

研究成果の概要（英文）：This study aims to build an AI that utilises text data from electronic medical records to provide diagnostic support for disease names based on descriptions in medical records.
Through this research, a methodology was established to extract disease and symptom names from text data in electronic medical records using natural language processing technology, and to automatically generate structured data (case matrix) of diseases and symptoms for each case. In addition, using the case matrix as training data, a machine learning model for disease classification (disease classification AI model) that can detect specific cardiovascular diseases with a maximum reproduction rate of 87% could be constructed.

研究分野：医療情報学

キーワード：電子カルテデータ　自然言語処理　非構造化データ　診断支援AI　疾患判別モデル　症例マトリクス　テキストデータ構造化　類似症例予測AI

１．研究開始当初の背景

昨今の「データ革命」の機運の高まりにより、医療分野でも医療の質・安全の追求のため、診療での AI 活用が大きく期待されている。近年、検査画像への AI 活用は著しい進歩が見られるが[1]、真の診断支援は、検査の前、つまり問診や診察段階における診断仮説の構築時にこそ大きな便益がある。このためには、診断の 76%は病歴から得られるという報告[2]からも、患者の病歴や経過などが一貫して詳述された電子カルテの「テキストデータ」の活用が成否を握る。しかしながら、自然言語処理によるテキスト前処理がボトルネックとなっており、特に、従来の形態素解析による前処理は、「辞書」の作成・整備に多大な工数がかかり活用の大きな妨げとなっている。また、非構造化データのまま処理されるため、アドホックなテキスト分析に留まり、疾患の予測モデル構築には繋げられていない。

そこで本研究では、電子カルテのテキストデータの診療活用を実現するために、辞書不要の新しい自然言語処理技術（InterSystems 社 IRIS NLP）を活用して、極めて効率的なテキスト前処理手法およびデータの構造化手法を開発するとともに、実用レベルの精度を持った疾患分類の機械学習モデルの構築を目指した。

２．研究の目的

本研究では、「診断支援 AI に向けた電子カルテテキストデータ構造化手法と疾患分類 AI モデル化手法の構築」を目的とし、下記 2 つの研究を行った。

① 電子カルテのテキストデータから、最新の自然言語処理手法を用いて患者別の疾患名・症状名を抽出し、自動的にマトリクス形式（**症例マトリクス**）に構造化する手法の確立

② 症例マトリクスを学習データとして、実用レベルの精度を持った疾患分類の機械学習モデル（**疾患分類 AI モデル**）を構築する手法の確立

一般に医師の診断精度は単独で 60%程度、複数人では 80%超とされるため[3]、本研究ではモデルが「もう一人の医師の目」としての役割を担いうる 60%の精度を目指した。

３．研究の方法

本研究では、電子カルテデータとして、比較的可読性の高い退院サマリを用い、①退院サマリの構造化、②症例マトリクス構築、③疾患分類 AI モデル構築の流れで進めた（図 1）。なお、モデルの確からしさは、見逃しに強いモデルを目指すため「再現率（recall）」を重視して評価するとともに、診断ガイドラインや医師の意見などをもとに医学的見地からの妥当性も評価した。

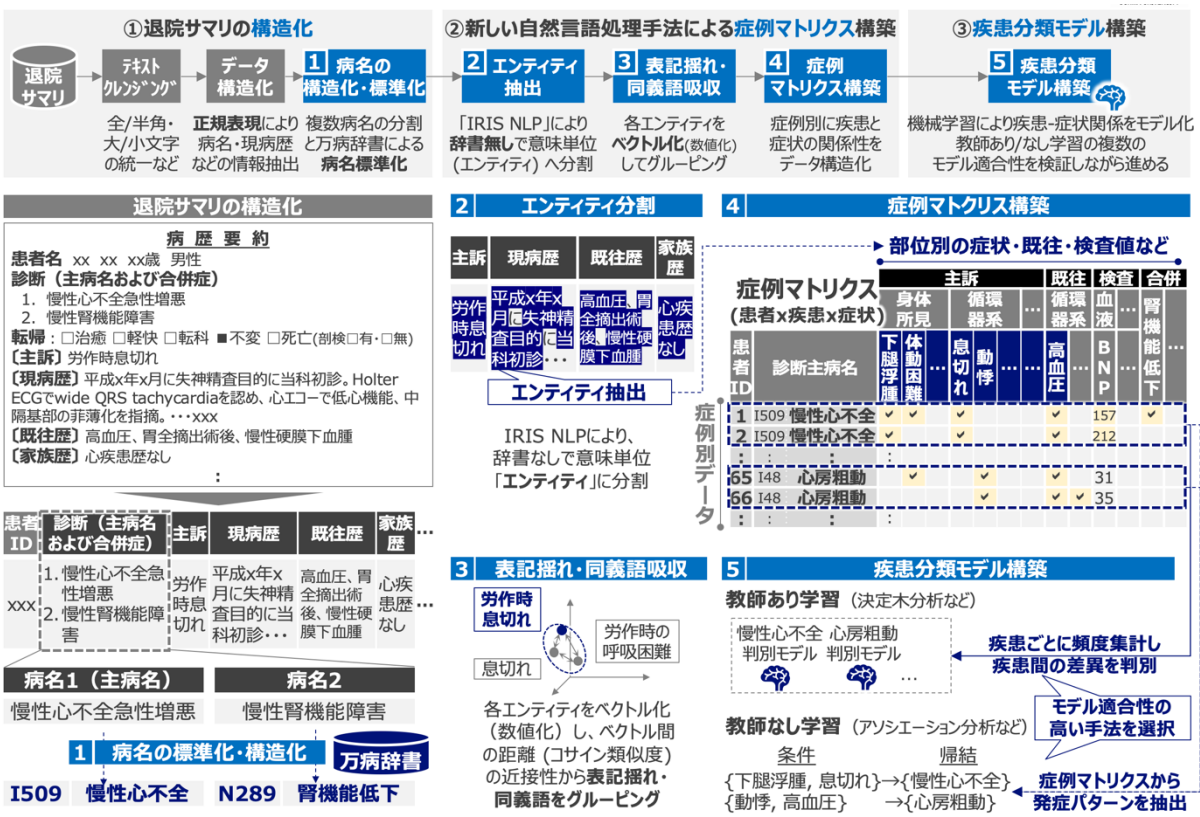

図1 本研究の全体イメージ

４．研究成果

①退院サマリの構造化
群馬大学医学部附属病院院の電子化された 10 年分の退院サマリ（見込み約 100,000 件）から正規表現（文字列のパターン抽出を行うための手法）により各セクション（病名・現病歴・既往歴など）のテキスト記述をセクション別に抽出・構造化する自動化手法を構築した。この際、疾患名も自由テキストで記述されているため、正規表現により複数病名は分割した上で、標準病名とカルテ内での出現名とを紐付けた「万病辞書」[4]を活用して病名の標準化を行うことで、病名の表記揺れを集約し、より精緻な構造化データを抽出することができた。

②症例マトリクスの構築
抽出データをセクション別に自然言語処理エンジン IRIS NLP に入力して「エンティティ」と呼ばれる意味単位に自動分割した後、各エンティティを列方向に構造化して症例ごとに出現有無を整理することで、症例マトリクスを構築した。循環器内科などの 10 年分の退院サマリに対して症例マトリクスを構築し、疾患別に頻度集計をとったところ、疾患特異的なエンティティ（特徴語）が抽出され、本アプローチの有効性が示された。

③疾患分類 AI モデルの構築
疾患判別および、類似症例予測の機械学習モデル構築を試み、いずれも一定の精度を持つモデルが得られ、本研究のアプローチの有用性が示された。

■疾患判別モデルの構築：当院 10 年分の退院時サマリーから抽出された症例マトリクスを活用し、主病名が主要な循環器疾患 8 種である症例に対して、症例マトリクス内の疾患名と特徴語（当該症例の退院サマリーより抽出された主に症状などの疾患特異的な単語）との関係性を機械学習手法（ランダムフォレスト）を用いて学習させた。学習データに対しては、特定の循環器疾患を最大で再現率 87%で検出可能なモデルを構築できた。

■類似症例予測モデルの構築：症例マトリクスに対して、「疾患別に頻出症状パターンを抽出する問題」と捉え、教師なし学習のアソシエーション分析（顧客の頻出購買パターン分析などに用いられる、大量データから項目間の共起性により関係性を見つけ出す手法）を適用し、疾患−症状関係のモデル化を試みた（出力例：{条件：動悸，高血圧}→{帰結：心房粗動}）。疾患判別モデルと同様に、主病名を循環器疾患 8 種に限定した上で、複数回入院患者を除いた約 1,000 例分の症例マトリクスを学習データとした。この症例マトリクスに対して、アソシエーション分析を適用し、類似症例予測のフレームワークを構築した。入力となる 1 症例分の単語集計結果（＝症例ベクトル）と、症例マトリクス内の各症例ベクトルとの相関係数を算出し、症例マトリクスの中から類似の症例を抽出する構成とした。性能評価として、症例マトリクスから 1 症例のみを抽出し、これを未知の症例ベクトルとして与えて、残りの症例から 類似症例を抽出する実験を、全症例に対して逐次行った。結果として、約 3 割の症例で、基準（相関係数 0.3 以上かつ、一致した特徴語数が 3 語以上）を満たす類 似症例が抽出され、約 1 割の症例では、類似症例の主病名が入力症例の主病名と一致した。

今後はより実運用に近づけるよう、モデルチューニングおよび、モデル汎用性の向上に注力していく。

＜引用文献＞
[1] Kudo SE., et al., Clin Gastroenterol Hepatol., S1542-3565(19): 30997-8 (2019)
[2] M. C. Peterson et al., West J Med, 156(2): 163–165. (1992)
[3] Barnett ML., et al., JAMA Netw Open., 2(3):e190096 (2019)
[4] Aramaki E., et al. MEDINFO2017, Vol. 217 (2017)

５．主な発表論文等

〔雑誌論文〕　計7件（うち査読付論文　6件／うち国際共著　0件／うちオープンアクセス　1件）

１．著者名 野口 怜， 鳥飼 幸太， 齋藤 勇一郎	４．巻 Vol.42 (Suppl.)
２．論文標題 電子カルテテキストから構築した症例マトリクスによる協調フィルタリングベースの類似症例予測	５．発行年 2022年
３．雑誌名 医療情報学	６．最初と最後の頁 908-911
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 －

１．著者名 鳥飼 幸太，野口 怜，松山 龍之介，齋藤 勇一郎	４．巻 Vol.42 (Suppl.)
２．論文標題 HL7 FHIR / Web アークテクチャ / 医療サイバーセキュリティ / 稼働継続性を重視した Virtual Fault Tolerance (VFT)コンセプトに基づく 病院情報システムの設計と実装	５．発行年 2022年
３．雑誌名 医療情報学	６．最初と最後の頁 741-742
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 －

１．著者名 Rei Noguchi	４．巻 1
２．論文標題 GunNLP at the NTCIR-16 Real-MedNLP Task: Collaborative Filtering-based Similar Case Identification Method via structured data “Case Matrix”	５．発行年 2022年
３．雑誌名 Proceedings of the 16th NTCIR Conference on Evaluation of Information Access Technologies	６．最初と最後の頁 349-352
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 オープンアクセスとしている（また、その予定である）	国際共著 －

１．著者名 野口 怜， 鳥飼 幸太， 齋藤 勇一郎	４．巻 Vol.41 (Suppl.)
２．論文標題 電子カルテテキストから構築した症例マトリクスによる協調フィルタリングベースの類似症例予測	５．発行年 2021年
３．雑誌名 医療情報学	６．最初と最後の頁 928-931
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 －

１．著者名 鳥飼 幸太，野口 怜，松山 龍之介，白戸 悠貴，齋藤 勇一郎	４．巻 Vol.41 (Suppl.)
２．論文標題 e-Path 3層構造とFHIRを擁する医療ロジックの共通化モデルに適したIn-Process Clinical Intelligence(IPCI)の設計と実装	５．発行年 2021年
３．雑誌名 医療情報学	６．最初と最後の頁 858-860
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 －

１．著者名 野口 怜 ， 鳥飼 幸太 ， 齋藤 勇一郎	４．巻 Vol.40 (Suppl.)
２．論文標題 オープンデータを活用したCOVID-19罹患者における症状・経過のテキスト分析	５．発行年 2020年
３．雑誌名 医療情報学	６．最初と最後の頁 504-509
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 －

１．著者名 野口 怜 ， 鳥飼 幸太 ， 齋藤 勇一郎	４．巻 48
２．論文標題 電子カルテデータ構造化による症例マトリクスと疾患判別モデルの臨床導入意義	５．発行年 2020年
３．雑誌名 月刊新医療	６．最初と最後の頁 45-49
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 無
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9件（うち招待講演　0件／うち国際学会　1件）

１．発表者名 野口 怜， 鳥飼 幸太， 齋藤 勇一郎
２．発表標題 電子カルテテキストから構築した症例マトリクスによる協調フィルタリングベースの類似症例予測
３．学会等名 第42回医療情報学連合大会
４．発表年 2022年

１．発表者名
鳥飼 幸太，野口 怜，松山 龍之介，齋藤 勇一郎

２．発表標題
HL7 FHIR / Web アークテクチャ / 医療サイバーセキュリティ / 稼働継続性を重視した Virtual Fault Tolerance (VFT)コンセプトに基づく 病院情報システムの設計と実装

３．学会等名
第42回医療情報学連合大会

４．発表年
2022年

１．発表者名
Rei Noguchi

２．発表標題
GunNLP at the NTCIR-16 Real-MedNLP Task: Collaborative Filtering-based Similar Case Identification Method via structured data “Case Matrix”

３．学会等名
The 16th NTCIR Conference（国際学会）

４．発表年
2022年

１．発表者名
野口 怜， 鳥飼 幸太， 齋藤 勇一郎

２．発表標題
退院サマリから構築した症例マトリクス学習モデルと経過記録テキストからの疾患判別予測

３．学会等名
第3回 日本メディカルAI学会学術集会

４．発表年
2021年

１．発表者名
野口 怜， 鳥飼 幸太， 齋藤 勇一郎

２．発表標題
電子カルテテキストから構築した症例マトリクスによる協調フィルタリングベースの類似症例予測

３．学会等名
第41回医療情報学連合大会

４．発表年
2021年

１．発表者名
鳥飼 幸太，野口 怜，松山 龍之介，白戸 悠貴，齋藤 勇一郎

２．発表標題
e-Path 3層構造とFHIRを擁する医療ロジックの共通化モデルに適したIn-Process Clinical Intelligence(IPCI)の設計と実装

３．学会等名
第41回医療情報学連合大会

４．発表年
2021年

１．発表者名
野口 怜 ， 鳥飼 幸太 ， 齋藤 勇一郎

２．発表標題
電子カルテデータを構造化した症例マトリクスによる疾患判別モデルの構築と表記揺れ集約の効果

３．学会等名
第24回日本医療情報学会春季学術大会

４．発表年
2020年

１．発表者名
野口 怜 ， 鳥飼 幸太 ， 齋藤 勇一郎

２．発表標題
診療記録の一次利用に向けたテキスト前処理フレームワークおよび症例マトリクスの提案と評価

３．学会等名
第48回日本Mテクノロジー学会大会

４．発表年
2020年

１．発表者名
野口 怜 ， 鳥飼 幸太 ， 齋藤 勇一郎

２．発表標題
オープンデータを活用したCOVID-19罹患者における症状・経過のテキスト分析

３．学会等名
第40回医療情報学連合大会

４．発表年
2020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

７．科研費を使用して開催した国際研究集会

〔国際研究集会〕　計0件

８．本研究に関連して実施した国際共同研究の実施状況

共同研究相手国	相手方研究機関