ポスター発表
[PB] ポスター B

[PB-02] 診療リアルワールドデータを用いた癌患者情報抽出手法の開発と精度評価
Development and Accuracy Evaluation of Cancer Patient Information Extraction Method Using Clinical Real-World Data

○串間 宗夫[1]、長谷川 義行[2]、野末 卓[2]、岡崎 絵美[2]、古賀 久芳[2]、小川 泰右[1]、荒木 賢二[1]（1.宮崎大学医学部附属病院、2.株式会社NTTデータ）

○Muneo Kushima[1], Yoshiyuki Hasegawa[2], Suguru Nozue[2], Emi Okazaki[2], Hisayoshi Koga[2], Taisuke Ogawa[1], Kenji Araki[1]（1.University of Miyazaki Hospital, 2.NTT Data Corporation）

診療リアルワールドデータを用いた癌患者情報抽出手法の開発と精度評価

串間 宗夫[*1], 長谷川 義行[*2], 野末 卓[*2], 岡崎 絵美[*2], 古賀 久芳[*3], 小川 泰右[*1], 荒木 賢二[*1]

[*1] 宮崎大学医学部附属病院 病院 IR 部

[*2] 株式会社 NTT データ 製造 IT イノベーション事業本部

[*3] 株式会社 NTT データ 数理システムデータマイニング部

Development and Accuracy Evaluation of Cancer Patient Information Extraction Method Using Clinical Real-World Data

Muneo Kushima[*1], Yoshiyuki Hasegawa[*2], Suguru Nozue[*2], Emi Okazaki[*2], Hisayoshi Koga[*3], Taisuke Ogawa[*1], Kenji Araki[*1]

[*1] Institutional Research, University of Miyazaki Hospital

[*2] Manufacturing IT Innovation Sector, NTT Data Corporation

[*3] Data Mining Division, NTT Data Corporation

抄録:[目的] 本研究は、製薬企業のニーズが強い遺伝子検査結果や癌のステージ分類・副作用などの患者情報の抽出技術を確立することで、患者情報の二次利用に向けた抽出技術実現可能性検証を行う。[方法] 宮崎大学医学部附属病院電子カルテデータが保存されているデータウェアハウスの診療リアルワールドデータの中から癌患者データを用い、癌患者情報抽出手法として、特徴の異なる、ルールベース(正規表現マッチ)と機械学習(系列ラベリング)の 2 種類の手法を開発し精度評価を行なった。[結果] ルールベースでは、人によるフラグ確認を踏まえると 100%の正解率で患者情報を抽出できている。一方、機械学習の精度評価としては、電子カルテ全文を対象とし、教師データ、検証データを作成した結果、99.74%の正解率で患者情報が抽出できている。[結論] ルールベースと機械学習のより一層の精度評価とその課題解決が必要であるが、想定した癌患者情報が抽出できることがわかった。

キーワード 診療リアルワールドデータ、電子カルテ、ルールベース(正規表現)、機械学習(系列ラベリング)。

1. 目的

製薬企業へのヒアリングを通して、遺伝子検査結果や癌のステージ分類・副作用などの患者情報の抽出に強いニーズがあることが判明している。また、電子カルテのテキスト(診療録など)に患者情報が記述されていることは確認できたが、そのデータの抽出技術が確立しておらず、医療分野の研究開発に活用できていない現状がある。更に、日々の診療を通して蓄積された医療情報(診療リアルワールドデータ)の二次利用に向けた実現可能性検証が十分にされているとはいい難い。本研究は、製薬企業のニーズが強い患者情報を抽出する手法の実現性と有効性を評価することで、当該情報の抽出技術を確立し、二次利用に向けた実現可能性検証を行う。

2. 方法

患者情報を抽出する手法として、特徴の異なる、ルールベース(正規表現マッチ)と機械学習(系列ラベリング)の 2 種類の手法[1]の実現性・有効性を評価する。

1) ルールベース(正規表現マッチ)、概要:人手でルールを書いて処理する、特徴:処理がわかりやすい、処理イメージ例:HER2(+)として、【ルール適用(例)】、([Hh][Ee][Rr][2])[¥t]*(¥(陽性|¥(?¥+¥)?¥))
"Her2"や"HER2"は OK、後ろにスペース・タブがいくつあっても OK、"陽性"もしくは"(+)"や"+"も OK、¥1 と¥2 で"HER2"と"(+)"のペアを取得、新たなルールに弱い(メンテが必要)。

2) 機械学習(系列ラベリング)、概要:教師データを用意し、そこから正解を導くルールを導出する、特徴:総合的に判断するため、表記ゆれや新たなルールにある程度対応できる、処理イメージ例:HER2(+)として、【ルール適用結果(例)】
HER2…検査項目開始、(…検査結果開始、+ …検査結果途中、) …検査結果途中、上から見て"HER2"と"(+)"のペアを取得。

3) 患者情報の抽出方法
「患者・文書の特定・抽出」→「患者情報記載箇所の特定・抽出」→「患者情報の意味付け」→「患者情報の特定・抽出」の 4 段階で抽出する。

(1)患者・文書の特定・抽出:分析対象患者の 2018 年 01-03 月に記載された文書「入院時サマリ」「経過記録」「退院時サマリ」を特定・抽出する。

(2)患者情報記載箇所の特定・抽出 ： 患者情報「病名」「病期分類」「TNM 分類」「遺伝子検査結果」が記載されている箇所を特定・抽出する。
(3)患者情報の意味付け ： 患者情報記載箇所に対し、表記ゆれを考慮しつつ、単語と患者情報を対応付ける。
(4)患者情報の特定・抽出 ： 意味付け情報を基に、2種類の手法「ルールベース」「機械学習」を用いて、患者情報を特定・抽出する。

収集する患者情報は、電子カルテおよび医事システムから、患者属性情報：誕生年、性別、疾病情報：疾病名、入院外来区分、診療基本情報：入退院履歴、外来受診履歴、処方情報：処方日、入院・外来区分、処方医 ID、薬剤名、数量、単位、処方期間、診療行為情報：手術歴情報、検査結果、観察項目、文書情報：経過記録、看護記録、退院時サマリ、手術記録、についての医療情報を抽出し分析対象としている。

3. 結果

正規表現は、2018 年 2 月の電子カルテ（入院時サマリ、退院時サマリ、経過記録、経過記録（SOAP））をもとに正規表現を作成し、2018 年 3 月の退院時サマリで検証した。一方、機械学習では、2018 年 2 月の電子カルテ(正規表現と同様)をもとに教師データを作成し、2018 年 3 月の退院時サマリで検証した。抽出誤り削減のため、形態素解析で区切るパターンを増加した影響から、プログラム単体の抽出精度が落ちているが、人によるフラグ確認を踏まえると 100%の患者情報が抽出できている。抽出誤りの改善はされているが、2018 年 3 月には、 2018 年 1 月、2 月に出現していない患者情報が多く出現し、抽出漏れが多かったことが原因だと考えられる。精度向上には、網羅性の高い正規表現パターンを用意する必要がある。

機械学習では、電子カルテ全文を対象とした教師データ、検証データを作成し、99.74％の患者情報が抽出できている。原因として、電子カルテの全体の9割以上は抽出対象外の文字であり、学習モデルでは、基本的には、対象外と認識されていると考えられる。抽出対象の患者情報のうち、病名とTNM分類は正しく認識割合が多い。割合が高い理由として、stage や検査名の比較し、文字数が多いことや病名の場合は漢字が連続しており、TNM の場合はアルファベットと数字が交互に存在するなど、他の文字と比べ特徴が判断しやすいことが考えられる。stage や検査名の特徴を学習できるようモデルを修正することが精度向上の鍵である。図 1,2 に、アウトプットイメージを示す。

4. 考察

課題や解決方針を以下に示す。
(1)問題リストの途中での改行など、抽出ロジックが対応しておらず、患者情報記載箇所の抽出にモレが生じている、モレを定量的に把握できていない。
(2) 病名や遺伝子検査項目の表記ゆれ・複数出現など、ルール（ルールベース）・教師データ（機械学習）・抽出ロジックが不十分であり、患者情報の抽出にモレが生じている、モレを定量的に把握できていない。
(3) 現状、機械学習は遺伝子検査の ALK のみを対象としており、対象の拡大が必須。
(4)課題解決を踏まえた実現性や精度・運用の観点での有効性を評価できていない。

課題解決方針として、モレを定量的に計測し、モレが生じる箇所・規模・原因を特定する。効果的な手段を見極める。辞書・ルールベース・教師データ（機械学習）・抽出ロジックなどをチューニングする。等が必要である。

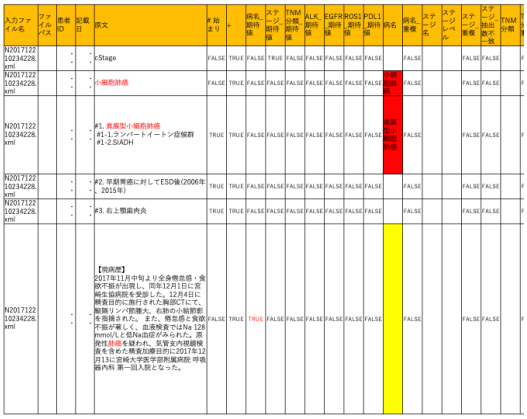

図 1 正規表現アウトプットイメージ

データ	正解値	予測結果	DN	IN	IR
#1	NO	NO	0.000349	2.13E-05	8.03E-05
異所性ACTH産生小細胞肺癌	DN	DN	0.992263	0.000119	5.39E-06
#1-1	NO	NO	0.000452	0.000212	0.001519
多発リンパ節転移	OTHER	OTHER	0.000637	5.86E-05	4.27E-05
右深頚リンパ節	OTHER	OTHER	0.001835	0.000169	3.58E-05
両側鎖骨上リンパ節	OTHER	OTHER	0.001293	0.000209	4.93E-05
縦隔リンパ節	OTHER	OTHER	0.000414	0.00023	3.18E-05
右肺門リンパ節	OTHER	OTHER	0.000545	0.000731	0.000195
#1-2	NO	NO	7.74E-05	0.000119	3.78E-05
多発肝転移	OTHER	OTHER	0.002649	6.54E-05	0.000844
#1-3	NO	NO	1.85E-05	5.20E-06	5.10E-06
心外膜転移の疑い	OTHER	OTHER	0.00264	0.0001	4.26E-05
#1-4	NO	NO	4.16E-06	2.80E-06	3.86E-06
転移性脳腫瘍疑い	OTHER	OTHER	0.005868	4.12E-05	2.99E-05

図 2 機械学習アウトプットイメージ

5. 結語

宮崎大学医学部附属病院電子カルテデータが保存されているデータウェアハウスの診療リアルワールドデータの中から癌患者データを用い、癌患者情報抽出手法として、特徴の異なる、ルールベース（正規表現マッチ）と機械学習（系列ラベリング）の 2 種類の手法を開発し精度評価を行なった。

正規表現と機械学習のより一層の精度評価とその課題解決が必要であるが、想定した癌患者情報が抽出できることがわかった。

参 考 文 献

[1] Muneo Kushima, Ryosuke Matsuo, Taisuke Ogawa, Yoshiyuki Hasegawa, Suguru Nozue, Emi Okazaki, Hisayoshi Koga, Kenji Araki, Development of Patient Information Extraction Method by Sequence Labeling using Electronic Medical Records, IEEE International Symposium on Multiple-Valued Logic, Miyazaki Japan, 2020.