平成３０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 基盤構築・人工知能実装研究事業））
分担研究報告書

機械学習を用いた症状記載の自動抽出に関する検討

研究分担者　竹村　匡正　兵庫県立大学大学院・応用情報科学研究科・教授
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
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

研究要旨
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。

A．研究目的

本研究では、これまで電子カルテシステムにおける SOAP 記載に対して、自然言語処理および機械学習技術を用いた自動的な内容把握に関する研究を進めてきた。これは、有害事象の自動抽出システムの構築として前年度に報告し、学会等への報告を行った。本年度は、これらの分析を踏まえた上で自動的な病名判定に関する研究を行った。

B．研究方法

1．データ抽出

国立循環器病研究センターでは、情報セキュリティの確保、診療情報の安全な利用のために 4 階層ネットワークの環境を構成している。第 1 層はインターネットに公開可能なデータ、第 2 層はセンター内に公開可能なデータ、第 3 層は研究に利用するデータ、第 4 層は診療のための電子カルテデータが格納される。第 4 層が病院情報システムのネットワークとなり、この第 4 層の最高機密の領域に DWH(Data Ware House)が置かれ、この DWH から臨床研究等に利用する診療情報を抽出することとなっている。

本研究の対象データについては、国立循環器病研究センターで稼働中の電子カルテシステムから記述情報（SOAP）および医事会計システムから DPC 制度における DPC データの様式 1 データをそれぞれの DWH を用いて抽出した。また、SOAP 記載データ自体は DWH に蓄積されていたものの、今回これらのデータを大規模に抽出することは、国立循環器病研究センターにおいても初めての試みであったため、SOAP データ抽出基盤を新たに構築した。

また、第 3 層、第 4 層のデータの抽出時には、データは原則匿名化を行う必要があるが、SOAP 記載の中には患者 ID や氏名が含まれている。そのため、**抽出した SOAP 記載情報は患者 ID や氏名について匿名化処理の必要がある。そこで、自然言語処理による匿名化を試みた。匿名化処理は、本研究の別の正解のひとつである IBM WATSON を用いた研究において構築した匿名化ツールを用いて**

行った。

最終的には、各入院毎に病名データを付与する必要があることから、患者 ID(DPCID)、入院日、退院日で突合を行い、1 患者 1 入院あたりの SOAP 記載に対応した病名の付与を行った。

2. 実験

実際に用いるデータとしては、国立循環器病研究センター病院において作成された、DPC 制度における様式1データ、および SOAP 記載データであり、国立循環器病研究センター倫理委員会の承認を得た。期間は、2012 年 9 月-2014 年 12 月であり、本助成で対象となっている MIDAS 研究の対象患者であり、患者数は 1248 人であった。SOAP 記載は 112,924 個であった。一方、国立循環器病研究センター病院において、本期間において入院 DPC ファイルより得られた入院数（主病名）は 23,374 個あり、入院患者数は 16,604 人であった。診断名自体は 1,885 種類であり、ICD10 コードとしては 785 種類であった。SOAP データと様式1の病名（入院）データと突合したところ、その結果、1074 入院の病名データと SOAP 記載データを突合することができた。そのうち、診断名は 89 種類、ICD-10 では 52 種類であった。

次に、SOAP 記載に対して医療辞書（約 30 万語）を搭載した形態素解析器（MeCab）によって形態素解析を行い、各 SOAP 記載をベクトル化した。その上で、各病名に対して RBF カーネルを用いた非線形 SVM（Support Vector Machine）を用いて他クラス分類を行った。また、最適なパラメータ（コストパラメータおよび RBF パラメータ）を決定するために、hyperopt を 100 回試行することによって探索した。

これらの学習結果を受けて、病名付与実験を行った。病名付与には、最低限の学習データがあると考えられたものを予測することとした。具体的には、主病名は症例が 10 以上のもの 22 疾患と、症例数が少ない疾患を集めた「その他」について病名付与実験を行った。また、ICD-10 については、症例が 10 以上のもの 13 疾患と症例数が少ない疾患を集めた「その他」について病名付与実験を行った。1074 例のうち、874 例で学習を行い、残りの 200 例で評価を行った。評価は one-vesus-one 方式による多数決による病名付与を行った。実装環境は python3/scikit-learn を用いた。主病名の分布および ICD-10 の分類は以下のようになった。

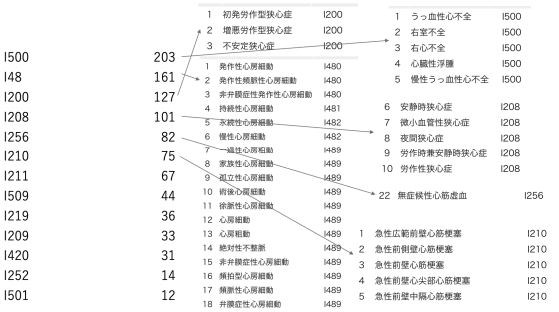
（倫理面への配慮）

人を対象とする医学系研究に関する倫理指針を遵守し研究を遂行する。

表１：主病名の出現頻度

うっ血性心不全	184	特発性拡張型心筋症	20
不安定狭心症	126	慢性心不全	17
労作性狭心症	100	急性前壁中隔心筋梗塞	17
無症候性心筋虚血	82	急性前側壁心筋梗塞	17
心房細動	63	心不全	16
急性下壁心筋梗塞	60	持続性心房細動	12
発作性心房細動	52	左心不全	12
急性前壁心筋梗塞	40	非弁膜症性心房細動	11
急性心筋梗塞	34	慢性心房細動	10
狭心症	32	特発性拡張型心筋症の疑い	10
		陳旧性心筋梗塞	10
		急性心不全	10

表２：ICD-10 の出現頻度



C．研究結果

具体的には、主病名付与（23 クラス）については、正答率 32.5%、ICD-10 の付与（14 クラス分類）については 44.5%であった。しかし、詳細に病名付与を検討すると、例えば「うっ血性心不全」と病名付与が行われたものの、実際に医師が付与した病名としては 14 例存在し、急性心不全、陳旧性心筋梗塞 2 例、慢性心不全、急性心不全、不安定狭心症 3 例、左心不全 2 例、急性前側心筋梗塞、突発性拡張型心筋梗塞などであった。また「心房細動」と間違ったものは 7

例あり、そのうち発作性心房細動6例、その他1例であった。

D. 考察

病名予測については、それほど高いとは言えない結果であったが、これは、国立循環器病研究センター、特に母集団となる MIDAS 研究の対象となる患者さんの疾患データであったため、病名が表１にみられるように疾患の偏りがあったことが理由として考えられる。特に、結果で視たような「心房細動」として予測したものについて、実際に主治医が付与した病名は「発作性心房細動」であった。このような場合は極めてカルテ記載において相違が現れにくいことが考えられる。また、分類としての病名が細分化されることで、意味的に排他的とは言えない状況が作られており、先の心房細動についても、発作性心房細動は心房細動の一つの発現の様相であって、意味的に予測を間違えたとは言い難い状況であった。これは、本研究の半会議においても議論となり、機械学習が判定した結果の方が正しい（そもそも間違っていない）のではないか、という議論となった。

なお、追加実験として機械学習手法として、SVM における one-versus-the-rest, またディープラーニングを行ってみたが、同等の結果であった。ただし、one-versus the rest による結果が今回の主実験である one-versus-one と異なる予測をしていることが多く（200 例の予測のうち、one-vs-one の正解数 65, one-vs-the-rest の正解数 65 であるものの両方が正解した数は 43)、より多くの学習データを準備できればこれらの性能も近づく可能性が高いことから、抽出すべきデータの分布を考慮することで、より高い精度の結果を得られると考えられる。

E．結論

電子カルテシステムにおける SOAP 記載データを利用して、機械学習を用いて病名予測を行った。結果、循環器、特に心疾患の予測においても、機械学習がある程度正確に病態を判定できることが示唆された。

G．研究発表

1. 論文発表　　なし
2. 学会発表　　なし

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