小児慢性特定疾病データベースのリンケージと解析に関する研究

研究分担者　　森　臨太郎　（国立成育医療研究センター政策科学研究部 部長）
研究協力者　　盛一　享徳　（国立成育医療研究センター小児慢性特定疾病情報室 室長）

研究要旨

医療情報の二次利用の重要性が昨今注目されている。診療情報明細書（レセプト）データベースや小児慢性特定疾病登録データベースと言った医療情報データベースは、データベース同士のリンケージにより、その価値を高めることができると考えられるが、その際には疾患名によるリンケージができることが重要となる。レセプトデータベースは、同一の疾患概念であっても複数の傷病名により登録が可能となっていることから、レセプトデータベースを二次利用するに当たっては、レセプト傷病名の類型化は避けて通れない。しかしながら、レセプトデータベースは極めて巨大であることから、従来の文字列検索のアプローチでは、膨大な時間と労力が必要となり、取扱が極めて困難である。

本研究は、日本語に対応した自然言語解析の技術の一つである InterSystems IRIS Natural Language Processing (NLP) Japanese を利用し、レセプト傷病名の自動類型化が可能であるかを検証した。IRIS NLP Japanese は主たる意味を持つ語句を認識し、さらに語句の欠損や入替があっても標準病名との対応を自動的に行えることが分かった。IRIS NLP Japanese を利用した傷病名の類型化は、現実的な解法との一つとなり得る可能性があることが分かった。

A．研究目的

近年種々の医療関係の情報が電子化され、その二次利用の重要性が認識されてきている。小児期の慢性疾患患者が多く登録されている小児慢性特定疾病登録データベースと診療情報明細書（レセプト）データの突合により、新たな知見が見いだせる可能性が期待されるが、これらのデータベースをリンケージする際には、『疾患名』が重要となる。平成 27 年度以降は小児慢性特定疾病における対象疾病は、登録の基準となる疾患名が明確に定義され、登録データ上『疾患名』の揺れはかなり減少したと思われるが、レセプトデータベースは、同一の概念の疾患であっても複数の傷病名が登録可能であり、また例外的ではあるが、自由記載による傷病名の登録も許されている。従って、双方のデータベースをリンケージするにあたり、レセプトデータベースに記録されている傷病名をグループ化し、同一の疾患概念と思われるレコードをまとめることが必要となる。コンピュータにより傷病名を疾患概念ごとにグループ化するためには、予めどのような語句がどのような疾患概念に属しているかの定義をする必要があるが、膨大な数のレセプトデータを事前に解析し、グループ化に必要となる辞書やオントロジーを定義する作業が求められるが、これらは人力で行わざるを得ず、現実的に不可能である。

非構造化データの集合である一般の文章を機械で解析する自然言語解析の技術の一つで

ある InterSystems IRIS Natural Language Processing (NLP) Japanese は、従来の自然言語解析技術とは異なり事前に辞書を準備しなくとも解析が可能とされる。IRIS NLP Japanese は、文章（非構造化データ）を意味のあるデータ項目（構造化データ）に変換することのできる自然言語処理技術の一つであり、原文中から検出される「エンティティ」と呼ぶ語句単位を切り出し、そのエンティティ同士の関連性を「パス」と呼ぶデータ項目として自動的に算出する。文章を単語要素単位に細断し解析する従来の自然言語解析手法とは事なり、IRIS NLP Japanese は、一単語以上からなる「エンティティ」を検出する。エンティティは単語ではなく文法を基に、文中から人間からみても意味の残っている「エンティティ」の切れ目を見出す。さらにこの解析には、面倒で限界のある辞書やオントロジーを予め定義する必要がない。

一般的な文字列解析は、文字列の完全一致であれば容易に検索が可能であるが、文字列に記載揺れがある場合、検索対象とする文字列を決め部分一致をさせる必要がある。この場合検索対象とする文字列を予め辞書として用意する必要があるが、膨大なレセプトデータに任意の文字列として記録されている傷病名文字列に対応する辞書を作成することは、極めて困難である。もし IRIS NLP Japanese の技術が傷病名という短い語句の集合体においても有効に作用するのであれば、傷病名の類型化という、極めて重要であるにも関わらず現在のところは人力による確認しか解決策がない課題に対する画期的な解法となり得る可能性がある。そこで本研究は、予め辞書やオントロジーを用意する必要の無い IRIS NLP Japanese の技術を利用して、膨大な傷病名リストから、類似する傷病名をグループ化することが可能であるかの検討を行った。

B．研究方法

本研究では、神奈川県および県下 33 市町村ならびに神奈川県国民健康保険団体連合会の全面協力のもと、県内の国民健康保険におけるレセプトデータを用いて、レセプトデータベースにおける傷病名の登録状況を把握し、同一の疾患概念のレセプト傷病名を類型化することが技術的に可能であるかの検討を行った。平成 25 年 1 月審査分から平成 26 年 12 月審査分ならびに平成 28 年 1 月審査分から平成 29 年 9 月審査分までの 20 歳未満の被保険者に関する神奈川県国民健康保険レセプトデータを用いた。

1. レセプトデータにおける傷病名（レセプト傷病名）は、コード化されている。このため一般財団法人医療情報システム開発センター（MEDIS）による標準病名マスターと突合し、日本語病名への変換と ICD-10 コードへの紐付けを行った
2. レセプト傷病名のうち、コード番号 999 は傷病名の自由記載コードであり、自由記載用フィールドにレセプト請求を行った医療機関で入力された自由記載による傷病名が含まれている。コード化された他の傷病名と合わせ、コード 999 により自由記載された傷病名も本研究による類型化の対象とした
3. 全く文字列に類似性のない傷病名同士の関係性は、ICD-10 コードを利用してグループ化を行った

（倫理面への配慮）

研究の遂行にあたっては、「人を対象とする

医学的研究に関する倫理指針」（平成 26 年文部科学省・厚生労働省告示）を遵守するとともに、国立成育医療研究センター倫理審査（受付番号：1729）の承認を受けて行われた。

C．研究結果

IRSI NLP Japanese により標準病名と自由記載された傷病名との比較・類型化を行った。その結果、以下の様な結果が得られた。

全く事前の辞書準備が無い状態で、日本語としての語句の切れ目を正しく理解するとともに、欠けている語句を補完しての認識（表 1）、修飾語を伴う傷病名に対し、主たる語句を抽出して認識（表 2）、長音の有無を含む語句の入れ替えの認識（表 3）を行うことができた。

表 1 欠けている語句を補完して認識

自由記載傷病名	標準病名	ICD10コード
アミノ酸代謝異常	アミノ酸代謝異常症	E729
ダウン症	ダウン症候群	Q909
右心低形成	右心低形成症候群	Q226
フォンタン術後	フォンタン術後症候群	I971
Ｂ６欠乏症	ビタミンＢ６欠乏症	E531

表 2 主たる意味を持つ語句を認識

自由記載傷病名	標準病名	ICD10コード
おたふくかぜの疑い→反復性耳下腺炎	反復性耳下腺炎	K112
びらんを伴う難治性口内炎	難治性口内炎	K121
両（瘢痕期）未熟児網膜症	未熟児網膜症	H351

表 3 語句の順番の入れ替えを認識

自由記載傷病名	標準病名	ICD10コード
エーラース・ダンロス症候群（血管型）	血管型エーラス・ダンロス症候群	Q796
ファロー四徴症（極型）	極型ファロー四徴	Q213
急性腎炎（溶連菌感染後）の疑い	溶連菌感染後急性糸球体腎炎	N009

一方で、認識に誤りがあった場合も認められた。修飾語を主たる意味をもつものと誤認し判断しているものが多かった。多くは外傷や整形外科、耳鼻科、眼科の領域で使用される体の部位を主たる意味をもつと誤認するケースであった（表 4）。

表 4 主たる意味を持つ語句の誤認

自由記載傷病名	標準病名	ICD10コード
じんま疹（全身）	全身倦怠感	
肺動脈閉鎖症（純型閉鎖）術後	僧帽弁形成術後	

D．考察

膨大なレセプトデータに記録されている「傷

病名」を機械的にグループ化できるかどうかの検証を行った。自然言語解析の技術の一つである IRIS NLP Japanese を用いたところ、傷病名という短い語句集合体であっても、全くの辞書を準備していない状態で、予想以上に正しくエンティティの認識が成されることが分かった。自由記載される傷病名については、文字の欠損や挿入がどのように発生するかを事前に予測することは極めて難しく、今回の結果のように、自動的に語句や文字の欠損や入れ替えを乗り越えて類型化が可能であったことは、今後のレセプト解析において極めて有益な結果であると思われた。

今回誤った類型化が成された結果では、型だの部位や処置に関する語句を重要と判断して類型化を行っている事例が最も多かった。本研究では IRIS NLP Japanese の能力を判定するために、意図的に事前の辞書を全く準備しなかったが、IRIS NLP Japanese は用意された辞書を事前知識として利用する事も可能であることから、体の部位や処置が修飾語であるという情報を与えておけば、より精度の高い結果が得られる可能性があると思われた。

E．結論

自然言語解析の技術の一つである IRIS NLP Japanese 利用により、これまで取扱が困難であったレセプト傷病名について、実現可能な作業量で、疾患概念ごとの類型化が行える可能性が示された。

【参考文献】

1) InterSystems IRIS NLP Japanese の概要 Version 2.0. インターシステムズ・ホワイトペーパー.
2) Bronselaer A, et al. Concept-Relational Text Clustering. International Journal of Intelligent Systems. 2012;27:970-93.

F．健康危険情報

なし

G．研究発表

1．論文発表

なし

2．学会発表

1) 盛一享徳. Natural Language Processing (NLP) を利用した病名収集の試み. 第 44 回日本診療情報管理学会学術大会（2018 年 9 月 20 日～21 日、新潟）

H．知的財産権の出願・登録状況

なし

1．特許取得

なし

2．実用新案登録

なし

3．その他

なし