

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

カルテ文章からの自動抽出した病名のクリーニングに関する研究

研究分担者：河添悦昌 東京大学大学院医学系研究科 特任准教授

A. 研究目的

東大病院の電子カルテに記載された診療記録から症状・所見・疾患に関する単語を抽出し、そこから解析エラー(病名以外の文字列)を削除し、精査した病名リストを作成する。

B. 研究方法

昨年度、2010年1月1日から2016年12月31日の期間を対象として、東京大学医学部附属病院の電子カルテに記載された診療記録(合計約1870万件の診療記録)から症状・所見・疾患表現(以下単に病名と呼ぶ)を自動抽出した。

これは、奈良先端大学の荒牧研究室で開発した病名抽出ツール(mednlp parser v006)で処理を施すことによって行った。

この結果、18,691,219種類の病名が抽出された。ただし、これらすべてが本当の病名でなく、解析エラーも含まれるため抽出された表現には病名として不適切なものも存在する。

例えば、

「むせこむが」

「一歩進む」

「色えんぴつください」

「最近つりやすく」

「■さんはC型肝炎」 ■部には人名

これらについて体系立てて整理しながら、精査を行った。

(倫理面への配慮)

この実施に際しては、東京大学大学院 医学系研究科の倫理承認(承認番号:11446)得て行った。

C. 研究結果

C-1. 解析エラーの分類

解析エラーはおおまかには次のように分類できた。

(0) キーワード単独では、病的概念を示すこと

がはっきりしない場合

閉所 → 「閉所」だけでははっきりしないので不採用

(ただし、閉所恐怖 → 主訴・症状とわかるので採用)

(1) 品詞「名詞」以外で終わる文の扱い

自動抽出した病名を形態素解析にかけた結果、文末形態素の品詞は約9割(398177件)が名詞であった。

名詞以外のものを以下のように精査した。

(1-1) 動詞・形容詞・助動詞

動詞・形容詞・助動詞で終わる文は、基本形(終止形)で終わっていて、かつ、内容的に採用基準を満たすものであれば病名とみなす。

逆に、終止形以外は病名とみなさない。

例) 終止形の例

筋緊張強くなる

右足首がむくむ

熱感目立たない

会話問題ない

イライラ目立たず

例) 終止形以外の例

最近つりやすく

嘔吐され

ふらつきそうで

ふらつきなく

嘔吐無く

(1-2) 助詞

格助詞、並立助詞、接続助詞、および、疑問や禁止の終助詞で終わる場合は病名とみなさない。

例) 格助詞、並立助詞、接続助詞

しんこきゅうを

むせたり
ぼんやりとかすんで

確認や詠嘆の終助詞で終わる文は、所見や主訴（症状）を示すようなものであれば病名とみなす。
例)

動くとけっこうしんどいね
ふらふらしちゃうな
手がむくんでるね
痛くないんだけどさ

(1-3) 副詞

(1) 程度等を表す副詞で終わる表現は内容的に採用基準を満たせば病名とみなす
例)

腰痛少々
息切れがかなり
先がびりびり
食欲いまいち
違和感再度
耳鳴時々

(2) 病名・症状名が複数個含まれるものも病名とみなす
例)

立位保持可 閉脚ふらつき＋ 閉眼ふらつき＋
誤嚥性肺炎既往あり ときどきむせこむ
原発巣及び縦隔リンパ節転移
痰がらみと息苦しさ
うっ血性心不全／アドリアマイシン心筋症
「チクチク」「シクシク」
未分化癌 or 低分化癌
典型像ではない。脳腫瘍
肝硬変→肝癌
左下肢痛＞腰痛

(3) 情報が不足していそうなもの（「何が？」がわからない）は、基本的に病名とみなさず不採用とし、病的概念に関連していそうだと判断できるものは採用する。

例) 病名とみなさない例
いつもこうなんです
ありません
例) 病名とみなす例
つまりやすさ
剥がれやすさ
異常有り

(4) 記号やスペース
前後に余分な（無視できる）記号やスペースを含むものは病名とみなさない。

例) 先頭にスペース
 膝癌
例) 先頭に記号
「アスペルガー症候群
— S 状結腸穿孔
例) 末尾に記号
動悸??
下痢症状??????
食物アレルギー—

カッコ書きがある場合も、内容が採用基準を満たせば病名とみなす。

例) 右乳癌 (T2N1M0-Stage II B)
冠動脈有意狭窄 (+)
インフルエンザ (A 型)

カッコの中身が病名の補足、読みなど場合も、それを含めて病名とみなす。

髄膜 (脳) 炎
大藤病 (おおふじびょう)
例) カッコの中と外が「症状 (病名)」「病名 (患者の状態)」「部位 (本数)」などの関係
過排卵 (高エストロゲン血症)
乳頭部癌 (再発)
右肋骨 (2 本) 骨折

カッコの片方が欠落しているものは病名とみなさない。

例) 過排卵 (高エストロゲン血症
「むくみ
非潰瘍型)

(5) 一般的ではない表記、または、誤字と思われるものも病名とみなす。

例)
すいみん
うつびょう
たこつば型心筋症
(たこつぼの誤りと思われる)
このような基準で頻度 5 回以上の病名について全件を精査した。

D. 考察

データ抽出過程のため特になし.

E. 結論

データ抽出過程のため特になし.

F. 健康危険情報

特になし.

G. 研究発表

特になし.

H. 知的財産権の出願・登録情報

該当なし