[別添 4]

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

病名自動抽出のための辞書リソースに関する研究

研究分担者：若宮翔子 奈良先端科学技術大学院大学 研究推進機構

A. 研究目的

万病辞書を行方向および列方向へ拡張するとともに，辞書項目を精査する．

B. 研究方法

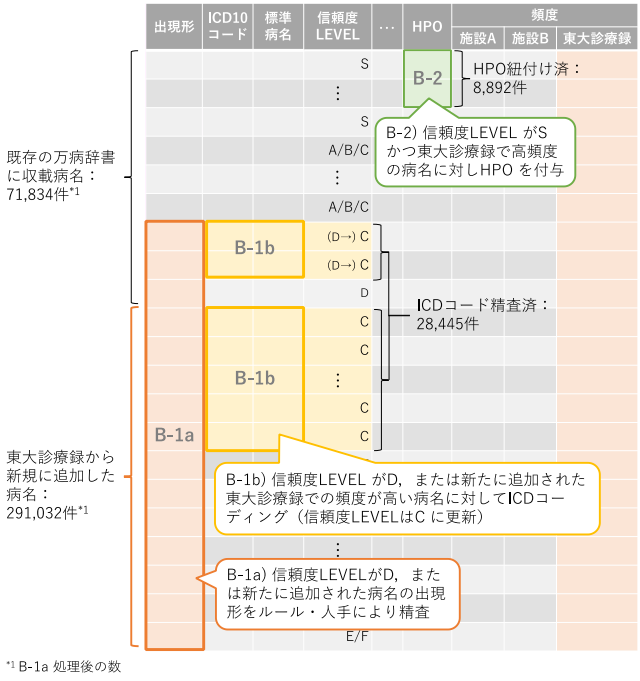

図 1．万病辞書の行方向および列方向への拡張の概要．灰色のセルは既存の万病辞書の項目，それ以外のセルは今回処理した項目を示す．

B-1） 万病辞書の行方向への拡張：新たな病名表現の追加

B1-a） 病名の追加と精査

平成 29 年度の分担研究「カルテ文章からの病名自動抽出に関する研究」にて，2010 年 1 月 1 日から 2016 年 12 月 31 日の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（以降，東大診療録と記載）における症状・所見・疾患（以降，単に病名と記載）に関する語（以降，出現形とも記載）を抽出し，東大診療録における出現頻度とともに追加した．なお，追加する語は以下のルールに基づきフィルタリングした．

追加病名フィルタリングルール：

・東大診療録のみに出現し，その頻度が 3 未満の病名のみは除去する．ただし，信頼度 LEVEL （病名に対する ICD コードや標準病名の確からしさ）が S の病名については頻度によらず全て収載する．

・「ふらつきなし」のように「万病辞書に収載済みの病名+なし（無し）」というパターン病名は除去する．

病名自動抽出では，明らかな抽出ミスや病名とはいえない語が多数抽出される．そこで，万病辞書に収載済みだが人手による精査が行われていない病名（信頼度 LEVEL が D）と新たに追加された病名の出現形を，ルールベースと人手によるチェックで精査した（図 1 の B1-a）．

B1-b） ICD コード・標準病名の付与と精査

新たに追加された語には，ICD コードと標準病名が付与されていない．そのため，万病辞書に収載済み病名の ICD コードと標準病名を元に，追加病名への自動コーディング処理を行った．この処理により ICD コードと標準病名が付与された病名の信頼度 LEVEL を E，付与できなかったものの信頼度 LEVEL を F として区別した．

次に，使用頻度が高い語について，人手によるコーディング（信頼度 LEVEL が F の病名），ならびに自動付与された ICD コードと標準病名の精査

（信頼度 LEVEL が D または E の病名）を行なった（図 1 の B1-b）．1 名の医療従事者によるコーディングまたは精査が行われた語の信頼度 LEVEL を C に変更した．なお，既存の万病辞書に収載済みの病名に対するコーディングガイドラインに準拠し，対応する ICD10 コードまたは標準病名がない場合には-1 を付与した．

B-2） 万病辞書の列方向への拡張：万病辞書と Human Phenotype Ontology (HPO) の病名の紐づけ

ICD10 コードや標準病名以外の病名分類として，ヒト疾患に関する表現型語彙について体系的に整備されており，国際的な症状データの標準として，他のオントロジーや知識ベースとの連携も進められている Human Phenotype Ontology (HPO) [1] を追加し，万病辞書を拡張した．このために，万病辞書の病名に対応する語を HPO で検索し，紐付けた．今回は，ICD10 対応標準病名マスター [2] に掲載されている 25,678 病名（万病辞書における信頼度 LEVEL が S）の一部と，HPO 日本語版 [3] の語を対応させた．図 2 は今回用いた HPO 日本語版の抜粋である．まず，万病辞書における出現形と HPO 日本語版の Japanese term (expert) の語との編集距離を求め，編集距離が閾値以下のものについては自動的に紐付けを行なった．そして，B-1） で追加した病名について，人手による紐付けと精査を行った．なお，この処理は一般のアノテーターが以下のルールに基づき行なった．

紐付け，および，表記ルール：
·対応する語が複数ある場合には，セミコロン（;）区切りで，すべて列挙する．
·部位が明らかに一致しない場合は対応なしとみなす．
·対応する HPO 病名がない場合は-1 を付与する．
·対応すると思うが自信がない場合は語尾にクエスチョンマーク（？）を付与する．
·ぴったり当てはまる語がない場合は，上位概念に当たる語やほぼ同義と考えられる語があれば割り当てる．
·HPO 日本語版の HPO: Japanese term (expert) 以外に記載されている語も参考にする．
·HPO 日本語版の「～エピソード」は「～の症状／発症」という意味で使われる医療用語と捉え，紐付け対象とする．
·万病辞書の病名より HPO の分類が細かい（部位の一部など）は紐付け対象外とする．例えば，「大腸癌」に対して「結腸癌」「胃腸癌」は対象外，「心不全」に対して「右室不全」も対象外とする．

また，割当に迷った項目については，医療従事経験者とやりとりしながら進めた．その結果作成されたルールの例を以下に示す．
·「～腫瘤」：HPO 日本語版の「～瘤」「～腫」を紐づけ対象とし，「～新生物」は紐づけしない．
·「～腫瘍」：HPO 日本語版の「～腫」「～新生物」を紐づけ対象とし，「～瘤」は対象外とする．
·「～ポリープ」：HPO 日本語版の「～腫」「～ポリープ」を紐づけ対象とし，「～新生物」「××瘤」は対象外とする．
·「～癌」：HPO 日本語版の「～癌」という表現で同義のものがない場合，「～新生物」「(悪性）～腫（瘍)」 などを紐づけ対象とする．

HPO ID	English term	Japanese term (expert)	Japanese term (Life Science Dictionary)	Japanese term (Mammalian Phenotype Japanese)	Japanese term (Google Translate)
HP:0002024	Malabsorption	吸収障害	吸収障害 OR 吸収不良 OR 吸収不全	NA	吸収不良
HP:0002033	Poor suck	吸啜不全	さしい OR 曾曾 OR 不得意 OR 不十分	NA	悪い吸う
HP:0006721	Acute lymphoblastic leukemia	急性リンパ性白血病	急性リンパ性白血病 OR 急性リンパ球性	白血病	急性リンパ芽球性白血病
HP:0008942	Acute rhabdomyolysis	急性横紋筋融解	急性 OR 急性型 OR 急性的 OR 良い OR	NA	急性横紋筋融解
HP:0006965	Acute necrotizing encephalopathy	急性壊死性脳症	急性壊死性脳症	NA	急性壊死性脳症
HP:0011949	Acute infectious pneumonia	急性感染性肺炎	急性 OR 急性型 OR 急性的 OR 良い OR	NA	急性感染性肺炎
HP:0200119	Acute hepatitis	急性肝炎	急性肝炎	NA	急性肝炎
HP:0006554	Acute hepatic failure	急性肝不全	急性肝不全	NA	急性肝不全
HP:0012388	Acute bronchitis	急性気管支炎	急性気管支炎	NA	急性気管支炎
HP:0006733	Acute megakaryocytic leukemia	急性巨核球性白血病	急性巨核球性白血病 OR 急性巨核球性	白血病	急性巨核球性白血病
HP:0100282	Acute colitis	急性結腸炎	急性大腸炎	大腸炎	急性大腸炎
HP:0011948	Acute respiratory tract infection	急性呼吸器感染症	急性気道感染 OR 急性呼吸器感染症	NA	急性呼吸器感染症
HP:0012467	Acute respiratory acidosis	急性呼吸性アシドーシス	急性 OR 急性型 OR 急性的 OR 良い OR	アシドーシス	急性呼吸性アシドーシス
HP:0011952	Acute aspiration pneumonia	急性誤嚥性肺炎	急性 OR 急性型 OR 急性的 OR 良い OR	吸引	急性吸引性肺炎
HP:0008281	Acute hyperammonemia	急性高アンモニア血症	急性 OR 急性型 OR 急性的 OR 良い OR	NA	急性高アンモニア血
HP:0004808	Acute myeloid leukemia	急性骨髄性白血病	急性骨髄性白血病	白血病	急性骨髄性白血病
HP:0004820	Acute myelomonocytic leukemia	急性骨髄単球性白血病	急性骨髄単球性白血病	白血病	急性骨髄単球性白血病
HP:0006573	Acute hepatic steatosis	急性脂肪肝	急性 OR 急性型 OR 急性的 OR 良い OR	脂肪肝	急性肝性脂肪症
HP:0011128	Acute esophageal necrosis	急性食道壊死	急性 OR 急性型 OR 急性的 OR 良い OR	壊死	急性食道壊死
HP:0001919	Acute kidney injury	急性腎外傷	急性腎障害 OR 急性腎傷害	NA	急性腎障害
HP:0004836	Acute promyelocytic leukemia	急性前骨髄球性白血病	急性前骨髄球性白血病	急性前骨髄球性白血病	急性前骨髄球性白血病
HP:0007131	Acute demyelinating polyneuropathy	急性脱髄性ポリニューロパチー	急性 OR 急性型 OR 急性的 OR 良い OR	NA	急性脱髄性多発神経障害
HP:0004845	Acute monocytic leukemia	急性単球性白血病	急性単球性白血病	白血病	急性単球性白血病
HP:0000371	Acute otitis media	急性中耳炎	急性中耳炎	NA	急性中耳炎
HP:0007280	Acute infantile spinal muscular atrophy	急性乳児脊髄性筋萎縮	急性 OR 急性型 OR 急性的 OR 良い OR	筋萎縮症	急性小児脊髄性筋萎縮症
HP:0008682	Acute tubular necrosis	急性尿細管壊死	急性尿細管壊死 OR 急性腎尿細管壊死	壊死	急性尿細管壊死

図 2. HPO 日本語版の抜粋

（倫理面への配慮）
本研究については課題名「電子的診療録の自動構造化機能を有した入力ソフトウェアの開発研究」で，奈良先端科学技術大学院大学情報学系の倫理審査に申請し，申請が受理されている．

C. 研究結果

B-1） 万病辞書の行方向への拡張：新たな病名表現の追加

B1-a） 病名の追加と精査

東大診療録から自動抽出された病名を単純に追加した結果，総病名数は160万件以上となった．これに追加病名フィルタリングを適用した結果，病名の総数は約47万件となった．このうち，既存の万病辞書に収載済みだが人手による精査が行われていない病名（信頼度LEVEL がDの36,611件）と新たに追加された病名の出現形（信頼度LEVEL が

Eの242,437件とFの154,363件）の計433,411件を，ルールベースと人手によるチェックで精査した．この結果，約105,000件（信頼度LEVELがDが28件，Eが61,920件，Fが43,848件）が削除され，324,748件が病名の出現形として残された．すなわち，東大診療録から新たに291,032件の病名の出現形が追加された．

B1-b）ICDコード・標準病名の付与と精査

人手によるICD10コードの付与や精査が行われていない病名（信頼度LEVELがD，E，F）を東大診療録頻度が高いものから順に，医療従事者1名がICD10コードと標準病名の付与ならびに精査を行なった．この結果，信頼度LEVELがDの5,565件，Eの11,026件，Fの11,854件に対して人手によるICD10コードの付与ならびに精査を行った．この結果，計28,445病名の信頼度LEVELがCとなった．図3に信頼度LEVELごとの件数を示す．

今回のコーディングおよび精査では，10,701病名のICD10コードに-1が付与された．ICD10コードに-1が付与された東大診療録頻度上位10病名は，「問題なし」，「膨張」，「合併症」，「転倒」，「発作」，「潰瘍」，「有害事象」，「危険行動」，「苦痛」，「炎症」であり，病気の原因あるいは結果に関する語や，複数の部位に関わる語などが見られた．

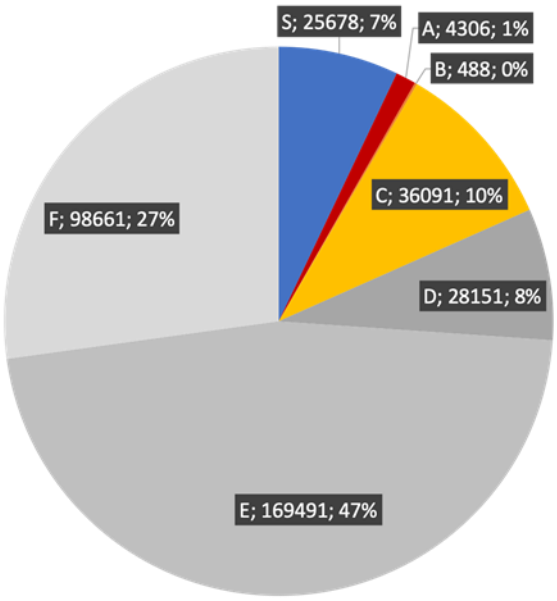

図3．信頼度LEVELごとの件数．データラベルは「信頼度LEVEL；件数；パーセンテージ」を表す．

B-2） 万病辞書の列方向への拡張：万病辞書とHuman Phenotype Ontology（HPO）の病名の紐づけ

万病辞書における出現形とHPO日本語版のJapanese term（expert）の語との編集距離を求めたところ，編集距離が0，すなわち，完全一致した病名は967であった．これら以外の信頼度LEVELがSの病名のうち，B-1）で追加した東大診療録頻度が5以上の7,925病名に対し，人手でHPO病名を付与した．このとき，アノテーターの作業効率化のために，編集距離が1のHPO病名を候補として示した．この結果，3,995病名にHPO病名が紐付けされた．このうち，アノテーターが対応すると思うが自信がない（確信度が低い）とした紐付けは299件であった．

D. 考察

B-1） 万病辞書の行方向への拡張：新たな病名表現の追加

今回の処理により，東大診療録から 291,032 件の病名の出現形が新たに追加され，既存の万病辞書（74,729 件）から約 4.8 倍（362,866 件）に拡大することができた．今回追加した病名は東大病院診療録から抽出されたものであり，万病辞書が対象とする「臨床現場で実際に使われる病名」をほぼ網羅できたと期待される．ICD コード・標準病名の付与と精査はコストがかかる作業であり，継続して実施する必要がある．

B-2） 万病辞書の列方向への拡張：万病辞書とHuman Phenotype Ontology（HPO）の病名の紐づけ

信頼度 LEVEL が S の一部の病名に対する紐付けを行なったが，継続して実施する必要がある．また，確信度が低いとされた紐付け結果については，今後医療従事者による精査を行う計画である．

E. 結論

万病辞書の行方向および列方向へ拡張するとともに，辞書項目の精査を行なった．万病辞書の行方向への拡張では，新たな病名表現を追加するとともに，ICDコード・標準病名の付与と精査を行なった．結果として，291,032行が追加をされ，既存の万病辞書の収載病名（の出現形の）数を約4.8倍に拡大した．

万病辞書の列方向への拡張では，ICD10コードや標準病名以外の重要な病名分類オントロジー

としてHuman Phenotype Ontology (HPO) の病名列を追加し，標準病名マスタに掲載されている病名の一部との紐付けを行なった.

今回の行方向への拡張により，臨床現場で実際に使われる病名はほぼ網羅されたと期待されるため，各病名の出現形に付随する情報（ICD10, ICD11, MedDRA, HPOなど）の精査や追加を重点的に進める必要がある.

[参照文献]

[1] Sebastian Köhler, Nicole Vasilevsky, Mark Engelstad, Erin Foster, et al. The Human Phenotype Ontology in 2017, Nucl. Acids Res. (2017) doi: 10.1093/nar/gkw1039

[2] 標準病名マスター http://www.dis.h.u-tokyo.ac.jp/byomei/

[3] Human Phenotype Ontology (HPO) 日本語版 https://github.com/ogishima/HPO-japanese

F.健康危険情報

該当なし

G.研究発表

1．論文発表

該当なし

2．学会発表

Kaoru Ito, Hiroyuki Nagai, Taro Okahisa, Shoko Wakamiya, Tomohide Iwao, Eiji Aramaki: J-MeDic: A Japanese Disease Name Dictionary based on Real Clinical Usage, LREC 2018. (Miyazaki, Japan)

H. 知的財産権の出願・登録情報

該当なし