[別添 1]

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）

カルテ情報の自動構造化システムと疾患数理モデルの逐次的構築,
及び,
自動構造化機能を有した入力機構の開発に関する研究

平成 30 年度 総括・分担研究報告書

研究代表者 荒牧　英治

平成３１（2019）年３月

[別添 2]

目次

[別添 3]

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 総括研究報告書）
カルテ情報の自動構造化システムと疾患数理モデルの逐次的構築，及び，自動構造化機能を有した入力機構の開発

研究代表者　荒牧英治 奈良先端科学技術大学院大学 研究推進機構

研究要旨

電子カルテは患者情報が全て記録されているものの，非文法的かつ断片化した表現が多く自然言語処理を応用した利活用は困難であった．これを二次利用するため申請者等（申請者荒牧及び分担者河添が所属する研究室主宰者の大江ら）は，2008年から電子カルテから医療用語の自動抽出及び自動コーディングを行う研究に従事してきた．その成果は，日本内科学会の症例報告検索システムなどとして実用化され，現在も用いられている．本研究は，電子カルテの二次利用のさらなる実用化に向けて問題となる次の２つの課題を解決する．

（課題１）実用化可能な解析精度の達成

（課題２）電子カルテに組み込み可能な実装の開発

若宮翔子（奈良先端科学技術大学院大学 研究推進機構･特任助教）

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

これまで，医療医学用語をまとめる試みは数々なされてきた．これらは主に医師・研究者などの医療者が使う用語を対象としており，電子カルテなど医療現場で扱う技術は向上しつつある．万病辞書もその辞書の一つであり，かつてない多くの用語を収載している．その一方で，近年増えつつある患者が記述したテキストを扱うためには十分な用語が収載されていない．

本研究はこれまで予定より早く医療用語のデータの収集が進んでおり，今後は，コストの許す限り人手による精査を進めていくのみが課題であると考えている．そこで最終年度である本研究は，当初の計画にはなかったこの患者表現をも一部収載しようと発展的な研究を行った．

患者によるテキストは，「闘病記」，「病の語り」，「当事者ブログ」，「患者 SNS」，「patient narrative」など，さまざまな呼び方が存在するが，本書では，「患者テキスト」と呼ぶことにする． この患者テキストに記述される内容は，不安などの心の問題（36%），症状・副作用・後遺症（16%）家族，周囲の人との関係(10%)，告知やインフォームドコンセント・セカンドオピニオン(9%)，診断・治療(8%)であり[医療言語処理（荒牧英治著）より]，不安と並んで，医学的な情報も多い．

患者テキストを利用した自然言語処理研究は始まったばかりであり，その課題は明確に定まっていない．しかし，最初に解くべき問題として，医学知識が豊かであるとは限らない患者が記述する患者表現を正式な医学表現に変換することが挙げられる．これは，従来，自然言語処理で表記ゆれ解消と言われた技術と近いが，単なる語の言い換えレベルにとどまらず，語から句への変換が必要となる場合もある．

例えば，医療用語としては〈感覚鈍麻〉として記述されるべき症状が，患者テキストにおいては「指先がピリピリする」と記述されうる場合がある．このように，擬音語や擬態語（以降，

擬音語と擬態語を総じてオノマトペと呼称する）を用いた動詞句レベルや文レベルでの表現が頻用される．患者テキストを医療現場で活用するためには，これらの非医療用語を言い換える必要がある．

このような言い換え全般をカバーすることは現在の自然言語処理では困難であり，大規模な辞書が必要であると考えている．このような背景の中，本研究は，大規模な患者表現の辞書を構築する．

B.　研究方法

B-1. 症状オノマトペの収集

前述したように患者表現が医学表現と異なる点は次の２つである．

・語彙的ギャップ：擬音語擬態語の頻用，特にオノマトペに代表される表現．

・構造的ギャップ：複合名詞である医学用語が動詞句として表現される．

これら２つの違いを考慮して，オノマトペに特化した収集方法と，クラウドソーシングを用いて動詞句を含んだ表現を募集する方法の２つを併用して表現を収集し，標準病名との対応をとる．以降，前者を症状オノマトペ，後者を症状句と呼称する．

B-2.　部位リストから部位＋オノマトペ表現の自動収集

Google n-gramコーパスから「（部位）が（オノマトペ）する」という表現を収集する．

擬音語，擬態語はカタカナであるかどうかで判定する．人体部位は事前に作成した表１を用いた．

この結果，部位，オノマトペ，頻度の３つ組のリストが得られた．次に，これを精査し，不適切なものを除いたデータを構築した．収集したデータのイメージを把握しやくするため，人手で可視化したものを図１に示す．

B-3. オノマトペ表現の標準化

次にオノマトペ表現がどのような医学表現に対応するかを人手により紐づけた．

医学表現としては，標準病名とし，部位＋オノマトペのペアで標準病名との対応を得た．例えば，「頭・ガンガン」は標準病名＜頭痛＞，「耳・ガンガン」は＜耳鳴＞となる．曖昧性があり一意に決定できない場合は，複数の標準病名を列挙した．

表１：人体部位リスト（抜粋）

人体部位	オノマトペ
頭部	がんがん
大腿部	パンパン
上腕	プルプル
膝	バキバキ

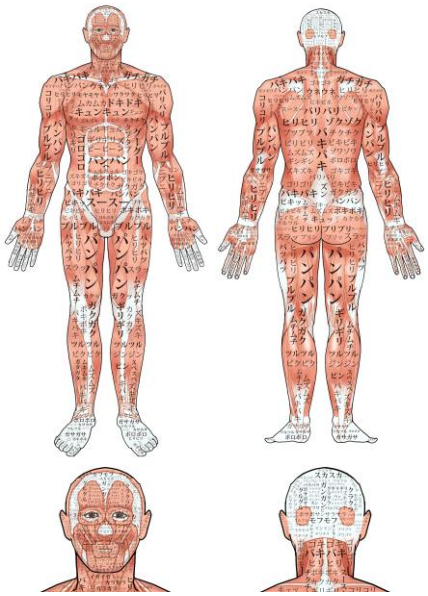

図 1: 部位ごとに可視化したオノマトペ

B-4. 症状句の収集とその方法

オノマトペと異なり，句の表現の形式は「AのB」「AがBする」「AとBがCする」など様々であり，事前に形式を決める収集はできない．患者表現を医学表現に紐付けるのではなく，逆に，医学表現を患者表現（形式は問わない）に言い換えることで収集を行った．医学表現としては＜標準病名＞を用い，クラウドソーシングにて収集した．収集方法としては，クラウドソーシングでは万病辞書の表現形から，内科学会頻度が25以上のものを対象に100名にアンケートを実施した．その際の教示は以下の通りである．

「症状を表す各単語について，他にどのような言い方（表現）ができるかを書いていただくタスクです．特定の地域や集団のみで使用されている表現（方言や隠語など）でもかまいません．またいくつ書いていただいてもかまいません．ただし，本調査は医療用語に関するものですので，それ以外の分野で同様の表記があるものに関しても，医療用語として認識してください．」

同一表現のものは一部省略（例：’すりガラス陰影’と’スリガラス陰影’）し，411単語のアンケートを９部にわけ，性別，年齢，出身地と，図２のような設問ごとに自由記述をしてもらった．収集したデータは句読点を削除し，同一表現ごとに集計したものを降順に並べた．漢字，ひらがな，カタカナは全て別表現として捉えている．参加者の性別，年代の内訳は図３に示すのとおりである．

図2 自由記述フォーマット

性別	人数
男性	568
女性	332
年代	人数
10 代	6
20 代	82
30 代	195
40 代	353
50 代	193
60 代	58
70 代	12
80 代以上	1

図3 参加者の性別と年代

（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

症状オノマトペ・症状句の収集の結果を，ウェブ上（http://sociocom.jp/~data/2019-pde/）に掲載した．また，本収集方法を論文として投稿予定である．

D. 考察

評価の困難さ：

本研究では，オノマトペに注目した方法とクラウドソーシングを用いた方法という２つの方法により，患者症状表現の収集を行った．患者表現収集の大規模な試みの事例は研究代表者らの知る限りなく，また，既存のリソースもないため，結果の評価は困難である．

例えば，クラウドソーシングで症状の表現を収取して本リソースの網羅性を評価しようとしても，これはクラウドソーシングによる手法と同じことであり，同じデータを再現してしまうだけである．今後は，構築方法とともに評価方法を検討する必要がある．

応用可能性：

本研究は患者症状表現の収集を行った．患者表現収集の大規模な試みの事例は乏しく，どれくらいをカバーすれば全体の何％をカバーするのかも分からず，また，何が実現できるのかも分からない．評価方法とともに小規模な応用を繰り返しながら，検討することも有効であると考えている．潜在的な応用先は次のようなサービスを考えている：

（１）スマートフォンやスマートスピーカーを用いた患者症状の抽出．

患者が日常的に用いるデバイスに日々蓄積される自然文から，患者の症状の抽出を行い，想定外の有害事象やアンメットニーズの発見につなげる．

（２）待ち時間の問診票

病院の待ち時間などに患者に問診票を記載する際に，その自然文を解析し，カルテに転送するなど，病院業務の軽減につなげる．

（３）医療者-患者間コミュニケーション支援

患者表現と医療用語を結びつけた辞書により，患者と医療者の双方が，相手側の用語を知り，コ

ミュニケーションを円滑にする教育効果が期待できる.

E．結論

本研究では，これまで大規模な収集が困難であった患者の症状表現の収集を行った．これは現在，ウェブ（http://sociocom.jp/~data/2019-pde/）にて公開している.

F. 健康危険情報

特になし.

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

特になし

[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

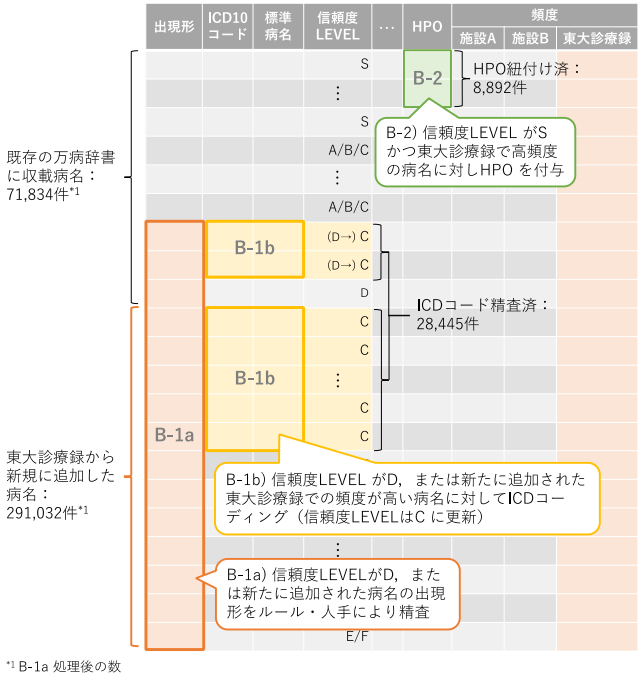

図 1．万病辞書の行方向および列方向への拡張の概要．灰色のセルは既存の万病辞書の項目，それ以外のセルは今回処理した項目を示す．

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．

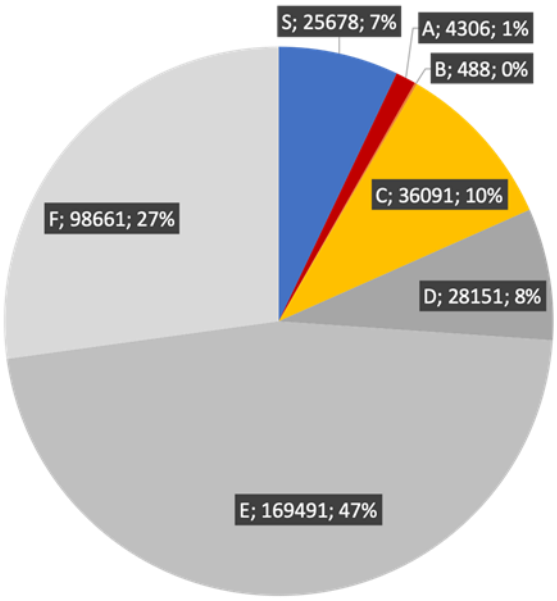

図3．信頼度LEVELごとの件数．データラベルは「信頼度LEVEL；件数；パーセンテージ」を表す．

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

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型肝硬変」■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
　膵癌
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 ⅡＢ）
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

［別添5］

研究成果の刊行に関する一覧表

論文

発表者氏名	論文タイトル名	発表誌名	巻号	ページ	出版年
該当なし					

令和元年５月２８日

厚生労働大臣
（~~国立医薬品食品衛生研究所長~~）　殿
（~~国立保健医療科学院長~~）

機関名　奈良先端科学技術大学院大学
所属研究機関長　職　名　学長
氏　名　横矢　直和　印

次の職員の平成３０年度厚生労働科学研究費の調査研究における、倫理審査状況及び利益相反等の管理については以下のとおりです。

１．研究事業名　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）

２．研究課題名　カルテ情報の自動構造化システムと疾患数理モデルの逐次的構築，及び，自動構造化機能を有した入力機構の開発

３．研究者名　（所属部局・職名）　研究推進機構　特任准教授

（氏名・フリガナ）　荒牧　英治　（アラマキ　エイジ）

４．倫理審査の状況

	該当性の有無		左記で該当がある場合のみ記入（※1）		
	有	無	審査済み	審査した機関	未審査（※2）
ヒトゲノム・遺伝子解析研究に関する倫理指針	□	■	□		□
遺伝子治療等臨床研究に関する指針	□	■	□		□
人を対象とする医学系研究に関する倫理指針（※3）	■	□	■	奈良先端科学技術大学院大学	□
厚生労働省の所管する実施機関における動物実験等の実施に関する基本指針	□	■	□		□
その他、該当する倫理指針があれば記入すること（指針の名称：　）	□	■	□		□

（※1）当該研究者が当該研究を実施するに当たり遵守すべき倫理指針に関する倫理委員会の審査が済んでいる場合は、「審査済み」にチェックし一部若しくは全部の審査が完了していない場合は、「未審査」にチェックすること。

その他（特記事項）

（※2）未審査に場合は、その理由を記載すること。
（※3）廃止前の「疫学研究に関する倫理指針」や「臨床研究に関する倫理指針」に準拠する場合は、当該項目に記入すること。

５．厚生労働分野の研究活動における不正行為への対応について

研究倫理教育の受講状況	受講 ■　未受講 □

６．利益相反の管理

当研究機関におけるＣＯＩの管理に関する規定の策定	有 ■　無 □（無の場合はその理由：　）
当研究機関におけるＣＯＩ委員会設置の有無	有 ■　無 □（無の場合は委託先機関：　）
当研究に係るＣＯＩについての報告・審査の有無	有 ■　無 □（無の場合はその理由：　）
当研究に係るＣＯＩについての指導・管理の有無	有 □　無 ■（有の場合はその内容：　）

（留意事項）　・該当する□にチェックを入れること。
・分担研究者の所属する機関の長も作成すること。

令和元年５月２８日

厚生労働大臣
（~~国立医薬品食品衛生研究所長~~）　殿
（~~国立保健医療科学院長~~）

機関名　奈良先端科学技術大学院大学
所属研究機関長　職　名　学長
氏　名　横矢　直和　印

次の職員の平成３０年度厚生労働科学研究費の調査研究における、倫理審査状況及び利益相反等の管理については以下のとおりです。

１．研究事業名　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）

２．研究課題名　カルテ情報の自動構造化システムと疾患数理モデルの逐次的構築，及び，自動構造化機能を有した入力機構の開発

３．研究者名　（所属部局・職名）　研究推進機構　特任助教

（氏名・フリガナ）　若宮　翔子　（ワカミヤ　ショウコ）

４．倫理審査の状況

	該当性の有無		左記で該当がある場合のみ記入（※1）		
	有	無	審査済み	審査した機関	未審査（※2）
ヒトゲノム・遺伝子解析研究に関する倫理指針	□	■	□		□
遺伝子治療等臨床研究に関する指針	□	■	□		□
人を対象とする医学系研究に関する倫理指針（※3）	■	□	■	奈良先端科学技術大学院大学	□
厚生労働省の所管する実施機関における動物実験等の実施に関する基本指針	□	■	□		□
その他、該当する倫理指針があれば記入すること（指針の名称：　　）	□	■	□		□

（※1）当該研究者が当該研究を実施するに当たり遵守すべき倫理指針に関する倫理委員会の審査が済んでいる場合は、「審査済み」にチェックし一部若しくは全部の審査が完了していない場合は、「未審査」にチェックすること。

その他（特記事項）

（※2）未審査に場合は、その理由を記載すること。
（※3）廃止前の「疫学研究に関する倫理指針」や「臨床研究に関する倫理指針」に準拠する場合は、当該項目に記入すること。

５．厚生労働分野の研究活動における不正行為への対応について

研究倫理教育の受講状況	受講 ■　未受講 □

６．利益相反の管理

当研究機関におけるＣＯＩの管理に関する規定の策定	有 ■　無 □（無の場合はその理由：　　）
当研究機関におけるＣＯＩ委員会設置の有無	有 ■　無 □（無の場合は委託先機関：　　）
当研究に係るＣＯＩについての報告・審査の有無	有 ■　無 □（無の場合はその理由：　　）
当研究に係るＣＯＩについての指導・管理の有無	有 □　無 ■（有の場合はその内容：　　）

（留意事項）　・該当する□にチェックを入れること。
・分担研究者の所属する機関の長も作成すること。

平成　31　年 2 月　8 日

厚生労働大臣　殿

機関名　東京大学

所属研究機関長　職　名　総長

氏　名　五神　真　　印

次の職員の平成 30 年度厚生労働科学研究費の調査研究における、倫理審査状況及び利益相反等の管理については以下のとおりです。

1．研究事業名　政策科学総合研究事業(臨床研究等 ICT 基盤構築研究事業)

2．研究課題名　カルテ情報の自動構造化システムと疾患数理モデルの逐次的構築、及び、自動構造化機能を有した入力機構の開発

3．研究者名　（所属部局・職名）　医学部附属病院・講師

（氏名・フリガナ）　河添　悦昌・カワゾエ　ヨシマサ

4．倫理審査の状況

	該当性の有無	左記で該当がある場合のみ記入 (※1)		
	有　無	審査済み	審査した機関	未審査 (※2)
ヒトゲノム・遺伝子解析研究に関する倫理指針	□ ■	□		□
遺伝子治療等臨床研究に関する指針	□ ■	□		□
人を対象とする医学系研究に関する倫理指針 (※3)	■ □	■	東京大学大学院医学系研究科	□
厚生労働省の所管する実施機関における動物実験等の実施に関する基本指針	□ ■	□		□
その他、該当する倫理指針があれば記入すること（指針の名称：　）	□ ■	□		□

(※1) 当該研究者が当該研究を実施するに当たり遵守すべき倫理指針に関する倫理委員会の審査が済んでいる場合は、「審査済み」にチェックし一部若しくは全部の審査が完了していない場合は、「未審査」にチェックすること。

その他（特記事項）

(※2) 未審査に場合は、その理由を記載すること。
(※3) 廃止前の「疫学研究に関する倫理指針」や「臨床研究に関する倫理指針」に準拠する場合は、当該項目に記入すること。

5．厚生労働分野の研究活動における不正行為への対応について

研究倫理教育の受講状況	受講 ■　未受講 □

6．利益相反の管理

当研究機関におけるＣＯＩの管理に関する規定の策定	有 ■　無 □(無の場合はその理由：　)
当研究機関におけるＣＯＩ委員会設置の有無	有 ■　無 □(無の場合は委託先機関：　)
当研究に係るＣＯＩについての報告・審査の有無	有 ■　無 □(無の場合はその理由：　)
当研究に係るＣＯＩについての指導・管理の有無	有 □　無 ■(有の場合はその内容：　)

(留意事項)　・該当する□にチェックを入れること。
・分担研究者の所属する機関の長も作成すること。