様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９　（共通）

科学研究費助成事業　　研究成果報告書

令和　５ 年　 ５ 月 ２０ 日現在

機関番号：１２６０１
研究種目：基盤研究(B)（一般）
研究期間：2020～2022
課題番号：２０Ｈ０４２７９

研究課題名（和文）遺伝性疾患のスクリーニングに向けた診療記録からの表現型の抽出と臨床応用評価

研究課題名（英文）e-Phenotyping from clinical text for hereditary disorders and feasibility evaluation for clinical applications

研究代表者
河添　悦昌（Kawazoe, Yoshimasa）

東京大学・医学部附属病院・特任准教授

研究者番号：１０６２１４７７

交付決定額（研究期間全体）：（直接経費）　　13,900,000円

研究成果の概要（和文）：指定難病151疾患362の症例報告テキストを収集し、70種の固有表現タグと35種の関係タグにより表現型をアノテートする基準を開発した。述べ数57,520件の表現型にアノテートを実施し、これら表現型を病名用語集（UMLS, HPO, MEDIS標準病名マスタ）の用語コードへの対応付けた。成果として、再配布の許諾が得られた179症例からなるコーパスを研究者らのHPで公開した。また、このアノテーションを再現する機械学習モデルを開発し精度評価を行った。固有表現抽出と関係抽出は比較的高い精度を示したが、表現型文字列をHPOコードに対応付ける精度は十分ではなく、今後の課題として残された。

研究成果の学術的意義や社会的意義
本研究は自然言語処理の基盤技術として、表現型（患者の状態）を抽出するための詳細なアノテーション基準を開発し、この基準でアノテートされた高品質なコーパスを構築・公開した。診療テキストを入力として、計算機がこのアノテーションを再現することで、患者の表現型（例えば、どの部位に症状が生じているのか、その症状は持続しているのか改善しているのかなど）を自動で抽出し集計できるようになる。機械学習による表現型の抽出は良好な性能を示したものの、抽出された表現型を医学用語集の用語に対応付けるエンティティリンキングの性能は十分ではないため、この性能を向上するための手法を開発することが今後の課題としてあげられた。

研究成果の概要（英文）：We collected case report texts for 362 cases of 151 designated intractable diseases and developed criteria for annotating phenotypes using 70 type of named entity tag and 35 type of relationship tags. We annotated 57,520 phenotypes and mapped these phenotypes to term codes in the disease name glossaries (UMLS, HPO, MEDIS standard disease name master). As a result, a corpus of 179 cases, for which permission for redistribution was obtained, was published on the researchers' website. A machine learning model was also developed to reproduce the annotations, and its accuracy was evaluated. Although the accuracy of unique expression extraction and relationship extraction was relatively high, the accuracy of mapping phenotype strings to HPO codes was insufficient and remains as future work.

研究分野：医療情報学

キーワード：診療記録　遺伝性疾患　表現型　自然言語処理　Phenotyping　Human Phenotype Ontology　Named Entity Recognition　Relation Extraction

様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９（共通）

１．研究開始当初の背景

- 遺伝性疾患は種類が多く頻度が低いことから、医師にとって未経験の疾患が多く存在するため、疾患の見落としが生じる可能性が高い。そのため、診療記録等から抽出した患者の表現型に関する情報と情報検索技術を活用して、候補となる遺伝性疾患や原因遺伝子を推定するための技術は重要なものとなる。これを実現するために、患者に観察される表現型を Human Phenotype Ontology（HPO）等の表現型オントロジーに対応付け、それに付随する情報を統計的に処理することで、候補となる疾患や原因遺伝子を順位付けして提示するアルゴリズムの開発は数多く行われている。しかしながら、本邦でこれを利用するためには、日本語で自然言語として記載された診療記録から複数の表現型を抽出し、HPO に含まれる約 16,000 の用語コードに対応付ける必要があり、これを行う技術は十分に確立されていない。

２．研究の目的

- 次の 3 点を目的とした。1) 日本語診療テキストから表現型（患者の状態）に関する表現を抽出し、既存の医学用語集に対応付けるための自然言語処理の基盤となるテキストコーパスを開発する。2) 開発したコーパスのアノテーションを再現する機械学習モデルを開発する。3) テキストから抽出した表現型から類似疾患をランキングして提示する既存アルゴリズムの性能を評価する。

３．研究の方法

- 本研究は東京大学大学院医学系研究科・医学部附属病院の研究倫理審査を受け、許可を得て実施した（審査番号：2019276NI, 承認日：2020/02/19）

1) NLP 技術基盤となるテキストコーパスの開発

- 実際の診療テキストを対象としてコーパスを開発することが理想であるが、個人情報保護の観点から公開のハードルが高いことや、病院独自の記録様式からテキスト部分のみを抽出すると内容が偏ることが問題となる。そのため、内容や様式が退院サマリに類似し、公開のハードルがより低い臨床医学系雑誌の症例報告テキストを利用した。希少・難治性疾患として、厚生労働省の指定難病 333 疾患を対象とし、タイトルに疾患名と「例」を両方含み、2000 年以降に出版された症例報告を J-STAGE で検索した。本文が公開されているものから 1 疾患あたり最大 4 件を限度として、症例報告の「はじめに」や「考察」の部分を除外し、「症例」セクションをコピー・ペーストによってテキストデータとした。目視で確認できる文字化けは修正したが、コピー・ペースト時に生じる空白は削除せずそのままとした。結果、151 疾患、延べ 362 症例報告のテキストを得て研究利用した。また、電子カルテシステムより指定難病 16 疾患 32 症例の退院サマリを抽出し、目視により匿名化した後に研究利用した。
- アノテーション基準は、まず仮の基準を作成し、362 症例報告テキストへのアノテーション実施と基準の修正を繰り返すことで構築した。基本方針として、症例報告テキスト中の情報をできるだけ漏らさずアノテートすることと、医学・医療的な意味を損なわない範囲で、タグの選択範囲を詳細化することとした。後者は、後段に続く用語の正規化（エンティティリンクング）を、より多くの用語集に対して行うことを可能にするための措置である。また、高度に専門的な知識を要するアノテートではなく、文字を頼りに判断できる情報をアノテートする方針とした（例：原因-結果関係は「呼吸困難による睡眠障害」の”による”のように、文字を頼りに判断可能な場合にアノテートするなど）。 膨大な数の固有表現タグと関係タグをアノテートする必要から、後述する機械学習モデルの出力をアノテータに提示し、必要に応じて修正することで、タグのつけ忘れや、タグ選択範囲の微細な違いを統一した。このような修正を繰り返し実施することで、コーパスの品質を向上させた。
- この基準によりアノテートされたテキスト中の個々の表現型に対して、3 種の医学用語集 UMLS(Unified Medical Language System、約 890 万用語)、HPO(Human Phenotype Ontology、約 16,000 用語)、MEDIS 標準病名マスタ、 約 27,000 用語）の各用語コードを人手により対応付けた。

2) アノテーションを再現する機械学習モデルの開発

- 開発したコーパスのアノテーションは、自然言語処理における固有表現抽出（NER）、関係抽出（RE）、ならびにエンティティリンキング（EL）を行なうことで再現できる。NER と RE を同時に行うモデルとして、BERT（Bidirectional Encoder Representations from Transformers）をベースとする Joint-NER-RE モデルを開発し性能を評価した。このモデルは、入力テキストを BERT 固有のトークンに分割し BERT に入力する。各トークンは BERT により固定長の埋め込みベクトルに変換され、このベクトルの系列を Conditional Random Field（CRF）に入力して NER タグを分類する。また、2 つの入力トークンに対応する NER タグの埋め込みベクトルとトークンの埋め込みベクトルとを Multilayer perceptron に入力

して関係タグを分類する。

- HPO に対するエンティティリンキングの方法を検討したところ、HPO が有する用語は約 16,000 と多く、機械学習によってこれを実現するためにはアノテーションが不足すると判断し、形態素解析辞書によってこれを実現する方法とした。

3) 表現型から類似疾患をランキングして提示する既存アルゴリズムの性能評価

- 既存の希少・遺伝性疾患のランキングアルゴリズムを実装したサービス（PubCaseFinder: https://pubcasefinder.dbcls.jp/）を利用した。このサービスは、症例に見られる表現型を HPO のコードとして入力すると、それらと関連性が高い順に希少・遺伝性疾患をランキングして提示する。提示する疾患の種別は異なる疾患用語集に基づいており、１つは OMIM（Online Mendelian Inheritance in Man)、もう一つは Orphanet（The Orphanet Rare Disease Ontology）である。疾患用語集として OMIM を選択した場合には、約 7700 種類の疾患を類似性の高い順に提示し、同様に Orphanet を選択した場合には、約 3600 種類の疾患を類似性の高い順に提示する。このサービスに指定難病 16 疾患 32 症例の退院サマリにアノテートされた表現型の HPO コードを入力し、実際の疾患が何位にランキングされるかを Mean Reciprocal Rank（MRR）によって評価した。Reciprocal Rank は情報検索システムの評価指標のひとつであり、ランクの逆数をスコアとする指標である。例えば、候補として提示された疾患のうち、実際の疾患が 20 位にランキングされた場合には 0.05 となる。この値は 0～1 の範囲をとり、1 に近いほど性能が高いと解釈できる。本研究では、1 疾患について 2 症例を用意したため、それぞれの症例の Reciprocal Rank を足して 2 で割ることで MRR を計算し、これを疾患のランキング性能として示す。

４. 研究成果

1) NLP 技術基盤となるテキストコーパスの開発

- 計 362 の症例報告を利用し 70 種の固有表現タグと 35 種の関係タグによって表現型をアノテートする基準を開発した。図 1 にアノテート基準の概要図を示す。指定難病 151 疾患 362 の症例報告テキストと、指定難病 16 疾患 32 症例の退院サマリに対してアノテーションを実施した。アノテートされた表現型の延べ数は 57、520 件となった（平均 145 件/文書）。また、これら表現型を 3 種類の病名用語集、UMLS、 HPO、 MEDIS 標準病名マスタ）の用語コードに人手により対応付けた。結果、対応付けられた表現型の延べ数は、UMLS が 46,690 件(81%)、HPO が 10,352 件(18%)、標準病名が 6,839 件(12%)となった。この結果から、約 890 万用語を有する UMLS であっても、テキストに出現する表現型の 81%をカバーするに過ぎず、テキスト中に出現する表現型（患者状態）の表現の多様さが確認された。成果物として、テキスト再配布の許諾が得られた 179 症例報告のアノテーション付きコーパスを iCorpus と名付け研究者らの HP で公開した。本コーパスは人手により詳細なアノテーションが付与された高品質なコーパスであるため、研究者による自然言語処理技術の開発を促進するものと考えている。

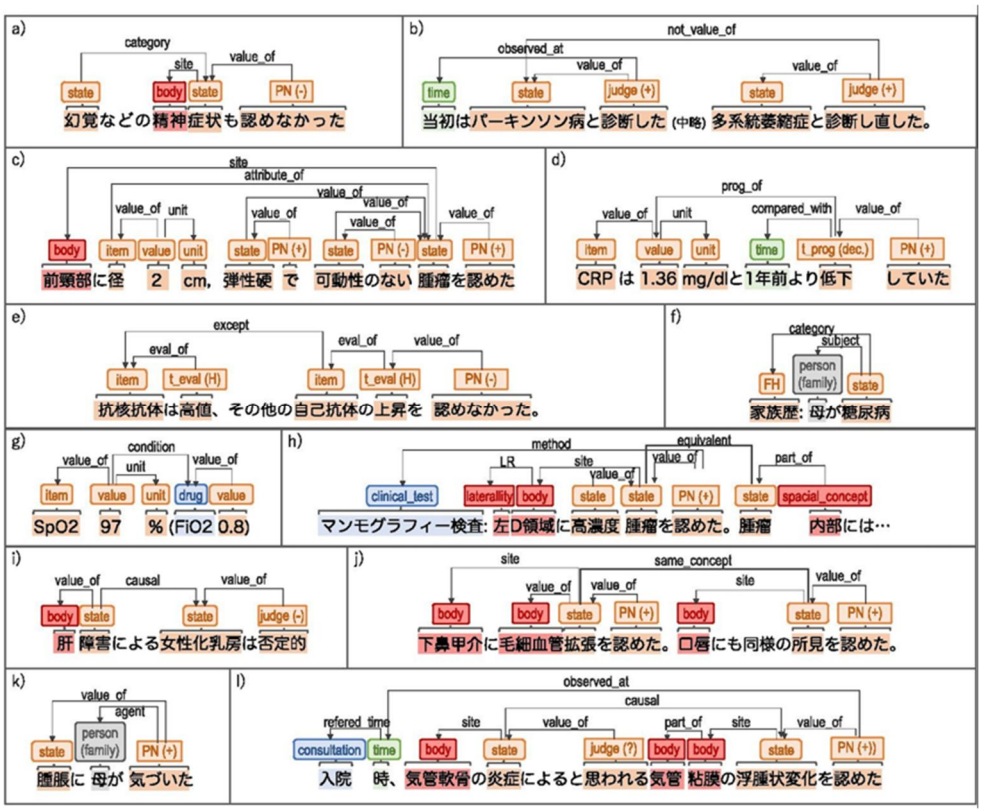

図 1. アノテート基準の概要図(J Biomed Inform. 2022 Oct;134:104200.より転載)

2) アノテーションを再現する機械学習モデルの性能

- Joint-NER-RE モデルの性能評価の結果、70 種の固有表現を抽出する NER の精度は 0.912（Micro-F1）と 0.601（Macro-F1）、35 種の関係を分類する RE の精度は 0.759（Micro-F1）と 0.611（Macro-F1）であった。このモデルを退院サマリに適用した場合、5%程度の精度が低下することが確認された。
- また、固有表現と関係の組みで同定される表現型文字列を HPO コードに対応付ける辞書ベースの手法を開発し、Recall：0.531、Precision：0.150、Micro-F1：0.234 の精度で HPO コードを同定できることを確認した。

3) 表現型から類似疾患をランキングして提示する既存アルゴリズムの性能評価

- 指定難病 16 疾患 32 症例の退院サマリにアノテートされた表現型の HPO コードを PubCaseFinder に入力し、退院サマリにおいて主病名とされた疾患がどの程度上位にランキングされるかを Mean Reciprocal Rank（MRR）によって評価した結果を表 1 に示す。表中のハイフン「-」は退院サマリ中の主病名が ORPHA もしくは OMIM の疾患名として登録されていなかったことを意味する。ORPHA と比較して OMIM は主病名が登録されていないケースが多く、これは OMIM が遺伝性疾患を取り扱うものであることが原因と考えられる。
- ORPHA においては、全身性強皮症（0.1875）、巨細胞性動脈炎（0.1834）、サルコイドーシス（0.1010）の順にスコアが高く、これはそれぞれ、平均して 5 位、5 位、10 位に主病名がランキングされたことを意味する。一方、OMIM においては、サルコイドーシス（0.0887）、ビタミン D 抵抗性くる病/骨軟化症（0.0884）、多発性嚢胞腎（0.0454）の順にスコアが高く、これはそれぞれ、平均して 11 位、11 位、22 位に主病名がランキングされたことを意味する。ORPHA と OMIM について、ハイフン「-」を除く疾患の平均 MRR は 0.0557（ORPHA）、0.0283（OMIM）であり、これはそれぞれ平均して 18 位と 35 位に退院サマリの主病名がランキングされたことを意味する。本研究は、実際の退院サマリを利用して、テキストに記録される表現型のみから既存アルゴリズムによって提示される候補疾患の妥当性を評価したものであり、申請者らの知る限りではあるが、本邦においては初の試みであると考えている。

表 1. PubCaseFinder を利用した希少・遺伝性疾患のランキング性能（Mean Reciprocal Rank）

告示番号	退院サマリ主病名（告示病名）	Mean Reciprocal Rank (ORPHA)	Mean Reciprocal Rank (OMIM)
35	天疱瘡	0.0007	0.0004
40	高安動脈炎	0.0167	0.0004
41	巨細胞性動脈炎	0.1834	0.0003
42	結節性多発動脈炎	0.0051	-
45	好酸球性多発血管炎性肉芽腫症	0.0744	-
50	皮膚筋炎／多発性筋炎	0.0159	-
51	全身性強皮症	0.1875	0.0013
54	成人スチル病	0.0247	-
55	再発性多発軟骨炎	0.0695	-
66	IgA 腎症	-	-
67	多発性嚢胞腎	0.0742	0.0454
84	サルコイドーシス	0.1010	0.0887
162	類天疱瘡	0.0008	-
220	急速進行性糸球体腎炎	0.0029	0.0017
222	一次性ネフローゼ症候群	-	-
238	ビタミン D 抵抗性くる病/骨軟化症	0.0229	0.0884
平均 MRR		0.0557	0.0283

- 上述の実験は人手によりアノテートされた表現型を利用したが、計算機で抽出した表現型とその HPO コードから同様のランキング性能を評価することが、本研究の当初のゴールであった。しかしながら、前述したようにテキストから抽出した表現型を計算機によって HPO コードに対応付けるエンティティリンキングの性能は未だ十分ではない（Micro-F1：0.234）ことから、この実験は見送ることとした。今後の課題として、エンティティリンキングの性能を向上するための手法を開発することが挙げられた。

５．主な発表論文等

〔雑誌論文〕　計3件（うち査読付論文　3件／うち国際共著　0件／うちオープンアクセス　1件）

１．著者名 Shinohara Emiko、Shibata Daisaku、Kawazoe Yoshimasa	４．巻 134
２．論文標題 Development of comprehensive annotation criteria for patients’ states from clinical texts	５．発行年 2022年
３．雑誌名 Journal of Biomedical Informatics	６．最初と最後の頁 104200～104200
掲載論文のDOI（デジタルオブジェクト識別子） 10.1016/j.jbi.2022.104200	査読の有無 有
オープンアクセス オープンアクセスとしている（また、その予定である）	国際共著 －

１．著者名 篠原 恵美子，河添 悦昌，柴田 大作，嶋本 公徳，関 倫久	４．巻 42(1)
２．論文標題 症例報告に対する網羅的な所見アノテーションのためのアノテーション基準の構築	５．発行年 2022年
３．雑誌名 医療情報学	６．最初と最後の頁 3～15
掲載論文のDOI（デジタルオブジェクト識別子） なし	査読の有無 有
オープンアクセス オープンアクセスではない、又はオープンアクセスが困難	国際共著 －

１．著者名 柴田 大作，河添 悦昌，嶋本 公徳，篠原 恵美子，荒牧 英治	４．巻 40(2)
２．論文標題 診療記録で事前学習した BERT による疼痛表現の抽出	５．発行年 2020年
３．雑誌名 医療情報学	６．最初と最後の頁 73-82
掲載論文のDOI（デジタルオブジェクト識別子） なし	査読の有無 有
オープンアクセス オープンアクセスではない、又はオープンアクセスが困難	国際共著 －

〔学会発表〕　計9件（うち招待講演　0件／うち国際学会　0件）

１．発表者名 柴田 大作，河添 悦昌，篠原 恵美子，嶋本 公徳
２．発表標題 詳細なアノテーション基準に基づく症例報告コーパスからの固有表現及び関係の抽出精度
３．学会等名 第41回医療情報学連合大会
４．発表年 2021年

１．発表者名
河添 悦昌，篠原 恵美子

２．発表標題
患者状態に関する網羅的なアノテーション基準とFHIR Conditionリソースとのマッピングの検討

３．学会等名
第41回医療情報学連合大会

４．発表年
2021年

１．発表者名
河添 悦昌，篠原 恵美子

２．発表標題
希少・難治性疾患を対象とした症例報告テキストコーパスの構築

３．学会等名
第41回医療情報学連合大会

４．発表年
2021年

１．発表者名
柴田 大作，河添 悦昌，篠原 恵美子，嶋本 公徳

２．発表標題
診療テキストの構造化に向けた症例報 告コーパスからの情報抽出

３．学会等名
第36回人工知能学会全国大会

４．発表年
2022年

１．発表者名
篠原 恵美子，河添 悦昌，柴田 大作，嶋本 公徳，関 倫久

２．発表標題
医療テキストに対する網羅的な所見アノテーションのためのアノテーション基準の構築

３．学会等名
第25回日本医療情報学会春季学術大会シンポジウム

４．発表年
2021年

１．発表者名
柴田大作，河添悦昌，篠原恵美子，嶋本公徳

２．発表標題
患者状態表現の病名交換コードへのマッピング

３．学会等名
第42回医療情報連合大会

４．発表年
2022年

１．発表者名
河添 悦昌，永島 里美，大江 和彦

２．発表標題
アレルギー情報の標準化を目指すJFAGYアレルゲン用語集とアレルゲンコードシステム

３．学会等名
第42回医療情報連合大会

４．発表年
2022年

１．発表者名
榎原 芽美，柴田 大作，篠原 恵美子，河添 悦昌，大江 和彦

２．発表標題
UMLSからの同義語を追加した形態素解析辞書を使用したPhenotypingの性能評価

３．学会等名
第27回日本医療情報学会春季学術大会

４．発表年
2023年

１．発表者名
Daisaku Shibata, Emiko Shinohara, Kiminori Shimamoto and Yoshimasa Kawazoe

２．発表標題
Towards structuring clinical texts: Joint entity and relation extraction from Japanese case report corpus

３．学会等名
MedInfo 2023, the 19th world congress on medical and health informatics

４．発表年
2023年

〔図書〕　計1件

１．著者名 河添 悦昌，篠原 恵美子	４．発行年 2022年
２．出版社 医歯薬出版	５．総ページ数 6
３．書名 医学のあゆみ283巻2号	

〔産業財産権〕

〔その他〕

症例報告コーパス（iCorpus）
https://ai-health.m.u-tokyo.ac.jp/home/research/corpus
医療AI開発学講座 - 症例報告コーパス
https://ai-health.m.u-tokyo.ac.jp/corpus

６．研究組織

	氏名 （ローマ字氏名） （研究者番号）	所属研究機関・部局・職 （機関番号）	備考
研究分担者	関　倫久 (SEKI TOMOHISA) (30528873)	東京大学・医学部附属病院・助教 (12601)	
研究分担者	篠原　恵美子 (SHINOHARA EMIKO) (40582755)	東京大学・医学部附属病院・特任助教 (12601)	

７．科研費を使用して開催した国際研究集会

〔国際研究集会〕　計0件

８．本研究に関連して実施した国際共同研究の実施状況

共同研究相手国	相手方研究機関