2C4-OS-21a-4

文献データに基づく症例検索システムの考察

A preliminary report toward a literature-based case discovery and case matching system

藤原 豊史[*1,2] Toyofumi Fujiwara
山本 泰智[*3] Yasunori Yamamoto
金 進東[*3] Jin-Dong Kim
高木 利久[*2] Toshihisa Takagi

[*1] 株式会社インテック Intec Inc.
[*2] 東京大学 The University of Tokyo
[*3] 情報・システム研究機構 ライフサイエンス統合データベースセンター Database Center for Life Science, Research Organization of Information and Systems

There are ~7,000 rare genetic diseases, and it is estimated that ~4% of newborns are affected by them. Unfortunately, less than half of them receive relevant diagnosis because the pathogenic genes of about half of rare genetic diseases are unknown. Recently, whole exome sequencing (WES) is played an important role in identifying pathogenic variants. The discovery of pathogenic variants by using WES typically requires confirmation of the common variant in multiple cases with the same rare genetic disease. However, the difficulty of finding relevant cases now remains as a major obstacle. Although some repositories such as PhenomeCentral have been developed to search for cases, they are facing a scalability issue because their population rely on submissions from clinicians and researchers. To address the scalability issue, we report a preliminary study toward utilizing more than one million records to be automatically extracted from case reports in MEDLINE.

1. はじめに

1.1 希少・難治性疾患の現状

希少・難治性疾患は，発病の機構が明らかでなく，治療方法が確立していない，そして，希少な疾病であって，長期の療養を必要とするもの，として定義されている[厚生労働省 2015]. Orphanet[Orphanet 2015]にはおよそ 7,000 の希少・難治性疾患が登録され，そのうちの 80%は子供の早い時期に発症する遺伝性疾患（染色体異常または遺伝子異常によって発症する疾患）である[Badapanda 2016]. 遺伝性疾患を治療するためには，まず診断を確定させる必要があるが，その症状からは病名を判断できず，診断不明とされる患者（未診断患者）が数多く存在する[小泉 2014]. そのため，日本医療研究開発機構（AMED）は，未診断患者の診断を行う IRUD 診療体制を全国で整えており[AMED 2015]，ゲノムを解析することで診断を行う遺伝子検査の導入を進めている. しかし，およそ 6,000～7,000 と見積もられている遺伝性の希少・難治性疾患のうち，その半数については原因となる遺伝子が同定されていないため[Boycott 2013]，遺伝子検査を適用することができない. 新生児のおよそ 4%は遺伝性の希少・難治性疾患に罹患しており，多くの患者が遺伝子検査を待ち望んでいる[Mutarelli 2014].

1.2 エクソーム解析による疾患原因遺伝子の同定

近年，単一遺伝子疾患（1 つの遺伝子の異常により発症する遺伝性疾患）の原因遺伝子の同定手法として，エクソーム解析が注目されている. エクソーム解析は全ゲノムのうち，エクソン配列（構造遺伝子の塩基配列のうちタンパク質合成の情報を持つ部分）に存在する変異を全て検出する手法である. ヒトゲノム中のエクソンを総計してもゲノム全体の 1%に過ぎないため，シークエンシングのコストや時間を大幅に削減できる. 一方で，ゲノムの一部分のみ（1%）の解析となるため，他領域の変異は検出できないが，単一遺伝子疾患に関係する変異の 85%はエクソン領域に生じている[Zang 2014]. 同一疾患と思われる複数の患者のエクソーム解析を実施し，共通して認められる変異が存在すれば，その変異を含む遺伝子を疾患原因遺伝子として同定する[Sarah 2009].

1.3 希少・難治性疾患を対象とした症例検索システム

エクソーム解析による手法では，2～3 名の患者から，疾患原因遺伝子が同定されることがある[Robinson 2011]. しかし，希少・難治性疾患はもともと患者数が少ないために，同一疾患と思われる類似の症例（疾患の症状の実例）を探すのに困難を伴う. 従来は，ジャーナルに掲載される症例報告や，未診断患者を報告するシンポジウムで情報の交換が行われてきたが，近年は，より直接的に情報を交換するために，希少・難治性疾患を対象とした症例検索システムが利用され始めている. 例えば，希少・難治性疾患の研究開発を進める国際的なコンソーシアム IRDiRC が推奨する PhenomeCentral には，1,250 件ほどの症例が登録されており（2015 年 4 月時点），疾患原因遺伝子の同定に役立っている[Buske 2015]. しかし，それら症例検索システムは医者や研究者が手動で登録した症例情報を対象に症例検索を行うため（以降，個別登録データに基づく症例検索システム，と呼ぶ），症例が大規模に集まりにくいという問題点がある. そのため，症例情報を保持する複数のデータベースを横断的に検索する取り組みも始まっている[Philippakis 2015].

1.4 文献データに基づく症例検索システム

我々は，症例が大規模に集まりにくいという「個別登録データに基づく症例検索システム」の問題点を補うために，MEDLINE[MEDLINE 2016]から取得した症例報告を対象に，類似症例を検索できる「文献データに基づく症例検索システム」を提案する（図 1）. 症例報告は，希少疾患に関する症例，未知の病態，および薬剤の副作用などを素早く共有したい場合に利用される文献の形式であり[Sudhakaran 2014]，MEDLINE には約 180 万件の症例報告が収納されている. 個別登録データに

連絡先：藤原 豊史，株式会社インテック，東京大学大学院 新領域創成科学研究科 メディカル情報生命専攻，fujiwara_toyofumi@intec.co.jp

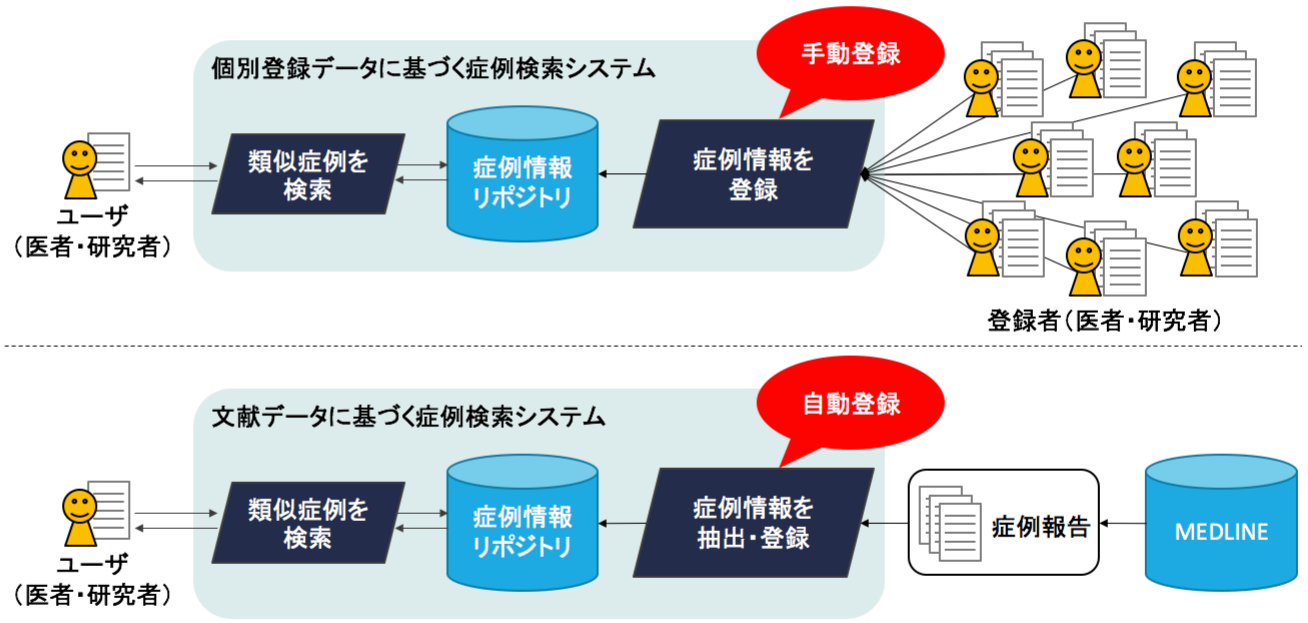

図 1.「個別登録データに基づく症例検索システム」と「文献データに基づく症例検索システム」の概要

基づく症例検索システムが手動で症例情報を登録するのに対して，我々が提案するシステムでは，約 180 万件の症例報告から自動で症例情報を抽出し，登録するため，症例を大規模に集めることができる．

2 章では症例報告の取得について述べ，3 章で症例情報を抽出するために利用するオントロジーおよびシステムついて紹介する．4 章ではオントロジーを用いた類似度計算手法を説明し．5 章で今回の調査で明らかになった課題を述べる．

2. 症例報告の取得

個別登録データに基づく症例検索システムは，医者または研究者が個別に保持している症例を収集対象とする．一方で，我々が提案するシステムは，既に MEDLINE に収められている症例報告を収集対象とする．以下に，PubMed を用いた症例報告の取得方法，および取得した症例報告について述べる．

2.1 PubMed を用いた症例報告の取得

PubMed[PubMed 2016]を用いて MEDLINE に含まれる症例報告を取得するには，症例報告に割り当てられた”Case Reports”タグを利用する．このタグは，PubMed の索引作成者がジャーナル毎の形式などを判断材料として手動で割り当てている．また，文献タイトルに”case report”または”case reports”を含む文献も症例報告として取得する．以下の PubMed クエリを用いて，症例報告リストを取得した．

- "case reports"[Publication Type] OR "case reports"[ti] OR "case report"[ti]

検索の結果，約 180 万件の症例報告のリストを取得し，それらのタイトルおよびアブストラクトを MEDLIINE から取得した．

2.2 MEDLINE に含まれる症例報告について

図 2 は MELINE に含まれる症例報告の年度別出版数である．出版数は年々増加する傾向にあり，近年では 1 年間に 5 万 5 千件ほどの症例報告が出版されている．

表 1 に各ジャーナルが今までに出版した症例報告について，その総数の上位 10 ジャーナルを示す．全ての症例報告を対象にした場合に加え，2000 年以降の症例報告のみを対象にした場合の上位 10 ジャーナルも示す．全ての症例報告を対象にした場合には，臨床医学ジャーナルとして知名度の高い Lancet や The New England journal of medicine が上位に位置している．一方で，2000 年以降の症例報告を対象にした場合では，症例報告のみを掲載対象とする，BMJ case reports や Internal medicine が上位に位置している．症例報告は，希少な疾患や未知の症例を報告するために被引用が少ない傾向にあり，近年では，Lancet などの臨床医学ジャーナルには掲載されにくくなっている[Nissen 2014]．

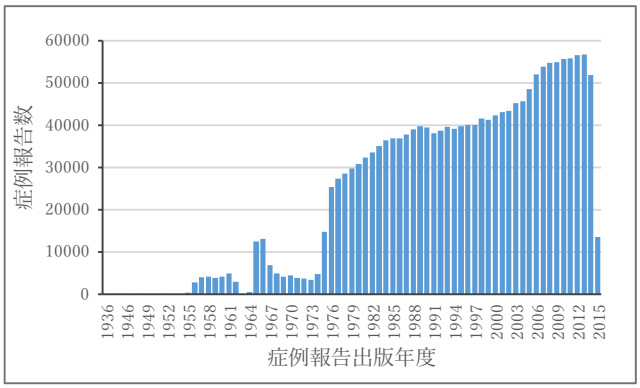

図 2. 症例報告の出版年度分布

3. オントロジーを用いた症例情報抽出

個別登録データに基づく症例検索システムでは，医者や研究者が患者の症状を元に，症例情報を手動で登録する．一方で，我々が提案するシステムは，症例報告から症例情報を自動で抽出する．以下に，症例情報の自動抽出に利用する希少・難治性疾患に関するオントロジー，および情報抽出システムについて紹介する．

3.1 希少・難治性疾患に関するオントロジー

オントロジーは，あるドメインの知識を概念化したものであり，概念間の関係が定義され，機械可読形式で表現されている[溝口 1999]．オントロジーの各概念には，それらを表現するためのラベルが付けられており，あるテキストに含まれる情報（オントロジーの概念）を抽出する際に利用することができる．

全ての症例報告		2000 年以降の症例報告	
ジャーナル名	症例報告数	ジャーナル名	症例報告数
Lancet (London, England)	10532	BMJ case reports	8480
The New England journal of medicine	9442	Internal medicine (Tokyo, Japan)	4756
BMJ case reports	8480	The Annals of thoracic surgery	4571
Neurology	7707	Gan to kagaku ryoho. Cancer & chemotherapy	3906
Clinical nuclear medicine	7704	Clinical nuclear medicine	3840
The Annals of thoracic surgery	7526	Neurology	3716
Southern medical journal	6898	American journal of medical genetics. Part A	3567
Nederlands tijdschrift voor geneeskunde	6815	International journal of cardiology	3305
Journal of the American Academy of Dermatology	6808	The New England journal of medicine	3179
Archives of dermatology	6772	Journal of the American Academy of Dermatology	2956

表 1. 症例報告件数上位 10 ジャーナル

希少・難治性疾患患者の症状を表すために，表現型の異常に関する Human Phenotype Ontology（HPO）[HPO 2016]が様々なリソースで利用されている[Buske 2015]. 表現型とは，生物のもつ遺伝的特徴が，形，色，大きさ，機能といった表面から観察できる形質として現れたものである. 例えば，「関節リウマチ」は複数の遺伝性疾患で観察され，表現型の異常として HPO に含まれる. HPO は，医学文献および Orphanet などの遺伝性疾患に関係するデータベースから抽出した用語を用いて構築されている. 現在，約 11,000 件の概念を含み，すべての概念は is-a の階層関係が定義されている. また，NCBO が運営する BioPortal[BioPortal 2016]には，HPO の他にも 515 件の生物医学オントロジー登録されており（2016 年 3 月時点），Orphanet Rare Disease ontology (ORDO)や Phenotypic Attribute and Trait Ontology（PATO）など，希少・難治性疾患に関連するオントロジーも数多く含まれている[Collier 2015].

3.2 オントロジーを用いた情報抽出システム

ライフサイエンス分野のオントロジーを用いて，テキストから情報抽出を行うシステムとして，NCBO Annotator[BioPortal 2016], OBO Annotator[Taboada 2014]，Bio LarK-CR[Groza 2015]，PubDictionaries[PubDictionaries 2016]などが存在する. NCBO Annotator はウェブサービスとして公開され，BioPortal に登録されているオントロジーを用いることができる. OBO Annotator は，HPO を用いた情報抽出に特化したシステムであり，ダウンロードをして利用することができる. Bio LarK-CR も同様に，HPO に特化したシステムであるが，公開されていないため，利用できない. しかし，MEDLINE の全文献を対象に HPO を用いて情報抽出した結果が PubAnnotation[PubAnnotation 2016]から「PubMed HPO」として公開されている. PubDictionaries は辞書を他のユーザと共有するためのプラットフォームで，オントロジーのラベルを辞書として登録することができる. また，登録されている辞書を用いて情報抽出を行うウェブサービスも提供している.

4. オントロジーを用いた類似度計算手法

PhenomeCentral は，HPO を用いて症例情報を登録し，検索クエリとなる症例も HPO で表現して，オントロジーを用いた類似度計算手法で症例間の類似度を求める. 我々が提案するシステムも同様の手法を用いて類似度を計算するが，HPO だけでなく，複数のオントロジーを用いた場合の類似度計算手法も取り入れる.

1つのオントロジーを用いた類似度計算手法は数多く開発されているが，Gene Ontology[GeneOntology 2016]を用いた類似度計算手法として simGIC[Pesquita 2007]が開発され，ライフサイエンスの分野で広く利用されている[Buske 2015]. この手法は，予め概念毎の Information Content（*IC*）を計算しておく（式 1）. これは，概念 c がコーパス中に出現する確率 *P(c)*を元に計算される. 例えば，OMIM に登録されている 7,000 件の遺伝性疾患に対する記述の中で，c が 10 件の記述に含まれていた場合，*P(c)*は 10/7000 となる.

$$IC(c) = -logP(c) \quad (1)$$

式 2 で，simGIC による症例 P と症例 Q の類似度 *Sim(P,Q)*を計算する. g^P および g^Q は，P および Q に割り当てられた c のセットを表す.

$$Sim(P,Q) = \frac{\sum_{c \in g^P \cap g^Q} IC(c)}{\sum_{c \in g^P \cup g^Q} IC(c)} \quad (2)$$

式 3 で，複数のオントロジーを用いた類似度の計算が可能である[溝口 2008]. *R(P,Q)*は，n 個のオントロジーを用いた場合の P と Q の類似度となる.

$$R(P,Q) = \sum_{i=1}^{n} w_i * Sim_i(P,Q) \quad (3)$$

$Sim_i(P,Q)$ は i 番目のオントロジーによって導き出された P と Q の類似度である. w_i は i 番目のオントロジーによる類似度への重みを示す. 重みを掛け合わせることにより，オントロジーごとの重要性に合わせた類似度結果を調整することができる.

5. 終わりに

「文献データに基づく症例検索システム」を提案するため本稿の調査を行ったが，今後検討すべき課題も明らかになったので，以下に述べる.

(1) アブストラクトを取得できない症例報告

約 180 万件の症例報告のうち，アブストラクトを含む症例報告は約 100 万件であった. アブストラクトがない症例報告は，症例が簡潔な文章と写真・図でまとめてあり，ページ数が少ないもの

ジャーナル名	症例報告数
Lancet (London, England)	9491
The New England journal of medicine	8917
Medicina clínica	4748
Archives of dermatology	4599
AJR. American journal of roentgenology	4563
Revista clínica española	4502
Presse médicale (Paris, France : 1983)	4484
JAMA	4196
Contact dermatitis	4097
Gastrointestinal endoscopy	3996

表 2. アブストラクトがない症例報告件数上位 10 ジャーナル

が多い. 表 2 は, 各ジャーナルが今までに出版したアブストラクトがない症例報告について, その総数の上位 10 ジャーナルを示す. アブストラクトがない症例報告については, 本文からの症例情報取得なども検討する必要がある.

(2) オントロジーを用いた情報抽出システムの結果の違い

PubDictionaries と OBO Annotator を使用して, ランダムに選択した 1 万件の症例報告を対象に, HPO を用いて情報を抽出した. 各症例報告から抽出した HPO の概念数の平均は, PubDictionaries が約 4.6 件で, OBO Annotator が約 2.3 件となった. 図 3 は症例報告から抽出した HPO 概念数の分布である.

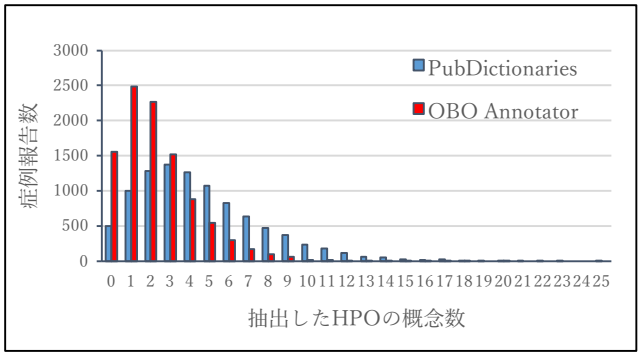

図 3. 症例報告から抽出した HPO 概念数の分布

また, 表 3 は NCBO Annotator, OBO Annotator, Bio-LarK-CR, PubDictionaries で, PubMed ID : 25351760 の書誌情報から HPO を用いて情報を抽出した結果である. 図 3 および表 3 の結果から, 同じオントロジーを用いても, 利用するシステムのアルゴリズムによって結果が大きく異なることがわかる. 希少・難治性疾患の類似症例検索にどのシステムの結果が有効であるか, 比較評価する必要がある.

システム名称	抽出した HPO 概念数
NCBO Annotator	9
PubDictionaries	8
Bio LarK-CR	3
OBO Annotator	2

表 3. 各システムが抽出した HPO の概念数

参考文献

[厚生労働省 2015] 厚生労働省: 難病医療費助成制度概要, http://www.mhlw.go.jp/file/06-Seisakujouhou-10900000-Kenkoukyoku/0000087752.pdf

[Orphanet 2016] Orphanet: http://www.orpha.net/consor/cgi-bin/index.html

[Badapanda 2016] Badapanda, Chandan., et al: Clinical & Medical Biochemistry : Open Access RareDDB : An Integrated Catalog of Rare Disease Database, Clinical & Medical Biochemistry, Open Access 2016 (2016).

[小泉 2014] 小泉二郎: 希少難病問題 : 置き去りの 6,700 疾患, 医学の歩み, 医歯薬出版 , (2014).

[AMED 2015] 日本医療研究開発機構: IRUD(未診断疾患イニシアチブ)について, http://www.amed.go.jp/content/files/jp/release/20150730.pdf

[Boycott 2013] Boycott, Kym M., et al: Rare-disease genetics in the era of next-generation sequencing: discovery to translation, Nature Reviews Genetics, 14. 10 (2013): 681-691.

[Mutarelli 2014] Mutarelli, Margherita., et al: A community-based resource for automatic exome variant-calling and annotation in Mendelian disorders, BMC Genomics, 15. 3 (2014): 1.

[Zang 2014] Zhang, Xuejun: Exome sequencing greatly expedites the progressive research of Mendelian diseases, Frontiers of Medicine in China, 8. 1 (2014): 42-57.

[Sarah 2009] Ng, Sarah B., et al: Targeted capture and massively parallel sequencing of 12 human exomes, Nature, 461. 7261 (2009): 272-276.

[Robinson 2011] Robinson, Pn., et al: Strategies for exome and genome sequence data analysis in disease-gene discovery projects, Clinical Genetics, 80. 2 (2011): 127-132.

[Buske 2015] Buske, J., et al: PhenomeCentral: A Portal for Phenotypic and Genotypic Matchmaking of Patients with Rare Genetic Diseases, Human Mutation, 36. 10 (2015): 931-940.

[Philippakis 2015] Philippakis, Anthony A., et al: The Matchmaker Exchange : A Platform for Rare Disease Gene Discovery, Human Mutation, 36. 10 (2015): 915-921.

[MEDLINE 2016] MEDLINE: http://www.nlm.nih.gov/pubs/factsheets/medline.html

[Sudhakaran 2014] Sudhakaran, Sivakumar and Surani, Salim: " The Role of Case Reports in Clinical and Scientific Literature ", Austin J Clin Case Rep, 1. 2 (2014): 1-2.

[PubMed 2016] PubMed: http://www.ncbi.nlm.nih.gov/pubmed

[Nissen 2014] Nissen, Trygve., et al: The clinical case report : a review of its merits and limitations, BMC Res Notes, 7. 216 (2014).

[溝口 1999] 溝口理一郎: オントロジーと知識処理, 大阪大学, (1999).

[HPO 2016 年] Human Phenotype Ontology: http://human-phenotype-ontology.github.io

[BioPortal 2016] BioPortal: http://bioportal.bioontology.org

[Collier 2015] Collier, Nigel., et al: PhenoMiner: from text to a database of phenotypes associated with OMIM diseases, Database, 2015 (2015): bav 104.

[Taboada 2014] Taboada, M., et al: Automated semantic annotation of rare disease cases: a case study, Database, 2014 (2014): bau045.

[Groza 2015] Groza, T., et al: Automatic concept recognition using the Human Phenotype Ontology reference and test suite corpora, Database, 2015 (2015): bav005.

[PubDictionaries 2016] PubDictionaries: http://pubdictionaries.org

[PubAnnotation 2016] PubAnnotation: http://pubannotation.org

[GeneOntology 2016] Gene Ontology Consortium: http://geneontology.org

[Pesquita 2007] Pesquita, C., et al: Evaluating go-based semantic similarity measures, InProceedings of 10thAnnualBio-OntologiesMeeting, 37. 40. (2007).

[溝口 2008] 溝口祐美子: オントロジーを用いた文書間類似度計算手法, 人工知能と知識処理, 108. 119 (2008): 87-92.