BiomedCurator：医学・生物学文献からの構造化データ抽出のためのデータキュレーションシステムの開発

Mohammad Golam Sohrab [1]　Khoa N. A. Duong [1]　池田 修己 [1]　Goran Topić [1]
夏目 やよい [2]　黒田 正孝 [2]　伊藤 眞里 [2]　高村 大也 [1]
[1] 産業技術総合研究所 人工知能研究センター
[2] 医薬基盤・健康・栄養研究所 AI 健康・医薬研究センター

{sohrab.mohammad, goran.topic, ikeda-masami, takamura.hiroya}@aist.go.jp
{natsume, m-kuroda, mari}@nibiohn.go.jp

概要

医学・生物学分野の論文から，事前に選択した属性（疾患，投与薬剤，投与量，投与期間，被験者の年齢，結果など）に関する情報を，構造化データとして抽出するキュレーションシステム BiomedCurator のを開発を行なった．本システムは，関係抽出のためのテキスト生成モデル，エンティティ認識，テキスト分類モデル，知識ベース上のエントリからの情報検索とエンティティ・リンキング，パターンに基づく複数属性の情報抽出を行う自然言語処理アプローチによって構成されている．BiomedCurator はウェブサイト（https://biomed-text.airc.aist.go.jp/biomedcurator/）において公開している．

1　はじめに

医学・生物学分野の論文や臨床試験の報告書には，適用された薬，対象疾患，投与量，投与期間，被験者の年齢，結果などの情報が提供されている．これらの情報は，創薬や医薬品開発のためのデータマイニングや統計分析に有用であり，通常は専門家が論文を読んで抽出し，図 1 に示すような属性と値が対応づけられた 2 次元のスプレッドシート形式の構造化データを手作業で作成している．本稿では，このような構造化データを自動生成（以下，データキュレーション）するシステムを提案する．

データキュレーションのタスクには，エンティティ認識，エンティティ・リンキング，関係抽出，テキスト分類など，さまざまな自然言語処理技術を用いたアプローチが必要となる．また，このタスクの特徴は，訓練データとなる，人間の専門家によってキュレーションされたデータがスプレッドシート形式で提供される一方，抽出すべき情報が論文中のどこの位置に記述されているかが明示されない．したがって，通常のエンティティ認識の学習データとは異なり，BIO タグは注釈されていない．また，キュレーションされたデータ中では，元の論文と異なる用語や表現が使われている可能性がある．

本稿では，これらの技術的課題を 2.2 節で述べる手法で解決する．さらに，特に肺疾患に関するドメインについて手法を実装し，ユーザーが指定する PubMed ID あるいは ClinicalTrials.gov ID に該当する文献から，61 個の属性を自動データキュレーションするシステム BiomedCurator について報告する．

2　関連研究

テキストマイニングを使用した医学・生物学情報の検索をサポートする方法およびウェブツールがいくつか提案されている．例えば，エンティティ情報の抽出を文書レベルで行う際に，生成的アプローチを採用したものなどがある [1]．文のエンコードから文書レベルのグラフを作成し，グラフのエッジ表現から文書レベルの関係を抽出する方法 [2] や，文書レベルの n 項関係を獲得する手法 [3] などが提案されている．また，pubmedKB [4] は，バリアント，遺伝子，疾患，および化学物質という 4 つの医学・生物学エンティティタイプ間のセマンティックな関係を抽出して視覚化するウェブサーバであり，多数の PubMed 抄録から意味関係の抽出が可能である．構造化された関係タプルを抽出するための文節抽出とメタパターン発見を組み込んだ CPIE（Clause+Pattern-guided Information Extraction）フレームワーク [5] や，BERT ベースの言語モデルを使用した遺伝子–疾患の関係抽出手法 [6]，文献から遺伝子型–表現型の関係抽出のためのパイプライン [7] が提案されている．

Reference Information			...	Intervention Characteristics					...	Disease Characreristics		...	Reference details		
reference_type	reference_id	associated_clinical trials	...	drug/therapy	refererence_drug therapy	dose	duration	CAS id	...	disease name	stage	...	source	...	Year
PubMed	23868010	UMIN000001779		Prednisolone	[NA]	80 mg	[NA]	50-24-8		Lung Cancer/Non-Small Cell	IIA, IIB		Full Text		2013
...															
PubMed	27924059	CEEOG0106, ML20033		Erlotinib	[NA]	150 mg	[NA]	183321-74-6		Lung Cancer	IIIB, IV		Full Text		2017
ClinicalTrial	NCT02759835	[NA]		Osimertinib	[NA]	80 mg	[NA]	1421373-65-0		Lung Cancer, Non-Small Cell	[NA]		Clinicaltrial-No result		2016
...															
ClinicalTrial	NCT02773238	[NA]		Radiotherapy	[NA]	[NA]	[NA]	[NA]		Lung Cancer, Non-Small Cell	IIB, IIIB		Clinicaltrial-No result		2016

図 1　スプレッドシート形式の構造化データの例．1 行目はカテゴリ，2 行目は属性を示す．

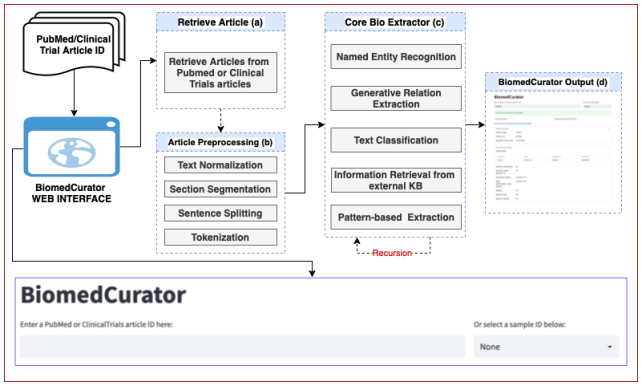

図 2　BiomedCurator のワークフロー．入力された PubMed あるいは ClinicalTrials.gov の ID をもとに，(a) から (d) のプロセスを通じて解析し，各属性の値として出力する．プロセス (c) の結果の一部は再利用され，他の属性を予測する入力機能として再結合されることを示す．

本稿で対象とするデータキュレーションでは，エンティティ抽出，関係抽出，エンティティリンキングなどに基づくより幅広い情報の抽出を行う．

3　提案システム

データ（3.1 節），キュレーションのための 5 つの主要な手法（3.2 節），そのウェブアプリケーション（3.3 節）の順で詳述する．

3.1　データセット

ここでは，肺がん，特発性肺線維症（IPF），間質性肺炎（IP）と線維症について過去 5 年間に発表された無料でアブストラクトと本文が入手可能な PubMed の論文を選択した．抽出する情報を 11 個のカテゴリに分類し，さらにサブカテゴリに分割して合計 61 個の属性を設けた．これらの属性は創薬の経験を持つ薬理学者がキュレータの生物学者との話し合いの中で決定した．次に，手作業でキュレーションを行なった．キュレーションには生物学者と編集者が参加し，品質保証と品質管理チェックを行った．構築した構造化データセットの概要を図 1 に示す [1)]．各カテゴリ，サブカテゴリ，属性の詳細な内容と使用するキュレーション手法は，付録およびウェブサイト [2)] に記載した．

1)　データセットの公開は検討中．
2)　https://github.com/aistairc/BiomedCurator

3.2　キュレーション手法

3.2.1　関係抽出のための生成的アプローチ

本稿の関係抽出は，抽出すべき情報の論文中の出現位置が訓練データ中で明示されていない，論文中とは異なる用語や表現が構造化データで使われる可能性がある，という技術的課題がある．そのため，生成問題として n 項関係抽出タスクを定式化する．入力テキストに対し，予め定めた構造を表す単語系列を生成する．例えば，*eligible patients received up to six cycles of pemetrexed, 500 mg/m(2) plus cisplatin, 75 mg/m(2) (day 1) or gemcitabine, 1000 mg/m(2) (days 1 and 8) plus cisplatin, 75 mg/m(2) (day 1). os and toxicity were assessed.* という入力に対し，次のような系列を生成する：

```
[start]
[drug] gemcitabine [/drug] [dose] 1000 mg/m2 [/dose]
[and]
[drug] cisplatin [/drug] [dose] 75 mg/m2 [/dose]
[or]
[drug] pemetrexed [/drug] [dose] 500 mg/m2 [/dose]
[and]
[drug] cisplatin [/drug] [dose] 75 mg/m2 [/dose]
[end]
```

これは，（gemcitabine, 1000 mg/m2）と（cisplatin, 75 mg/m2），あるいは（pemetrexed, 500 mg/m2）と（cisplatin, 75 mg/m2）という薬品-投与量の関係を表す．

長い系列に対応できるエンコーダ・デコーダモデルとして，BigBirdPegasus [8] を採用し，上記の生成的アプローチによる関係抽出を行う．このアプローチを用いるのは，`dose`, `drug`, `route of administration` の 3 個の属性である．

3.2.2　エンティティ抽出

エンティティ抽出に帰着できる属性については，SciBERT [9] を用いた．特に，`ethnicity` については，同じラベルを持つ OntoNotes 5 [10] を訓練データとして SciBERT をファインチューニングした．`Biomarker name` については，同様に BioNLP13CG [11] でファインチューニングした．それ以外の属性については，擬似訓練データに基づく distant supervision を用いた．具体的には，構造化

データのエンティティと表層的に類似している文を取得し，文字列マッチングにより擬似的にラベル付けした．

3.2.3 テキスト分類

いくつかの属性については，SciBERT により入力テキストをエンコードし分類する方法を用いた．また，他の属性の値が手がかりになる属性については，RandomForest[3] に基づく分類器を用いた．例えば，association については，marker_type, marker_nature, phenotype の属性値を分散表現で表し連結したベクトルを入力とし RandomForest を用いた．

3.2.4 正規表現

reference_id, grade, stage, total_sample_number などの属性については，正規表現により値を抽出した．詳細はプロジェクトページに記述した．

3.2.5 外部知識ベースからの情報検索

CAS ID, ChEMBL ID, DrugBank ID, Entrez ID, Uniprot ID, HGVS Name, Rs ID, KEGG Pathway Name の 8 個の属性については，外部知識ベース[4] から取得した．

3.3 ウェブアプリケーション

BiomedCurator のワークフローを図 2 に示す．ユーザから入力された PubMed あるいは ClinicalTrials.gov の ID をもとに，(a) オンラインデータベースから対応する論文を取得し，(b) 前処理を経て，(c) さまざまな種類の情報を抽出するように設計された 5 つの主要なアプローチに適用される．最後に，抽出された情報が返され，(d) 各属性の値を表示する．

4 実験

4.1 データ

PubMed (2,570 論文) と ClinicalTrials.gov (2,371 論文) から構築したキュレーションデータで実験を行った．データセットの内訳と統計値を表 1 に示す．

3) https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html

4) https://commonchemistry.cas.org, https://go.drugbank.com, https://www.ncbi.nlm.nih.gov/gene/, https://www.genecards.org, https://www.uniprot.org/uniprot/, https://www.ncbi.nlm.nih.gov/CBBresearch/Lu/Demo/LitVar/api.html, https://www.ncbi.nlm.nih.gov/CBBresearch/Lu/Demo/LitVar/api.html, https://www.genome.jp/kegg/pathway.html

データの前処理は 4 つのステップから成る：(1) テキスト正規化，(2) 文分割，(3) 節分割および段落分割，(4) トークン化．

テキスト正規化 XML タグの削除や，連続する空白を単一の空白にする作業，特殊記号を空白にするする作業を行う．その後，ftfy [12] を用いて NFC 形式でユニコード正規化を行う．

文分割 テキスト正規化後に，GENIA 文分割器[5] を用いて文分割を行う．

節分割および段落分割 drug と dose の関係など，薬品-投与量の関係など，単一文でなく複数文に渡って記述されている属性があるため，入力テキストを節や段落の単位に分割する．まず，XML 形式で与えられている入力テキストのメタデータを用いる．分割に利用できるメタデータがない場合は，Abstract, Introduction, Method, Approach, Results などのキーワードに基づくルールにより分割を行う．

トークン化 最後に BigBirdPegasus の PegasusTokenizer と SciBERT の BertTokenizer を用いてトークン化を行う．

4.2 計算

すべてのモデルは AdamW [13] を用い学習率 3e–5 で最適化した．関係抽出のための生成モデルは，8 GPU 上で total batch size 32 で 50 エポック訓練した．また，詳細は割愛するが，カリキュラム学習を採用している．エンティティ抽出モデルは，単一 GPU 上で batch size 32 で 5 エポック訓練した．計算には，NVIDIA A100 for NVLink 40GiB を用いた．

5 結果と考察

表 2 に提案手法による PubMed データセットに含まれる 17 個の属性へのデータ抽出性能を，適合率 (P)，再現率 (R) および F スコア (F) で示した．F スコアに基づくほとんどの属性への抽出性能は良好に機能していることを示していた．一方，属性 duration, grade, disease_name, phenotype への性能は他に比べて低かった．disease_name のモデルは，distant supervision による擬似訓練データで学習されているが，病名には同義語が多いことからこの擬似訓練データにノイズが多く含まれていることが再現率低下につながったと考えられる．対照的に，単一回答のみで構成される PubMed データセットに含まれる 6 個の属性については，十分に高い結果が得られた (表 3)．

ClinicalTrials.gov データセットへのデータ抽出の

5) http://www.nactem.ac.uk/y-matsu/geniass/

表 1 PubMed と ClinicalTrials.gov から選択収集した各キュレーションデータセットの内訳および統計

データセット	Split	#Docs	Avg. Tokens/Doc	Avg. Sec./Doc	Avg. Para./Doc
PubMed	Train	1542	3296.52	10.90	37.89
	Dev	514	3037.13	10.38	35.32
	Test	514	3277.14	10.87	36.96
ClinicalTrials.gov	Train	1421	1395.08	8.41	15.34
	Dev	475	1296.97	8.39	15.00
	Test	475	1343.81	8.43	15.07

表 2 PubMed データセットに含まれる 17 個の属性（複数回答）へのデータ抽出の評価結果

属性名	P	R	F (%)
associated_clinical trials	54.24	60.38	57.10
Relation of (drug/therapy-dose)	53.77	50.59	52.13
duration	4.91	38.57	8.71
Cell line/Model Name	44.07	41.67	42.83
study_type	85.80	82.43	84.08
ethnicity	27.72	73.29	40.23
grade	10.53	21.43	14.12
phase	18.07	82.86	29.67
disease_name	91.67	5.66	10.66
stage	44.34	70.19	54.35
association	97.00	97.00	97.00
phenotype	9.09	45.19	15.14
p_value	33.37	32.68	33.02
application	94.00	95.00	94.50
allocation	39.02	72.73	50.79
masking	37.50	46.15	41.38
authors	86.35	87.35	86.85

表 3 PubMed データセットに含まれる 6 個の属性（単一回答）へのデータ抽出の評価結果

属性名	Accuracy (%)
type of alteration	87.44
phenotype_alteration	93.00
significance	99.95
author_conclusion	100.00
title	95.33
year	99.61

表 4 ClinicalTrials.gov データセットに含まれる 21 個の属性（複数回答）へのデータ抽出の評価結果．BNAMIR は "biomarker name as modified in reference" の省略表記

属性名	P	R	F (%)
Relation of (drug/therapy-dose)	38.36	33.15	35.5
duration	2.53	81.82	4.90
disease_name	61.66	73.44	67.04
disease_sub_category	11.75	51.43	19.13
stage	23.44	85.80	36.82
BNAMIR	1.29	7.38	2.20
phenotype	22.43	51.03	31.16
total_sample_number	74.95	75.11	75.03
patient_number (case)	74.95	75.11	75.03
age(case)	87.77	87.96	87.86
gender(case)	99.37	100.00	99.68
ethnicity (case)	55.56	71.43	62.50
sponsor & collaborator	66.67	88.93	76.20
phase	97.31	97.97	97.64
inclusion_criteria	73.05	83.41	77.89
authors	95.61	94.37	94.99
intervention_model	96.25	96.48	96.37
masking	97.92	98.38	98.15
primary_purpose	98.84	99.07	98.96
association	98.00	98.00	98.00
application	99.00	99.00	99.00

表 5 ClinicalTrials.gov データセットに含まれる 3 個の属性（単一回答）へのデータ抽出の評価結果

属性名	Accuracy (%)
trial_status	56.21
title	93.89
year	86.11

評価結果を，表 4 および 5 に示す．属性 `duration`，`disease_sub_category`，`BNAMIR` を除くほとんどの属性に対する抽出性能は良好である．属性 `duration` への性能の低さは，このゴールドエンティティは通常，1〜3 桁の数字の後に時間の単位を表す単語が続くことが原因である（例：24 hours，120 days, 2 weeks, 3 months など）．情報の正規化，言い換え，対象の論文以外からの情報の付加が専門家による手動のキュレーションデータに含まれていた場合，モデルによる論文内容からの検索と抽出精度を低下させ，評価が困難であった．

6 おわりに

医学・生物学文献からのデータキュレーションシステム BiomedCuration の開発を報告した．本システムは，PubMed および ClinicalTrials.gov に由来する論文から自然言語処理アプローチによる解析を通じて情報を抽出し，61 個の属性の値として出力する．2 つのデータセットに対する F スコアと精度による性能評価から，システムが機能していることを示す結果が得られた．また，本システムはインタラクティブなウェブアプリケーションとして公開している．今後は，システムのさらなる改善のために，n 項関係抽出などの新しい機能の実装を計画している．

謝辞

本稿で紹介した研究は，官民研究開発投資拡大プログラム（PRISM）「新薬創出を加速する症例データベースの構築・拡充／創薬ターゲットの推定アルゴリズムの開発」の助成を受けたものです．

参考文献

[1] Kung-Hsiang Huang, Sam Tang, and Nanyun Peng. Document-level entity-based extraction as template generation. In **Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing**, pp. 5257–5269, Online and Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics. https://aclanthology.org/2021.emnlp-main.426.

[2] Fenia Christopoulou, Makoto Miwa, and Sophia Ananiadou. Connecting the dots: Document-level neural relation extraction with edge-oriented graphs. In **Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)**, pp. 4925–4936, Hong Kong, China, November 2019. Association for Computational Linguistics. https://aclanthology.org/D19-1498.

[3] Robin Jia, Cliff Wong, and Hoifung Poon. Document-level n-ary relation extraction with multiscale representation learning. In **Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)**, pp. 3693–3704, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. https://aclanthology.org/N19-1370.

[4] Peng-Hsuan Li, Ting-Fu Chen, Jheng-Ying Yu, Shang-Hung Shih, Chan-Hung Su, Yin-Hung Lin, Huai-Kuang Tsai, Hsueh-Fen Juan, Chien-Yu Chen, and Jia-Hsin Huang. pubmedKB: an interactive web server for exploring biomedical entity relations in the biomedical literature. **Nucleic Acids Research**, Vol. 50, No. W1, pp. W616–W622, 05 2022. https://doi.org/10.1093/nar/gkac310.

[5] Xuan Wang, Yu Zhang, Qi Li, Yinyin Chen, and Jiawei Han. Open information extraction with meta-pattern discovery in biomedical literature. **Proceedings of the 2018 ACM International Conference on Bioinformatics, Computational Biology, and Health Informatics**, 2018.

[6] Chuan Deng, Jiahui Zou, Jingwen Deng, and Mingze Bai. Extraction of gene-disease association from literature using biobert. **The 2nd International Conference on Computing and Data Science**, 2021.

[7] Wenhui Xing, Junsheng Qi, Xiaohui Yuan, Lin Li, Xiaoyu Zhang, Yuhua Fu, Shengwu Xiong, Lun Hu, and Jing Peng. A gene–phenotype relationship extraction pipeline from the biomedical literature using a representation learning approach. **Bioinformatics**, Vol. 34, No. 13, pp. i386–i394, 06 2018. https://doi.org/10.1093/bioinformatics/bty263.

[8] Manzil Zaheer, Guru Guruganesh, Avinava Dubey, Joshua Ainslie, Chris Alberti, Santiago Ontañón, Philip Pham, Anirudh Ravula, Qifan Wang, Li Yang, and Amr Ahmed. Big bird: Transformers for longer sequences. **CoRR**, Vol. abs/2007.14062, , 2020. https://arxiv.org/abs/2007.14062.

[9] Iz Beltagy, Kyle Lo, and Arman Cohan. SciBERT: A pretrained language model for scientific text. In **Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)**, pp. 3615–3620, Hong Kong, China, November 2019. Association for Computational Linguistics. https://aclanthology.org/D19-1371.

[10] Sameer S. Pradhan, Eduard Hovy, Mitch Marcus, Martha Palmer, Lance Ramshaw, and Ralph Weischedel. Ontonotes: A unified relational semantic representation. In **International Conference on Semantic Computing (ICSC 2007)**, pp. 517–526, 2007.

[11] Sampo Pyysalo, Tomoko Ohta, Rafal Rak, Andrew Rowley, Hong-Woo Chun, Sung-Jae Jung, Sung-Pil Choi, Jun'ichi Tsujii, and Sophia Ananiadou. Overview of the cancer genetics and pathway curation tasks of BioNLP shared task 2013. **BMC bioinformatics**, Vol. 16, No. 10, pp. 1–19, 2015.

[12] Robyn Speer. ftfy. Zenodo, 2019. Version 5.5.

[13] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In **International Conference on Learning Representations, ICLR 2019, New Orleans, USA**, 2019. https://openreview.net/forum?id=Bkg6RiCqY7.

A 付録

表 6 各カテゴリとそのサブカテゴリである属性（1 列目と 2 列目）の詳細な内容（4 列目）および使用するアプローチの一覧（3 列目．PE：パターンに基づく抽出，RE：関係抽出，EE：エンティティ抽出，EL：エンティティ・リンキング，TC：テキスト分類）．FV（Fixed Value）はキュレートデータの特定の値，NA は現時点で未対応であることを意味する．

カテゴリ	属性名	NLPT	説明
Reference Information	reference_type	FV	Source of the article. (e.g.) PubMed or ClinicalTrial.gov.
	reference_id	PE	Unique PubMed ID or Clinical trial id of the curated document.
	associated_clinical trials	PE	Provides the associated clinicaltrial ids for which the results were published.
	s_no	NA	Each assertion has given a unique number.
Intervention Characteristics	drug/therapy	RE	Captured the list of authors focus drug/s of case group.
	reference_drug/therapy	RE	Captured the list of authors focus drug/s of reference group.
	treatment_details	NA	Detail description of the treatment, including but not limited to patient details, drug/therapy, dose/cycles, duration, route, schedule, analysis.
	dose	RE	It represents the concentration value of the drug used in the given reference.
	route of administration	RE	The route through which the drug is administered.
	duration	EE+PE	Time period of the treatment.
	CAS id	EL	Chemical abstracts service registry number of the drug.
	ChEMBL	EL	Unique id as provided by ChEMBL.
	drug bank id	EL	Drug bank id for the given drug.
	approved_drug	NA	Name of the drug which is approved by any approval authority.
	approval_authority	NA	Name of the organization/institution has the authority to approve the respective drug. (e.g.) FDA.
Disease Characteristics	disease_name	EE+PE	Name of the focused indication for which the biomarker was studied.
	disease_sub_category	EE+PE	Represents the subtype or any state of the disease mentioned in the given reference.
	Stage	PE	Stages of the disease. (e.g.) Stage I, II, III, IV, etc.
	Grade	PE	Grading of the disease. (e.g.) Grade I, II, III, IV, etc.
	Histopathology	EE	Additional details of the disease mentioned in the article. (e.g.) Stage, histopathology etc.
Biomarker Details	BNAMIR	EE	Complete name of the biomarker. Abbreviations are extended for ease of understanding.
	marker_type	EE	Represents the type of the biomarker based on the techniques used to measure the biomarker. (e.g.) Biochemical, Genomic, etc.
	marker_nature	EE	Represents the chemical nature of the biomarker based on the techniques used to measure the biomarker. (e.g.) Protein, Gene, Lipid, etc.
	Entrez ID	EL	Unique ID as provided by the NCBI Entrez gene database for each gene.
	UniProt ID	EL	Protein accession number of UniProtKB.
	type_of_variation	EL	Represents standard HGVS constructs unique for each variation.
	rs_id	EL	Represents the unique reference number for each SNP at a specific position. Taken from NCBI dbSNP site. (e.g.) rs763110.
	HGVS Name	EL	Field describes nucleotide/DNA (c.) change as per the HGVS format (the nucleotide/genomic numbering should be as in article only).
Biomarker association with outcomes	association	TC	Describes about the high level type/category of biomarker association with outcomes. Associations are of 5 types: Gene - drug relationships; Gene - gene interactions; Gene - pathway relationships; Gene - phenotype relationships; Gene - transcript information
	marker_alteration	EE	Represents the type of alteration or measurement done for biomarker. (e.g.) Gene expression, Polymorphism, Biomarker level, etc.
	type of alteration	PE	Represents the modification of the marker mentioned in the article. (i.e.) change of biomarker expression or levels. (e.g.) High, Low, Decreases, Association, Upregulation, etc.
	phenotype	TC	Biomarker associates with any phenotype character, end point, outcome, any physiological process and other biomarkers of the study sample.
	phenotype_alteration	PE	Represents the state of change for the outcome variables which are associated with the studied biomarker.
	significance	PE	Represents the level of significance of P value between different groups. (e.g.) Non-significant or Significant.
	p_value	PE	P-value (Significance) between the different groups for comparison of biomarker result values or any other values related to biomarker. (e.g.) P=0.016
Utility	application	TC	Denotes the utility of the biomarker for a given condition in a specific reference (either clinical trial or pubmed article).
	author_conclusion	TC	Represents the utility of the biomarker from the author's perspective in the given reference. Yes indicates that author, in the reference, supports the application of the biomarker for the given indication. No indicates that author in the reference does not support the application of the biomarker for the given indication.
	evidence_statement	NA	Gives the structured description of the application text of the biomarker in a given condition specific to each reference and clinical status.
Study characteristics	study_type (Clinical/PreClinical)	PE	Represents the status of the clinical study. (e.g.) Clinical, Preclinical, etc.
	Cell line/ Model Name	EE	Represents the cell lines used in the preclinical model/It represents the preclinical model. (e.g.) Mouse, rat etc.
	total_sample_number	PE	Denotes total number of participants from both study and reference sample group in a particular study.
	patient_number (case)	PE	To capture the study group sample size for the curated assertion from the article.
	patient_number (reference)	PE	To capture the reference group sample size for the curated assertion.
	age (case)	PE	Used to capture the study sample age from the article.
	gender (case)	PE	Used to capture the gender for studied samples from the article.
	ethnicity (case)	EE	This represents the nationality/ethnicity of the study group as stated in the article.
Trial level information	trial_status	PE	Current stage of a clinical study. (e.g.) Completed, Terminated, etc.
	sponsor & collaborator	PE	Sponsors/collaborators of the clinical study.
	phase	PE	Represents the clinical phase of the trial. (e.g.) 0, I, II, III, IV.
	inclusion_criteria	PE	Description on the Inclusion criteria for the patients in the clinical study.
	exclusion_criteria	PE	Description on the Exclusion criteria for the patients in the clinical study.
Study design	allocation	PE	Assigning trial subjects to treatment or control groups. (e.g.) Non-randomized, Randomised.
	intervention_model	PE	Type of intervention model from the study. (e.g.) Single Group Design, Parallel Design, Crossover Design and Factorial Design.
	masking	PE	Types of Masking include None, Open Label, Single and Double Blind Masking.
	primary_purpose	PE	Represents purpose of the study primarily under taken for the research.
Additional details	pathway_name	EL+PE	Names of the pathways in which a biomarker has a role. Taken from KEGG database.
Reference details	Source	FV	Whether the curated data is from full-text or abstract of the article or ClinicalTrials.gov.
	Title	PE	Title of the article.
	Authors	PE	Authors of the article.
	Article/URL	PE	Name of the journal or specific links from which the information is captured.
	Year	PE	Year in which the given article published. (i.e.) Article published year for PubMed articles and first received year is considered for ClinicalTrials.gov.