

2018/11/22

JAMI & JSAI AIM 合同研究会

希少疾患診断支援システム PubCaseFinderを支える オントロジーとオープンデータ

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
藤原豊史 @fujitoyo

ライフサイエンス統合データベースセンターのミッション

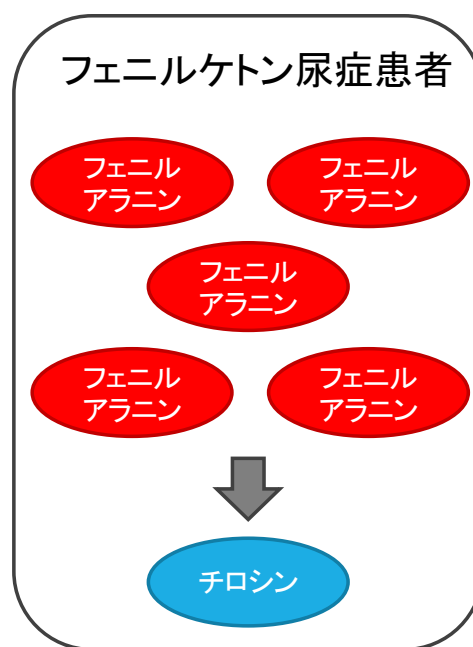
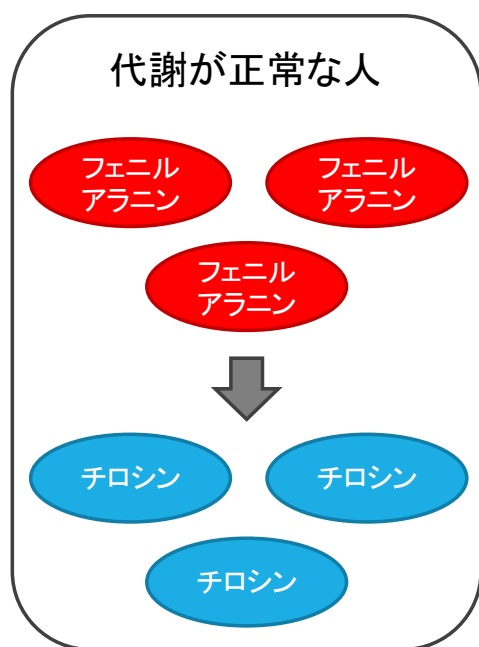
大学共同利用機関法人
情報・システム研究機構
データサイエンス共同利用基盤施設
ライフサイエンス統合データベースセンター

1. ライフサイエンスDBを有効利用し、知識発見に繋げるためのDB統合基盤技術の開発
2. ライフサイエンス知識を有効利用するための文献処理技術などの基盤技術開発とコンテンツの作成
3. ゲノムを含むオミックスデータを有効利用するための基盤整備

背景

□ 希少疾患患者に適切な治療を施すには早期の診断が重要

○ 先天性代謝異常症：フェニルケトン尿症



【症状】

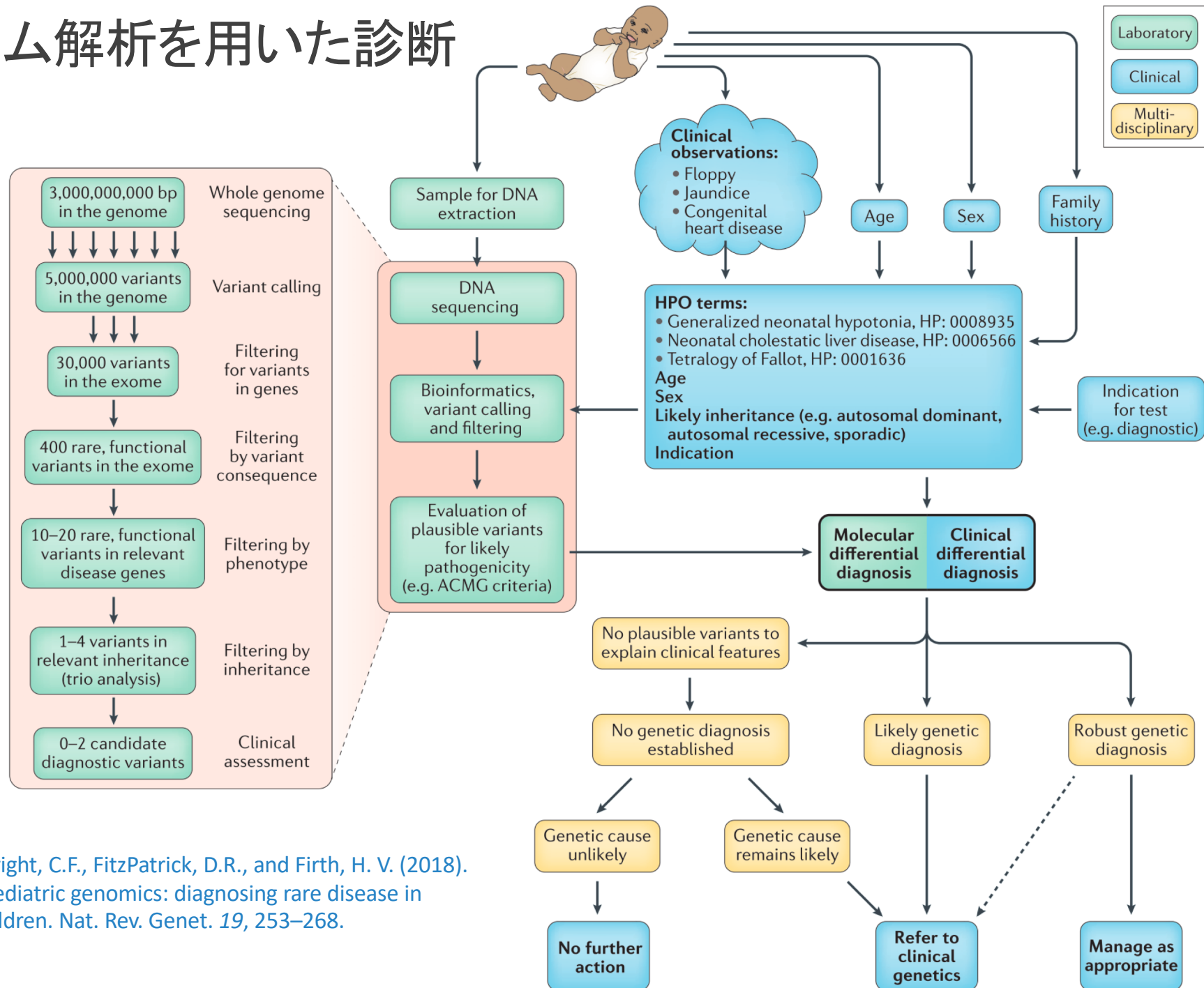
けいれん、発達障害など

食べ物から入る
フェニルアラニン
の量を減らす食事
療法を適用するこ
とで症状改善

○ 患者の約40%は最初の診断が間違っていた (EURORDIS, 2007)

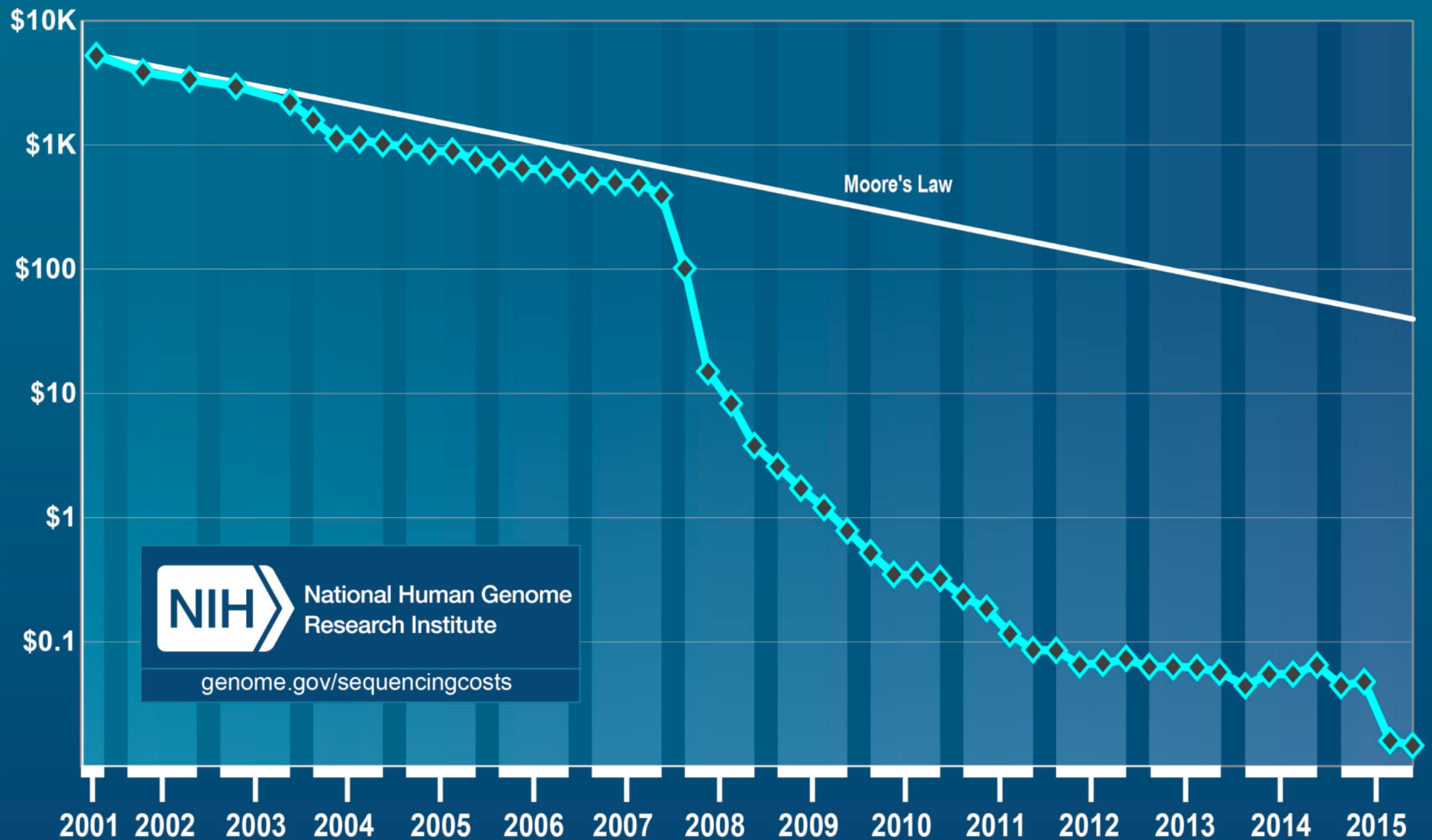
○ 患者の約半数は診断がつくまでに5年以上を要した (RD-UK, 2015)

ゲノム解析を用いた診断



出所: Wright, C.F., FitzPatrick, D.R., and Firth, H. V. (2018).
Paediatric genomics: diagnosing rare disease in children. Nat. Rev. Genet. 19, 253–268.

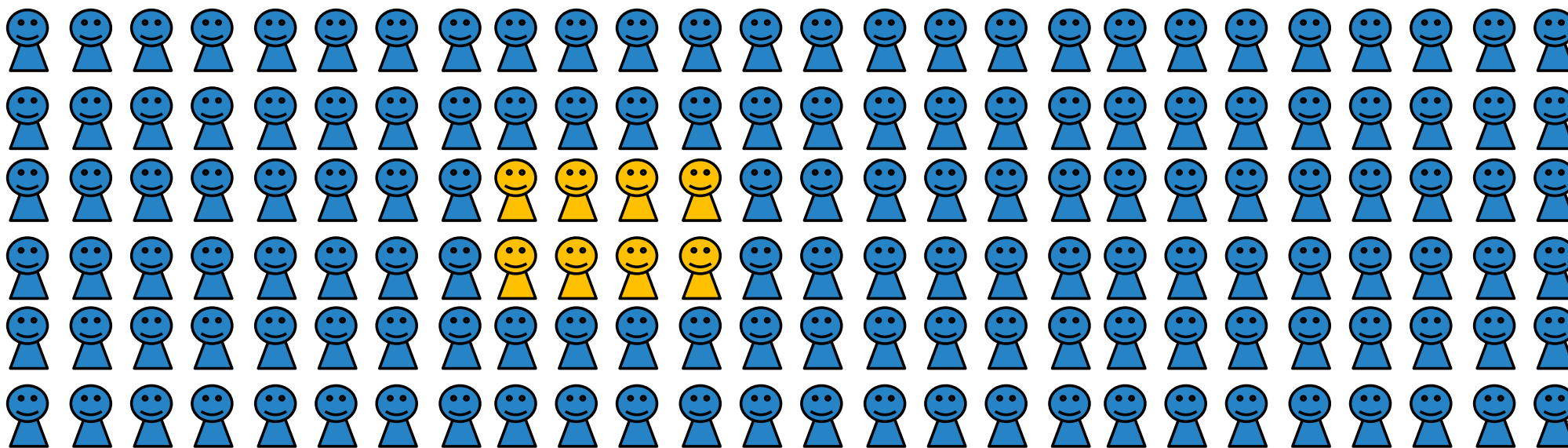
Cost per Raw Megabase of DNA Sequence



背景

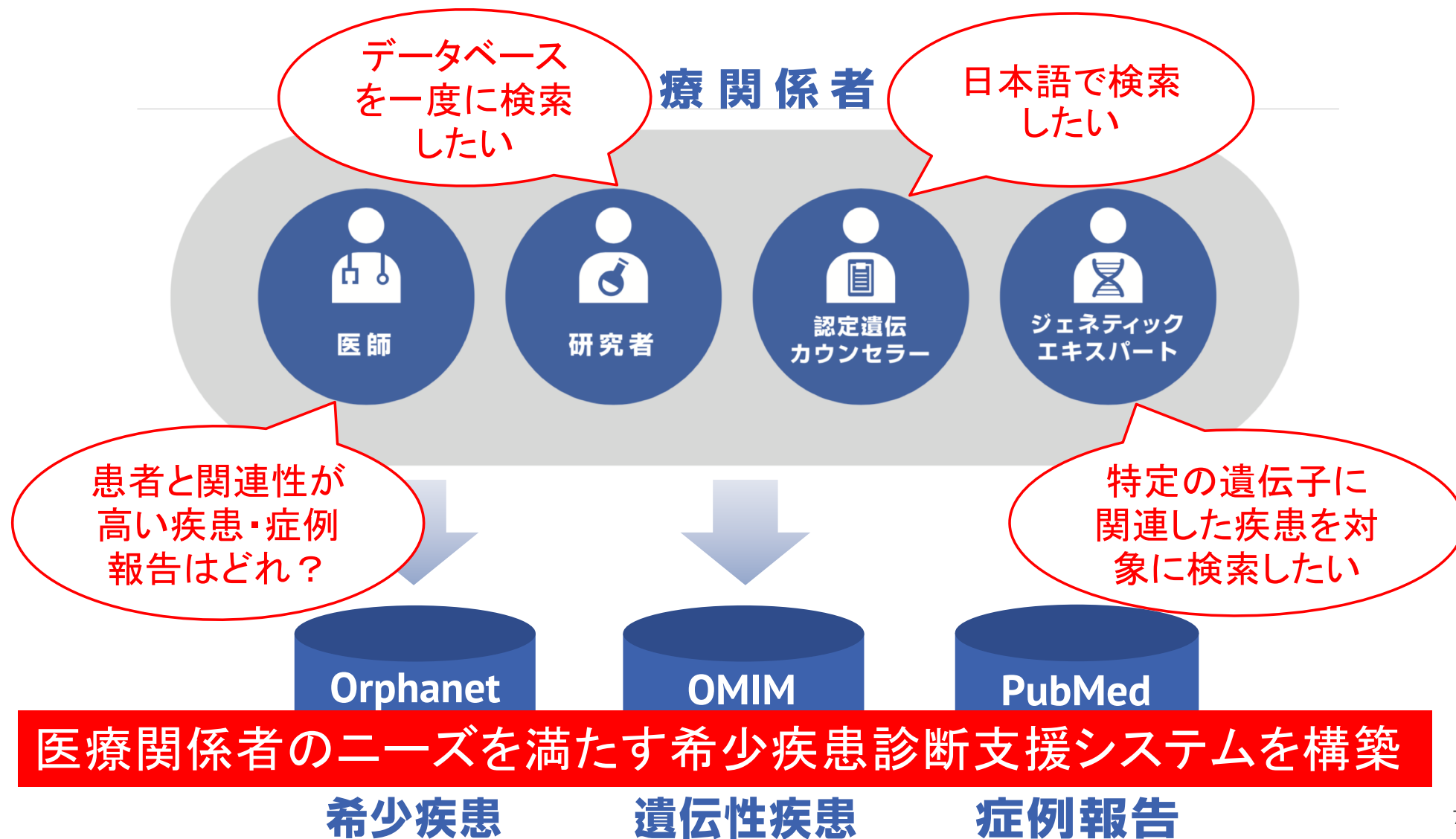
□ 希少疾患の未診断患者を対象に、次世代シーケンサーを用いたゲノム解析が実施されている (Sawyer, 2016)

○ 診断率は約25～40%



背景・目的

- 早期の診断には、医療関係者が**患者の症状**を元に、**疾患情報**や**過去の症例**を容易に検索できる環境整備が重要 (Sebastian, 2016)



希少疾患診断支援システム PubCaseFinder



クリア

疾患を検索

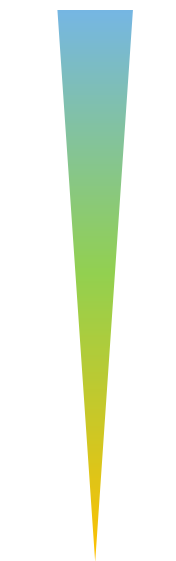
ex) 毛細血管拡張、知的障害、発作、舞蹈病、頭痛

PubCaseFinder を支える重要なリソース

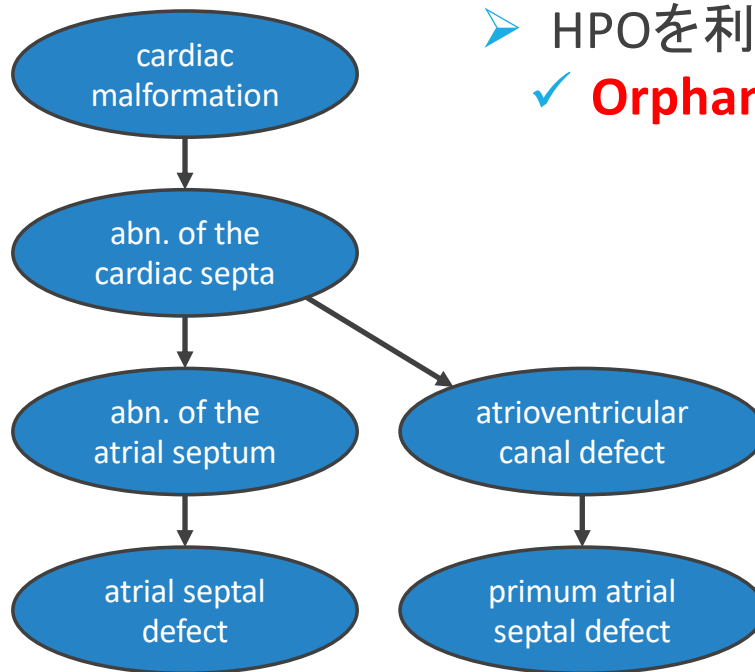
□ Human Phenotype Ontology (HPO)

- 希少・遺伝性疾患に関する症状を**約13,000件**収録 (Sebastian, 2016)

general



specific



- HPOを利用する多数のデータベース (Köhler, 2017)
- ✓ **Orphanet**, **OMIM**, ClinVar, MedGen, GARD... etc.

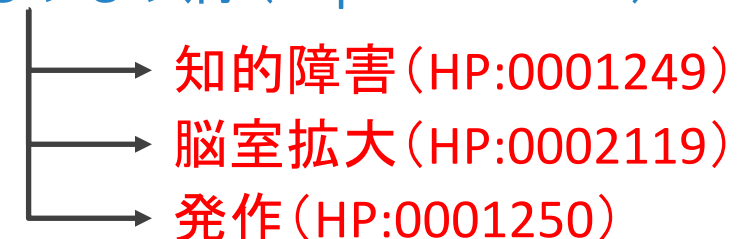


希少疾患



遺伝性疾患

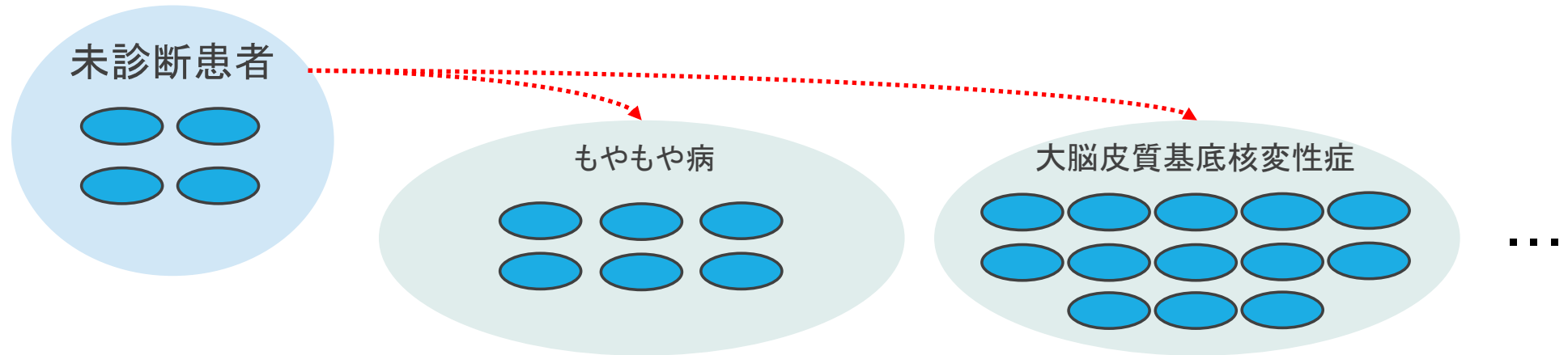
例) もやもや病 (Orphanet:2573)



代表表現: Atria septal defect (心房中隔欠損)
類義語: Atrial septum defect
類義語: ASD

症状セットの類似度計算手法

- 未診断患者と各疾患との類似度を症状セットを元に求める



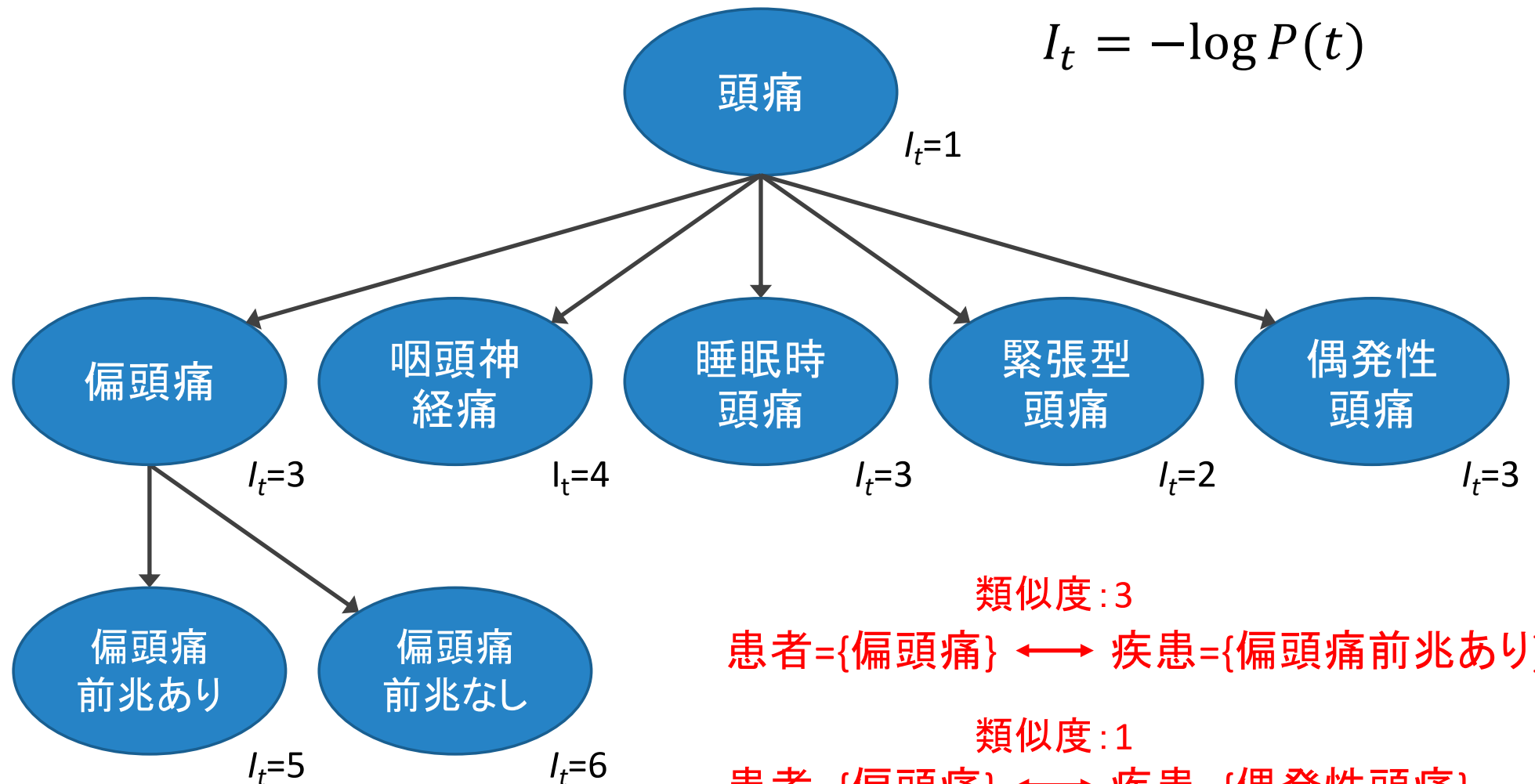
Measure	Equation	Variations	Reference
Resnik(a,b)	$\max_{t \in g^a \cap g^b} IC(t)$	Avg, Max	Rensik, 1995
Lin(a,b)	$\frac{2 * Resnik(a,b)}{IC(a) + IC(b)}$	Avg, Max	Lin, 1998
Jiang-Conrath(a,b)	$\frac{1}{IC(a) + IC(b) - 2 * Resnik(a,b) + 1}$	Avg, Max	Jiang, 1997
simGIC(P,Q)	$\frac{\sum_{t \in g^P \cap g^Q} IC(t)}{\sum_{t \in g^P \cup g^Q} IC(t)}$		Pesquita, 2007
GeneYenta(P,Q)	$\frac{\sum_{t \in T_c} R_t \times \max_{t' \in T_d} sim_{terms}(t, t')}{\sum_{t \in T_c} R_t \times I_t} \times 100$		Gottlieb, 2015

症状セットの類似度計算手法

□ 症状間の類似度

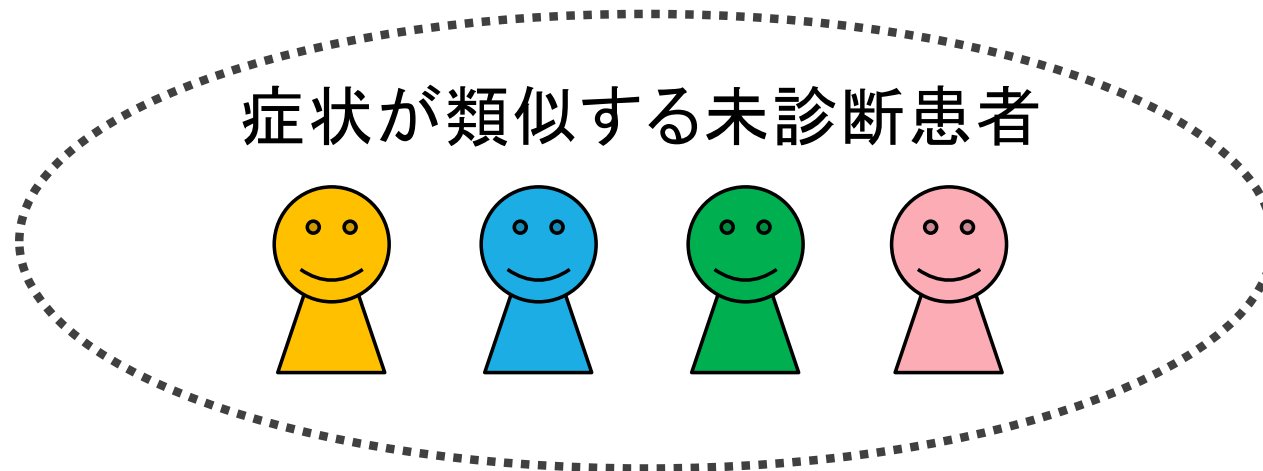
$$P(t) = \frac{|annot_t|}{|annot_{all}|}$$

$$I_t = -\log P(t)$$



HPOと類似度計算手法の活用事例

- 希少疾患分野では、HPOと類似度計算手法を用いて、症例データの共有が盛んに行われている
 - 症状が類似する複数の未診断患者が集まることで
 - 新規疾患の定義
 - 疾患原因遺伝子の同定



各未診断患者の症状はHPOで表現

HPOと類似度計算手法の活用事例

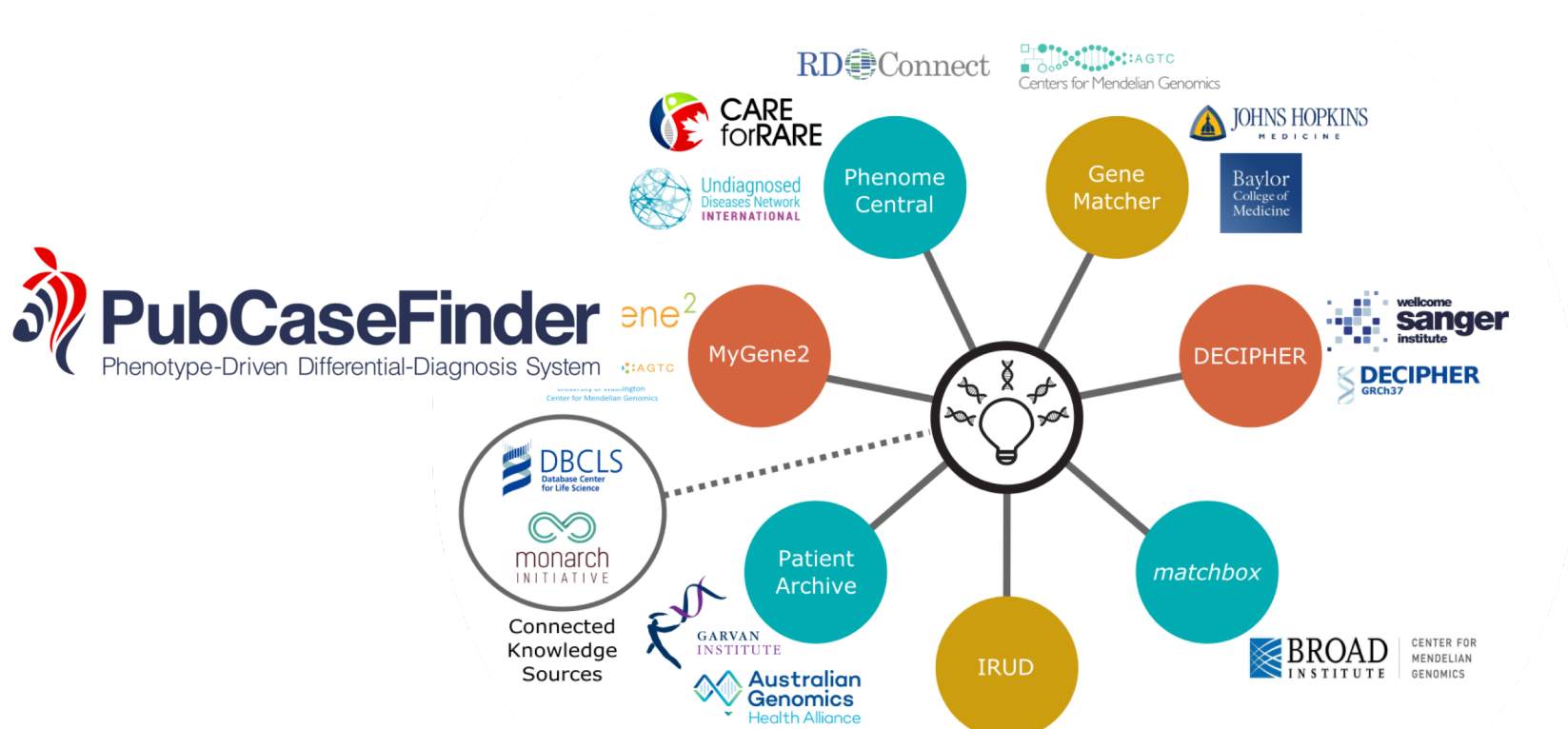
□ 現在、多くの症例データベースが症状をHPOで管理

Name	URL
PhenomeCentral	phenomecentral.org
DDD (Deciphering Developmental Disorders)	www.ddduk.org
DECIPHER (Database of genomic variation and Phenotype in Humans using Ensembl Resources)	decipher.sanger.ac.uk
ECARUCA (European Cytogeneticists Association Register of Unbalanced Chromosome Aberrations)	http://umcecaruca01.extern.umcn.nl:8080/ecaruca/ecaruca.jsp
The 100 000 Genomes Project	https://www.genomicsengland.co.uk/
Geno2MP (Exome sequencing data linked to phenotypic information from a wide variety of Mendelian gene discovery projects)	http://geno2mp.gs.washington.edu
NIH UDP (Undiagnosed Diseases Program)	available via phenomecentral.org
NIH UDN (Undiagnosed Diseases Network)	available via phenomecentral.org
HDG (Human Disease Gene Website series)	www.humandiseasesgenes.com
Phenopolis (An open platform for harmonization and analysis of sequencing and phenotype data)	https://phenopolis.github.io
GenomeConnect (Patient portal developed by ClinGen (67))	www.genomeconnect.org
FORGE Canada & Care4Rare Consortium	available via phenomecentral.org
RD-Connect	platform.rd-connect.eu
Genesis	thegenesisprojectfoundation.org

HPOと類似度計算手法の活用事例

❑ 希少疾患は症例数が少ないため、世界規模で症例データを共有する必要がある

○ 多くのプロジェクトが症例データ共有プロジェクト Matchmaker Exchangeに参加 (Orion, 2016)

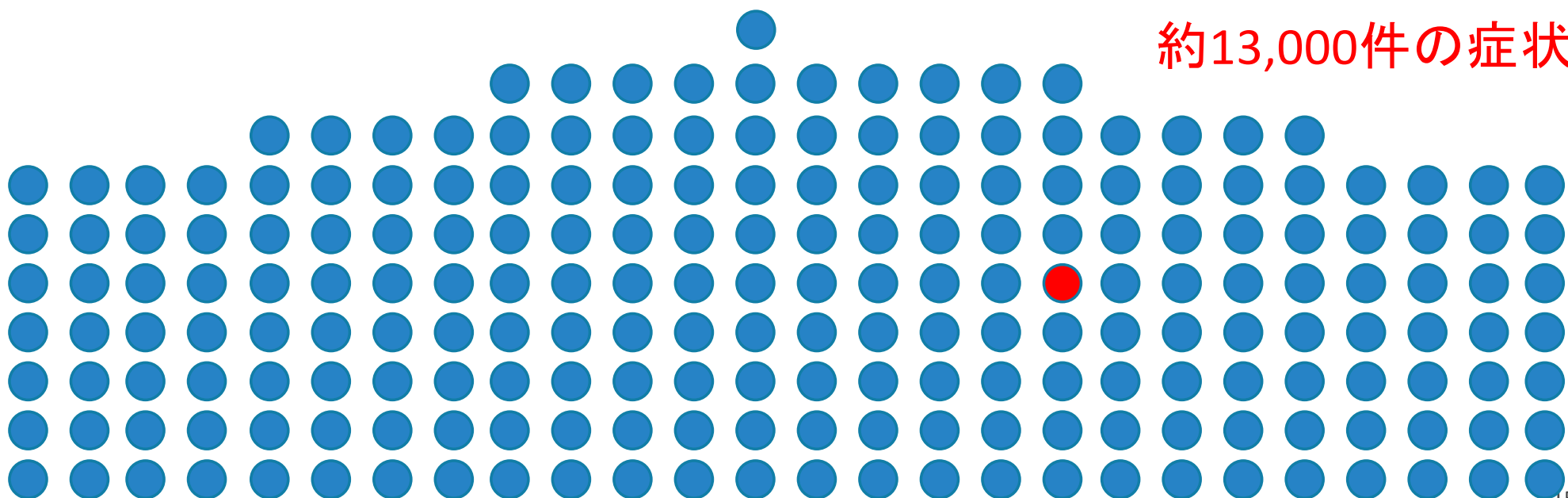


100以上の疾患原因遺伝子が同定された

問題点1: HPOの用語を探すのが困難



顔	?
HP:0010628 顔面麻痺	
HP:0007209 顔面麻痺	
HP:0010828 片側顔面スパスム	
HP:0000324 顔面非対称	
HP:0000282 顔面浮腫	
HP:0011331 片側顔面萎縮	



約13,000件の症状

HPOの用語を探すのが困難

□ HPO用語探索支援ツール「PhenoTouch」を開発

The screenshot displays the PubCaseFinder web application interface. At the top, the logo "PubCaseFinder" is visible, along with the tagline "Phenotype-Driven Differential-Diagnosis System". The page includes a navigation bar with links for "ヘルプ", "利用規約", "API", and "お問い合わせ". The main content area features a search bar labeled "患者の徴候および症状" with a language dropdown set to "JPN". Below the search bar, several HPO terms are listed: "HP:0001009 毛細血管拡張", "HP:0001249 知的障害", "HP:0001250 発作", "HP:0002072 舞蹈病", and "HP:0002315 頭痛". A "Clear" button is present next to the search bar. Below the search bar, there are three panels: "上位概念" (Superior Concept) showing "神経系生理の異常" with a count of 26 and buttons for "追加" and "置換"; "HP:0002315 頭痛" (Selected HPO Term) showing details for the term, including its ID, symptoms in Japanese and English, a definition, and a synonym; and "下位概念" (Subordinate Concept) showing a list of related terms with counts and buttons for "追加" and "置換".

PubCaseFinder
Phenotype-Driven Differential-Diagnosis System

(最終更新日: 2018年9月25日) ヘルプ 利用規約 API お問い合わせ

患者の徴候および症状 JPN ▼

HP:0001009 毛細血管拡張 × HP:0001249 知的障害 × HP:0001250 発作 × HP:0002072 舞蹈病 × HP:0002315 頭痛 ×

Clear OK Cancel

上位概念

神経系生理の異常 26 追加 置換

HP:0001009 毛細血管拡張 ×

HP:0002315 頭痛 ×

HP:0002315 頭痛

追加 置換

HPO Id : HP:0002315

症状(日) : 頭痛

症状(英) : Headache

症状定義 : Cephalgia, or pain sensed in various parts of the head, not confined to the area of distribution of any nerve.

Synonym : Headaches

下位概念

追加 置換 0 Thunderclap headache

追加 置換 2 偏頭痛

追加 置換 0 喉頭神経痛

追加 置換 0 睡眠時頭痛

追加 置換 0 緊張型頭痛

追加 置換 0 群発(性)頭痛

追加 置換 0 頭痛 (褐色細胞腫を伴う)

HPOの用語を探すのが困難

- ヒトの3Dモデルをタッチすると、該当箇所に関連する症状の一覧を得ることができる

tokyogw.bits.cc

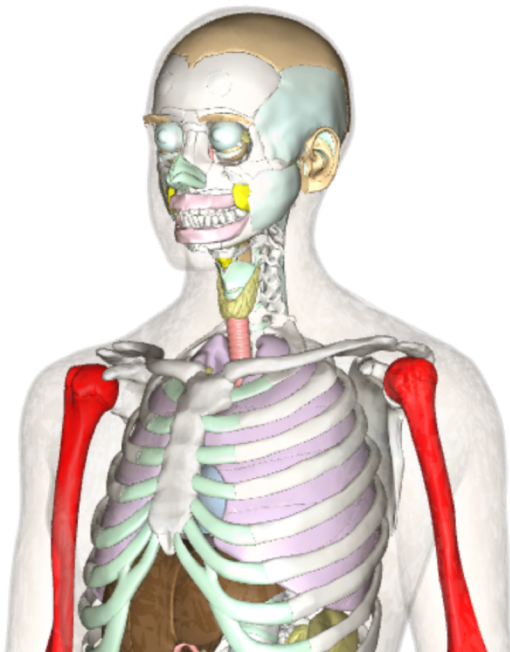
患者の徴候および症状 JPN ▼

HP:0001090 大きな眼 × HP:0009938 落ちくぼんだ頬部 × HP:0002057 目立つ眉間 × HP:000885 幅広い肋骨 ×

Clear OK Cancel

身体各部位を選択 Close

☒ Internal ☒ Bone ☐ Muscle



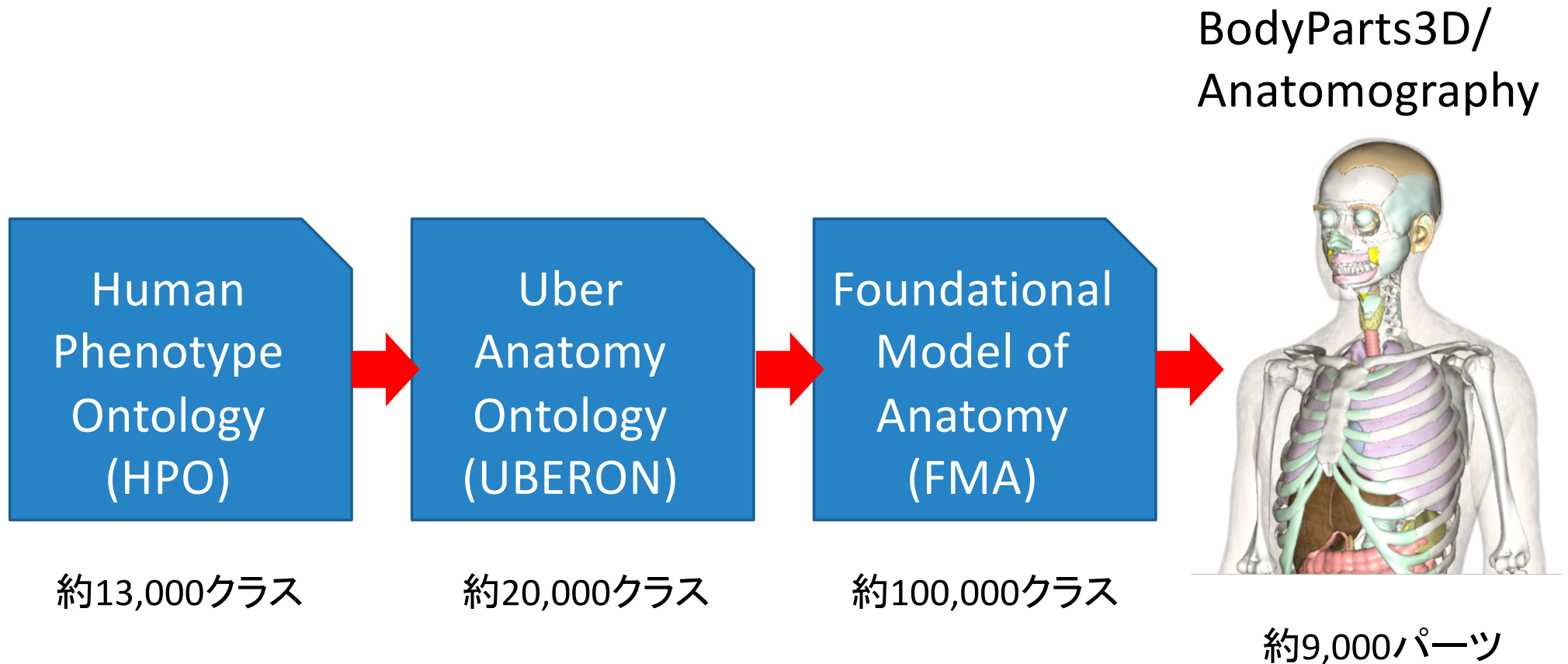
11 長骨
0 上腕骨

FMA:13303 上腕骨

追加	置換	0	三角形の上腕骨
追加	置換	11	上腕骨の異常
追加	置換	3	上腕骨無形成/低形成
追加	置換	3	上腕骨骨化異常
追加	置換	0	上腕骨骨粗鬆症
追加	置換	0	二分した上腕骨
追加	置換	0	先細りの上腕骨
追加	置換	1	変形した上腕骨
追加	置換	0	幅広い上腕骨

HPOの用語を探すのが困難

- オントロジーのクロスリファレンスを利用して、HPO termをヒトのBody Partsに対応付けた



問題点2: 疾患一症状関連情報の不足

- Orphanetの約7,000の希少疾患のうち、**症状が割り当てられているのは約2,500疾患のみ**

【順位】	【類似度】	【疾患】	【症状セット】			
1	0.9	ファブリー病	下痢	発疹	嘔吐	...
2	0.8	アジソン病	脱力感	筋力低下	易疲労感	...
3	0.6	遺伝性膵炎	腹痛	脂肪便	体重減少	...
4	0.5	ウィルソン病	黄疸	浮腫	食欲不振	...
⋮	⋮	⋮	⋮	⋮	⋮	⋮
2478	0.1	カナバン病	痙攣	症状が少ない		

Orphanet
の約60%

Orphanet

出所: Orphadata: Free access data
from Orphanet. © INSERM 1997.

疾患一症状
関連情報

← 不足

ランキング
対象外

ペータース異常

症状がない

文献から疾患に関連する症状を抽出

□ 日々大量に出版される文献から手動で情報を抽出するのは限界がある

○ 自動抽出を試みる



3000報以上/日

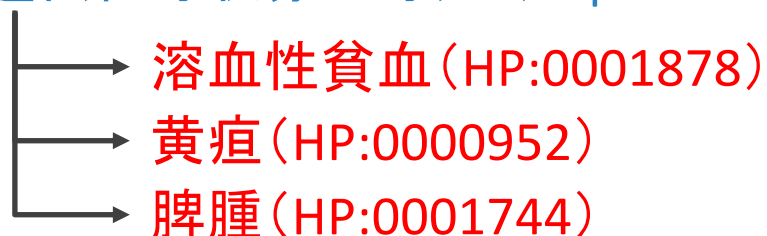
文献から疾患に関連する症状を抽出

□ 課題:「疾患－症状」関連データの不足が大きな課題

➤ 約100万件の症例報告から、テキストマイニング技術で自動取得



例) 遺伝性球状赤血球症 (Orphanet:822)



症例報告 PMID: 12355853

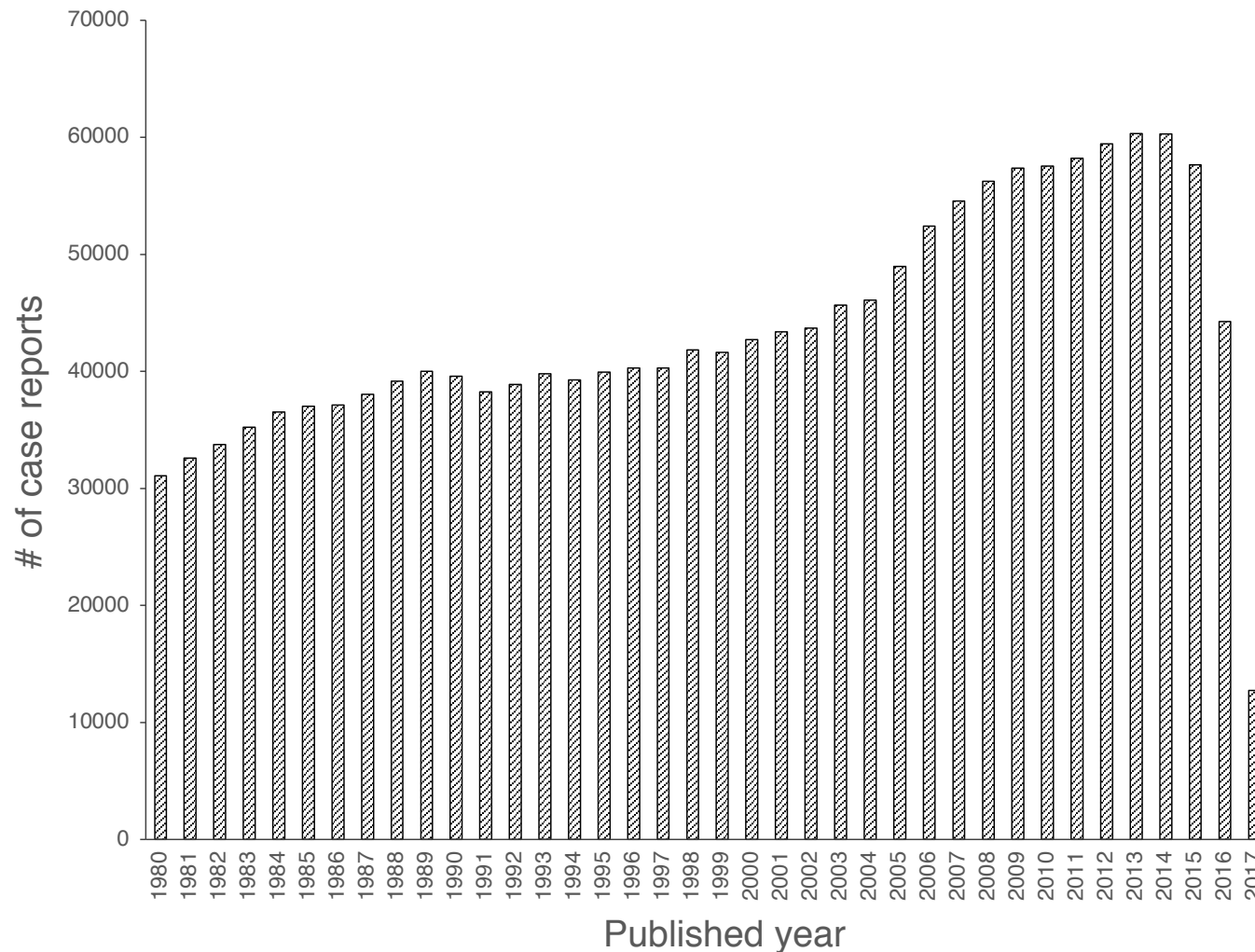
青: 疾患名 赤: 症状

Hereditary spherocytosis is a genetic, frequently familial hemolytic blood disease characterized by varying degrees of hemolytic anemia, splenomegaly, and jaundice. . . .



症例報告出版数

- Distribution of the number of case reports published per year in PubMed from 1980 to 2017



文献から疾患に関連する症状を抽出

□ オントロジーを用いたアノテーションツール

- ConceptMapper (Tanenblatt, 2010)、MetaMap (Aronson, 2010)、NCBO Annotator (Jonquet, 2009)

□ CRAFT Corpus を利用した、パフォーマンス比較 (Christopher, 2014)

- 8つのオントロジーにおいてF-measureを比較結果、7つのオントロジーでConceptMapperのF-measureが最も高かった

□ HPO gold standard (Tudor, 2015)を用いたツール評価 (藤原, JSAI2017)

System	F-measure	Precision	Recall
NCBO Annotator	0.51	0.54	0.47
MetaMap	0.56	0.51	0.61
ConceptMapper	0.52	0.52	0.51

System	Processing time (sec)
NCBO Annotator	206.0
MetaMap	351.0
ConceptMapper	4.3

100万件の症例報告
の処理に要する時間



17.7 (day)

5.2 (hour)

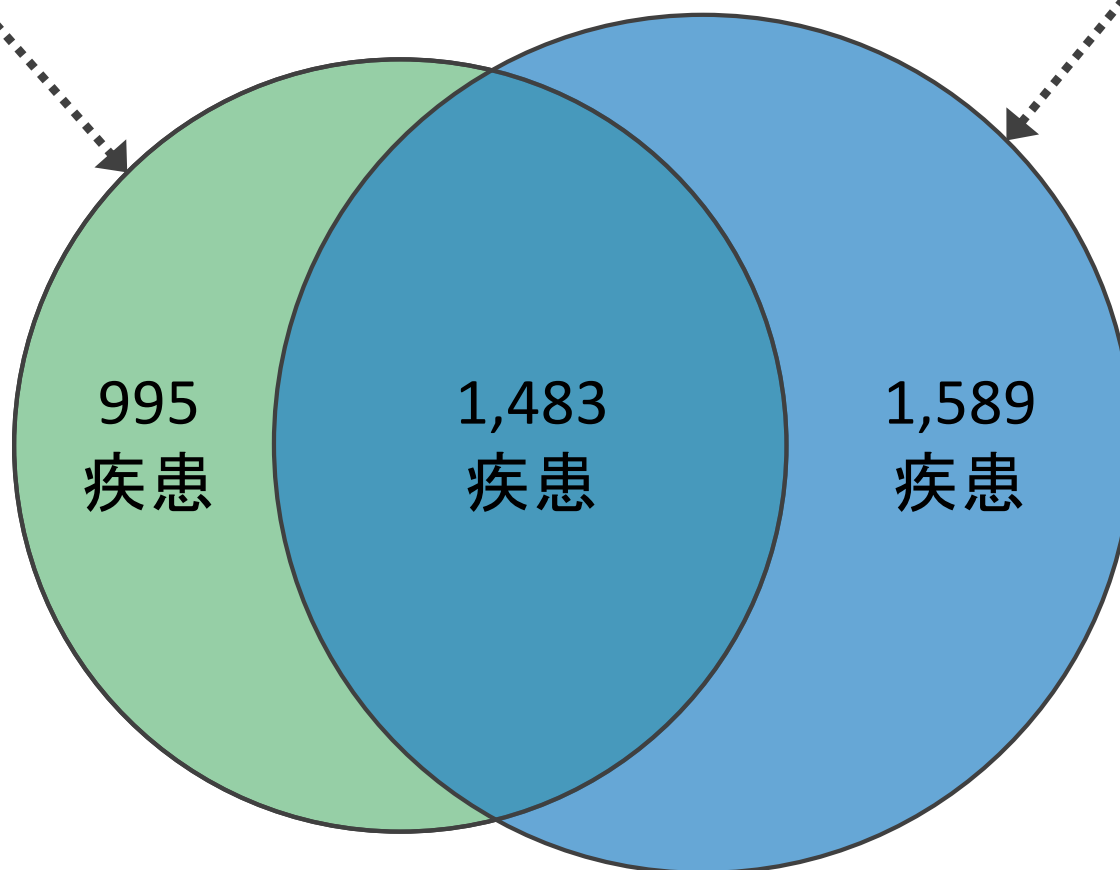
疾患一症状関連情報抽出 結果

【Orphanet】

2,478疾患に症状を付与

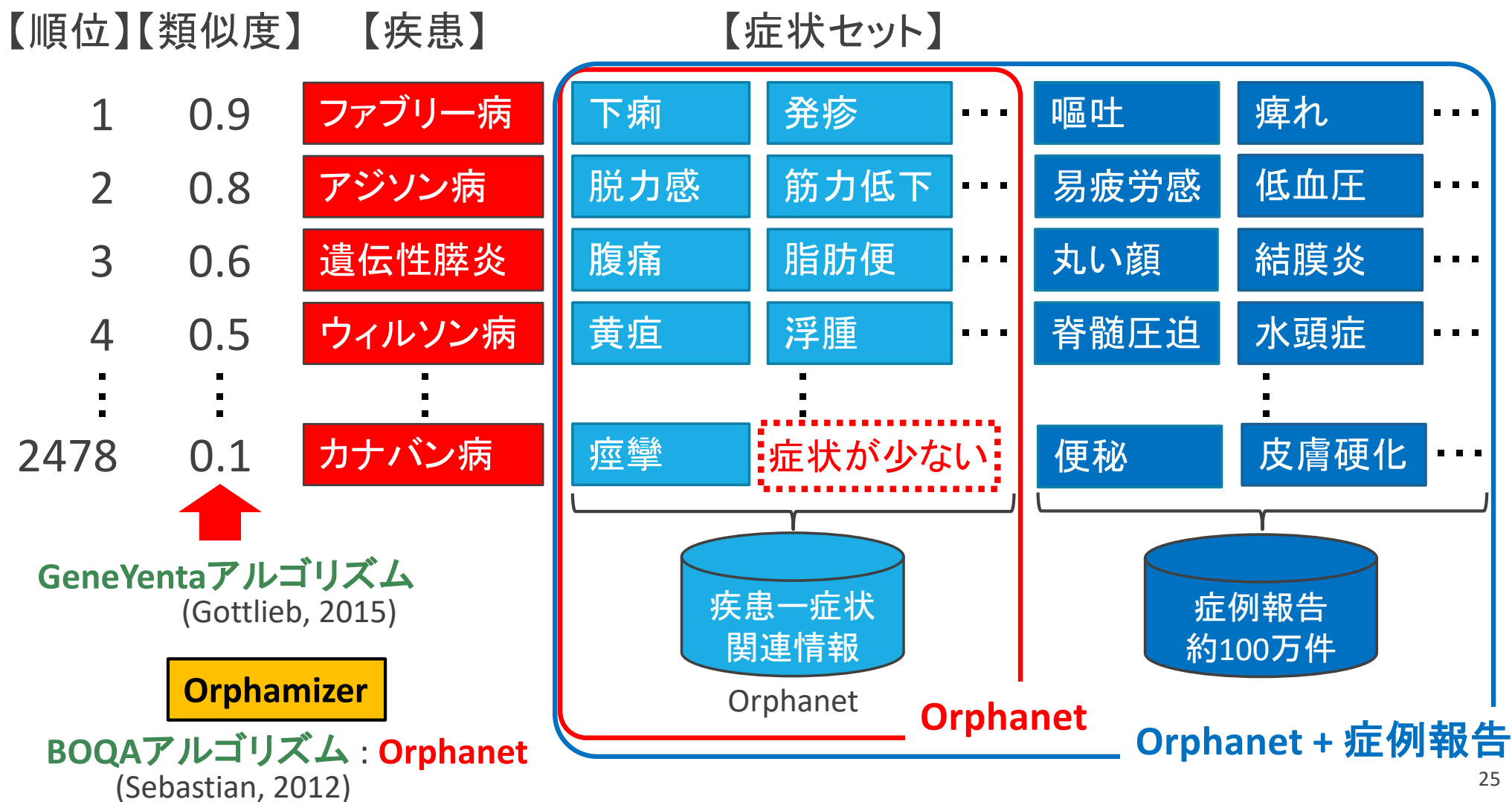
【症例報告】

3,072疾患に症状を付与



取得した疾患一症状関連情報を疾患ランキングシステムに活用

□ 症例報告から取得した疾患一症状関連情報は、ランキング精度の向上に寄与するか？



疾患ランキング精度の評価方法

□ 希少疾患の症例データセット(194件)で評価に利用

○ 診断結果(ORDO ID)、症状(HPO IDs)

□ Recallで疾患ランキング精度を評価

	症例 1	症例 2	症例 3
	順位 疾患名	順位 疾患名	順位 疾患名
	1 疾患A	1 疾患A	1 疾患A
診断結果 と一致	2 疾患B	2 疾患B	2 疾患B
	3 疾患C	3 疾患C	3 疾患C
	4 疾患D	4 疾患D	4 疾患D
Top5	5 疾患E	5 疾患E	5 疾患E
	6 疾患F	6 疾患F	6 疾患F
	⋮	⋮	⋮
	2478 疾患Z	2478 疾患Z	2478 疾患Z

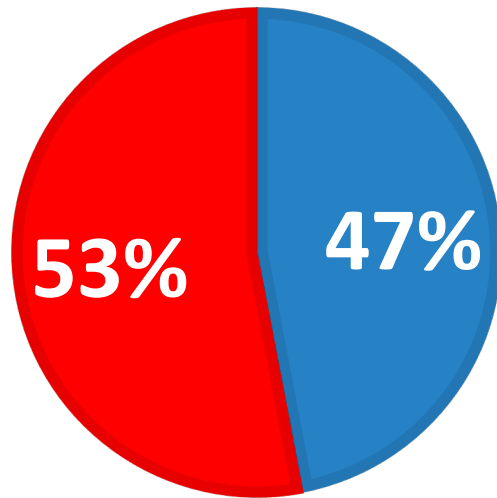
$$RecallX = \frac{TopX \text{ にランキングした症例数}}{\text{全症例数}}$$

疾患ランキング精度の評価

□ 評価用の135症例を利用して、トップ5のRecallを比較

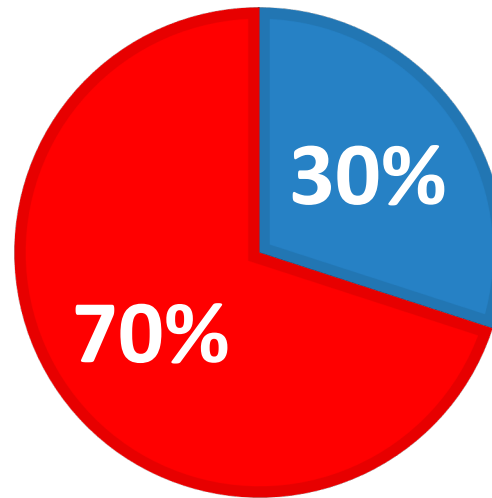
PubCaseFidner
(Orphanet + 症例報告)

■ 正解症例 ■ 不正解症例



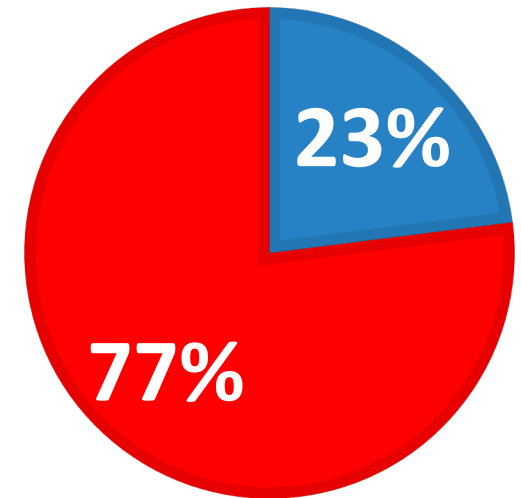
PubCaseFidner
(Orphanet)

■ 正解症例 ■ 不正解症例



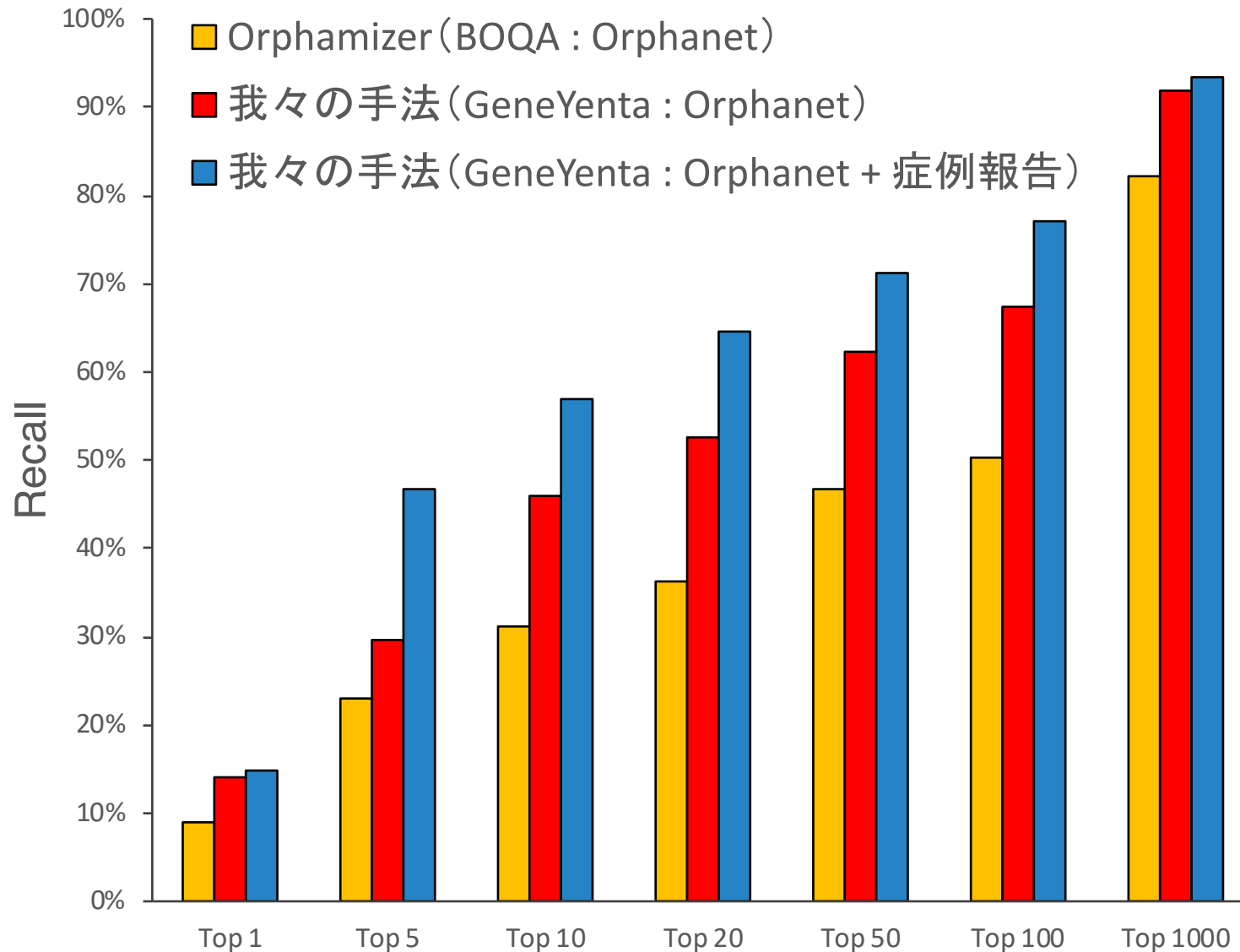
Orphamizer
(Orphanet)

■ 正解症例 ■ 不正解症例



疾患ランキング精度の評価

□ 評価用の135症例を利用して、Recall1,5,10,20,50,100,1000を比較



まとめ

医療関係者

複数のデータベースを一度に検索

日本語での検索可



医師



研究者



認定遺伝
カウンセラー



ジェネティック
エキスパート

高精度な疾患
ランキング機能
を提供

特定の遺伝子
に関連した疾患
を対象に検索可



PubCaseFinder

Phenotype-Driven Differential-Diagnosis System

Orphanet

約7000件

希少疾患

OMIM

約8000件

遺伝性疾患

PubMed

約100万件

症例報告



PhenoTouch

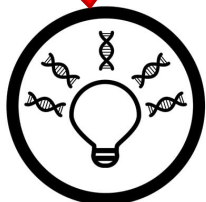
症状入力
支援機能



PubCaseSummary

検索結果
まとめ機能

GA4GHへの
貢献



Matchmaker
Exchange

API

謝辞

DBCCLS

金進東
山本泰智
片山俊明

東京大学

高木利久

東北大学

荻島創一

慶應義塾大学

小崎健次郎

国立遺伝学研究所

大久保公策
川本祥子

Gene42

Orion Buske

Care4Rare Canada Consortium

日立製作所

熊谷禎洋

BITS

武藤勇
山本利也

Genome.One

Tudor Groza

ご清聴ありがとうございました