Research Organization of Information and Systems
Joint Support-Center for Data Science Research
Database Center for Life Science

DBCLS
Database Center
for Life Science

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

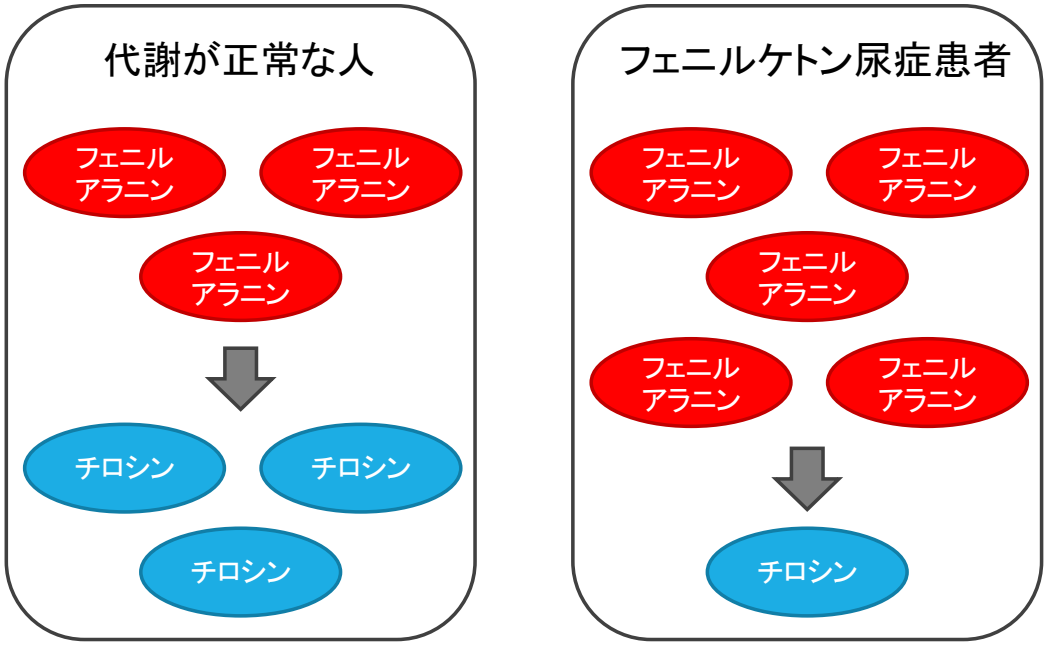
ライフサイエンス統合データベースセンターのミッション

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

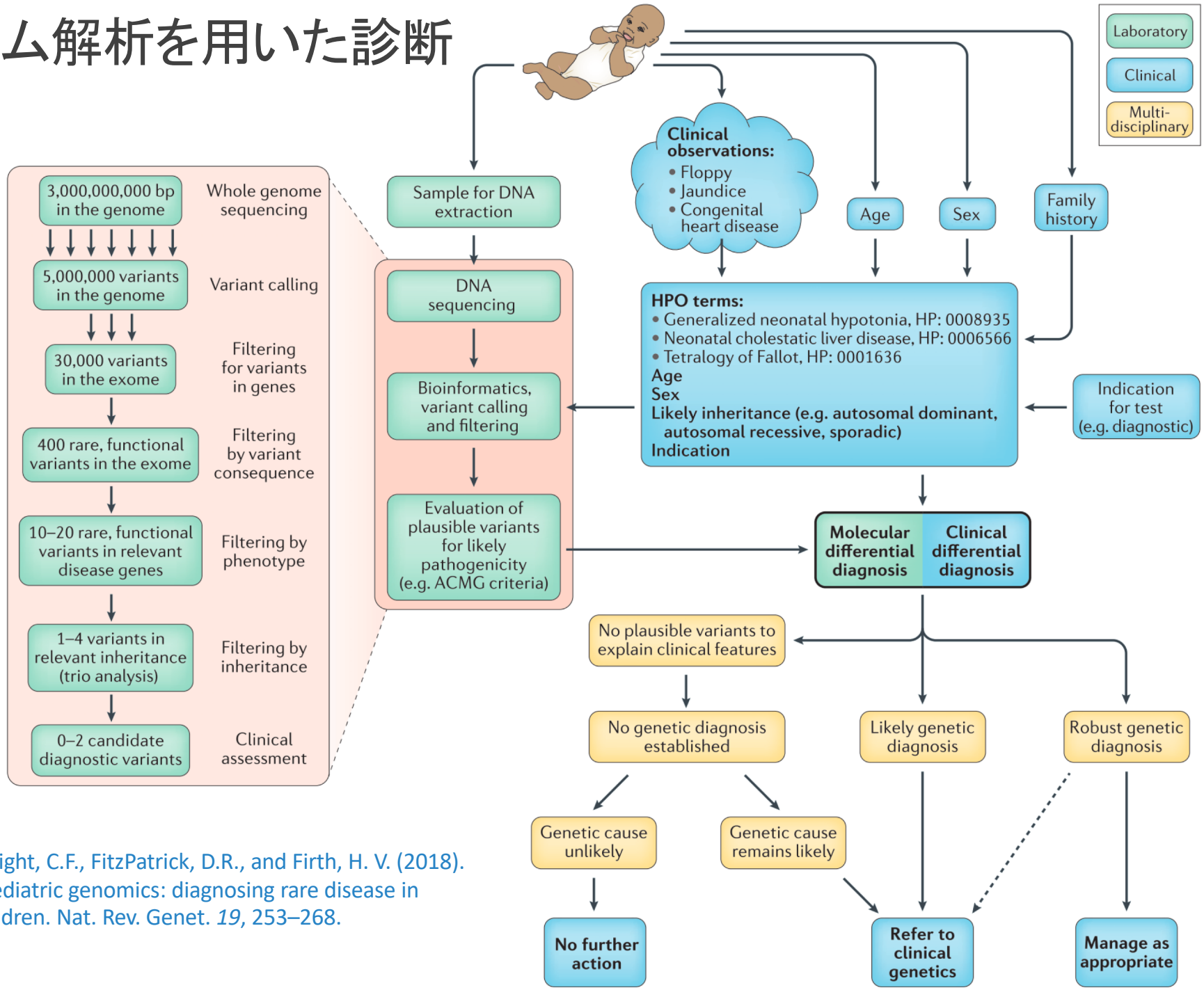


【症状】
けいれん、発達障害など

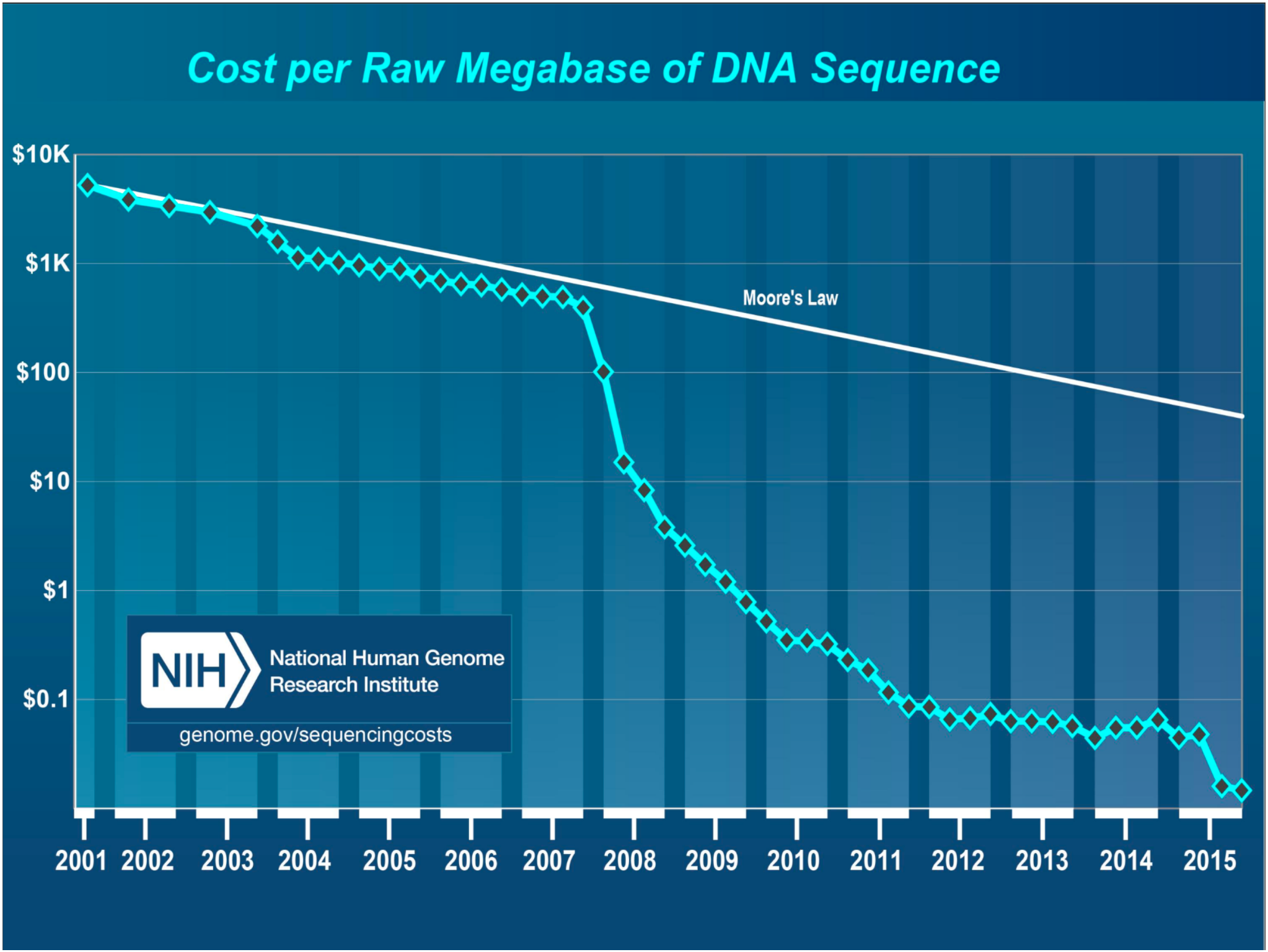
食べ物から入るフェニルアラニンの量を減らす食事療法を適用することで症状改善

  - 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

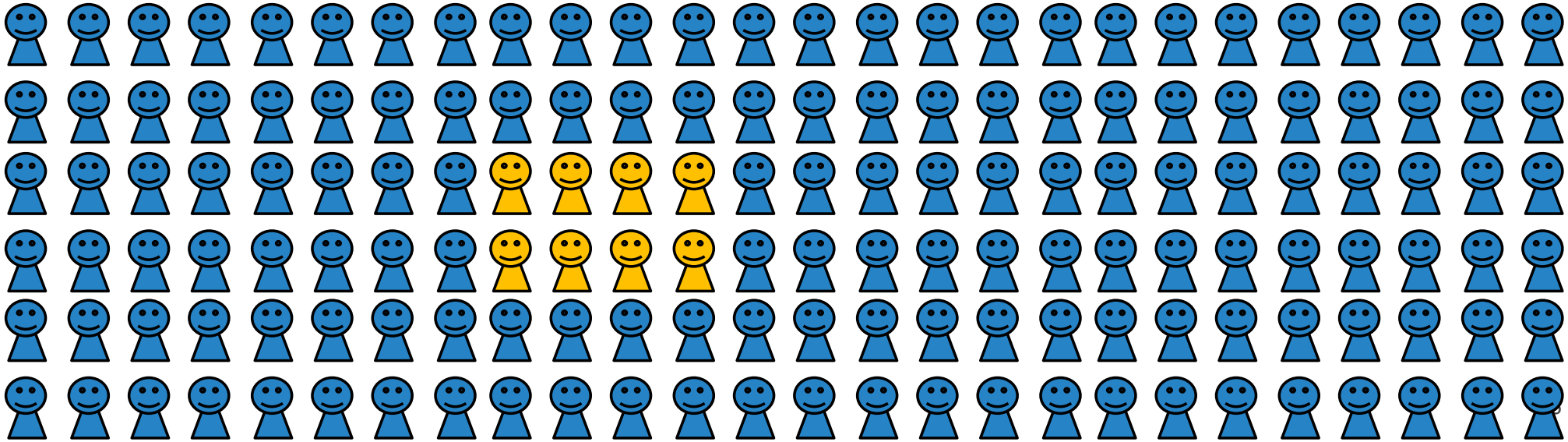


出所：Wright, C.F., FitzPatrick, D.R., and Firth, H. V. (2018). Paediatric genomics: diagnosing rare disease in children. Nat. Rev. Genet. *19*, 253–268.


Cost per Raw Megabase of DNA Sequence
$10K
$1K
$100
$10
$1
$0.1
Moore's Law
NIH
National Human Genome Research Institute
genome.gov/sequencingcosts
2001
2002
2003
2004
2005
2006
2007
2008
2009
2010
2011
2012
2013
2014
2015

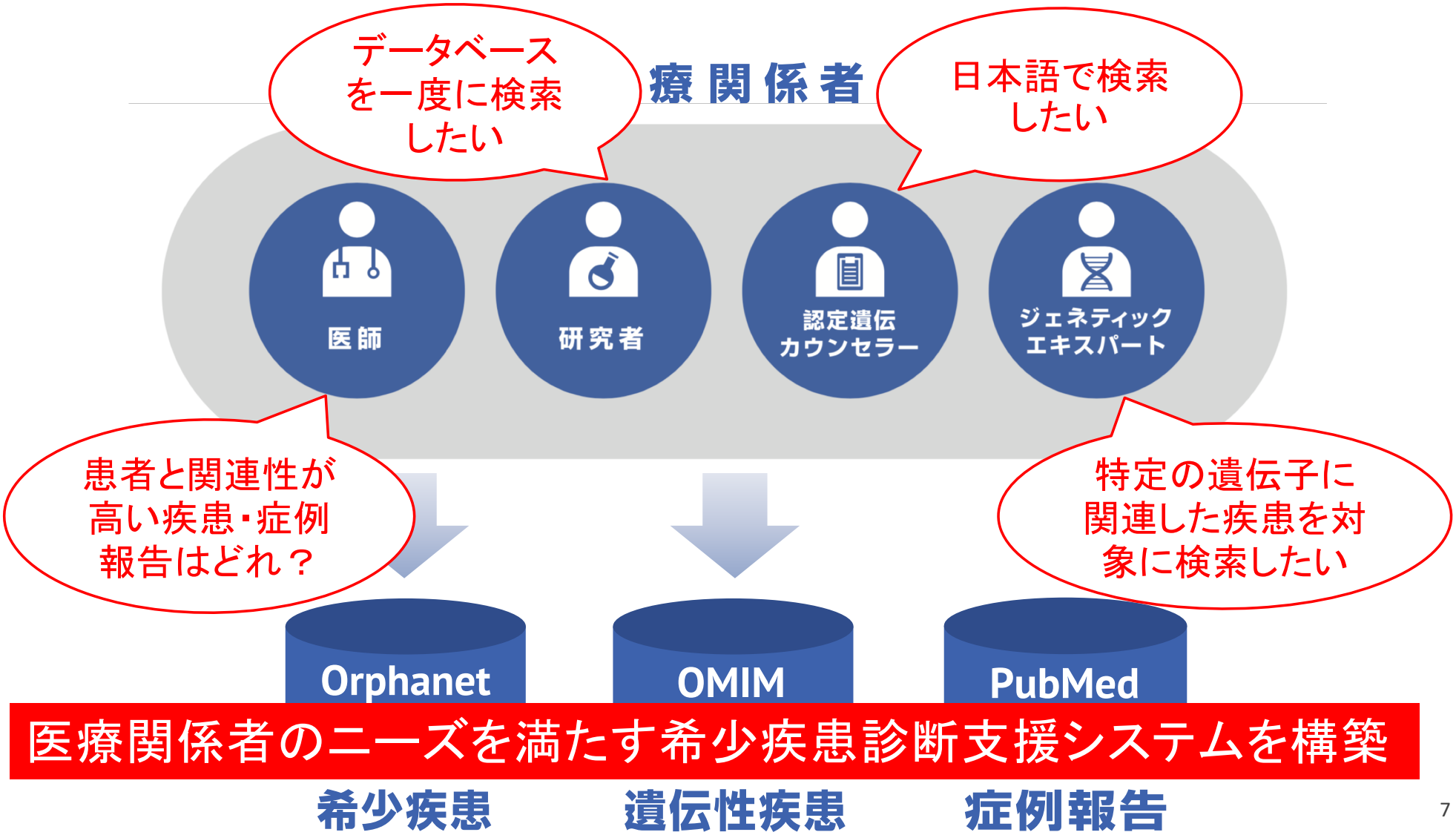
背景

- 希少疾患の未診断患者を対象に、次世代シークエ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%

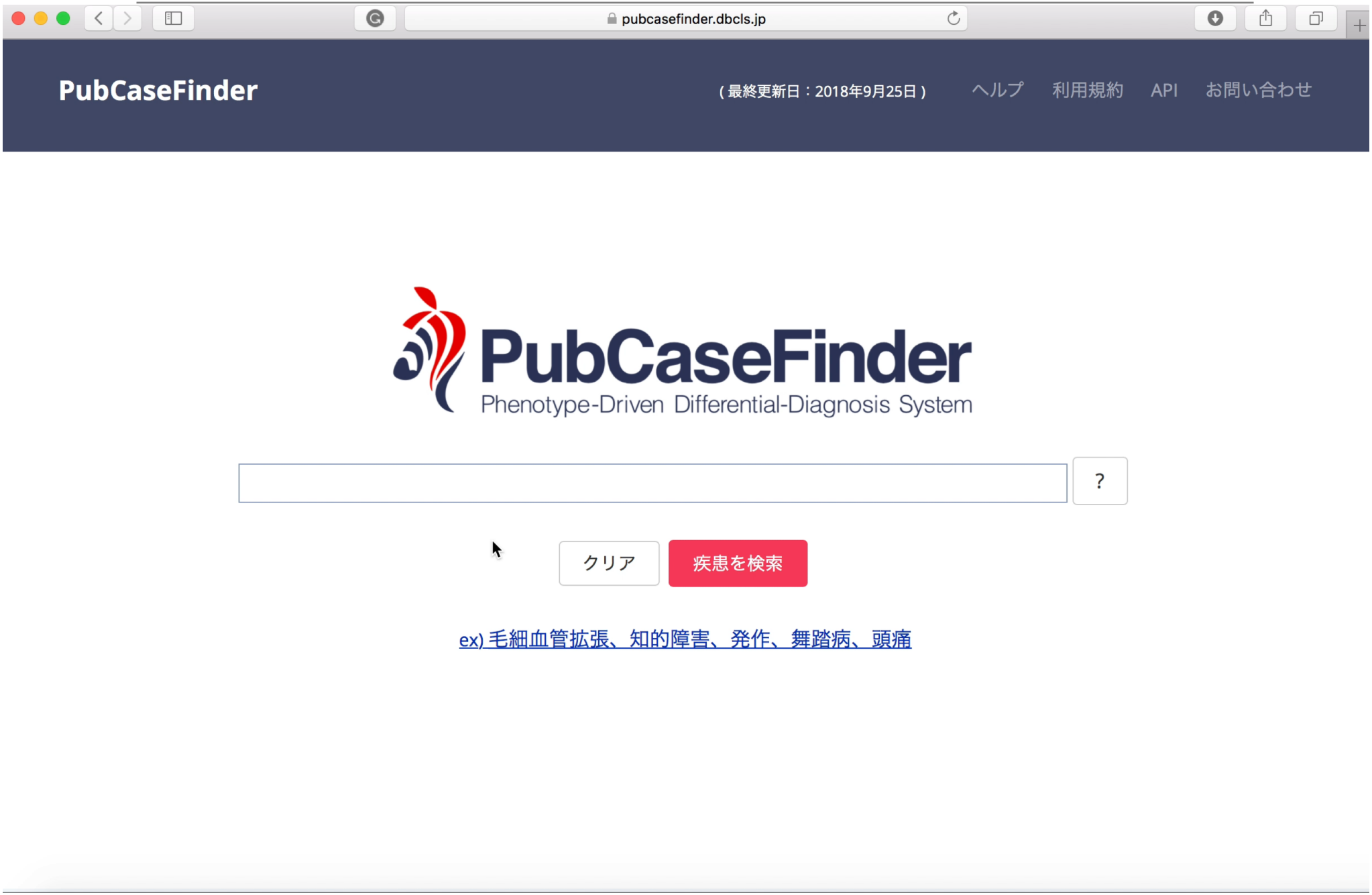


背景・目的

- 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）

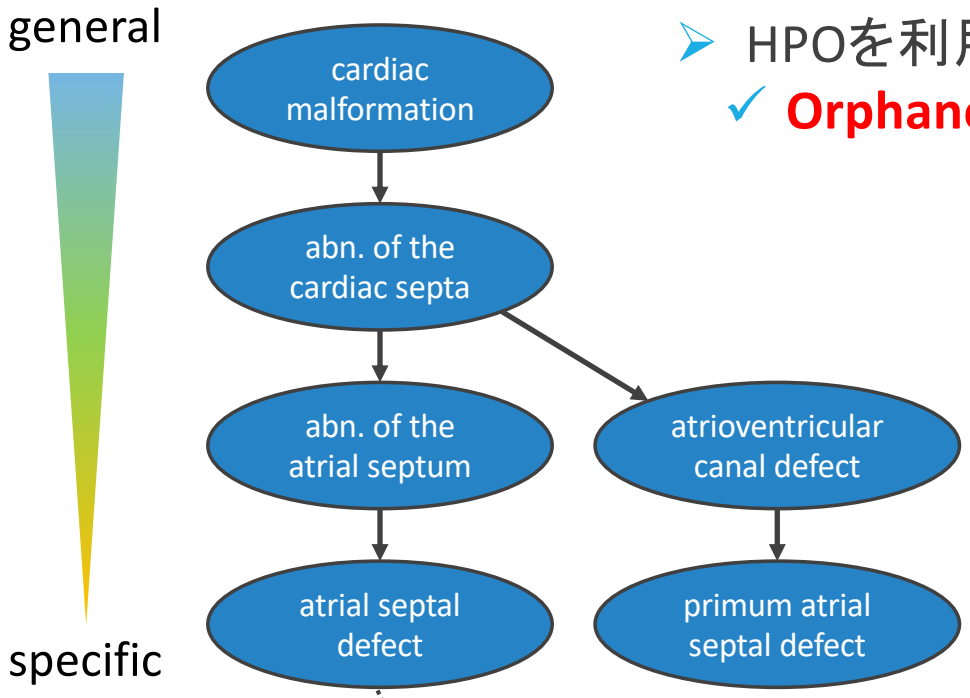


希少疾患診断支援システム PubCaseFinder



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）



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

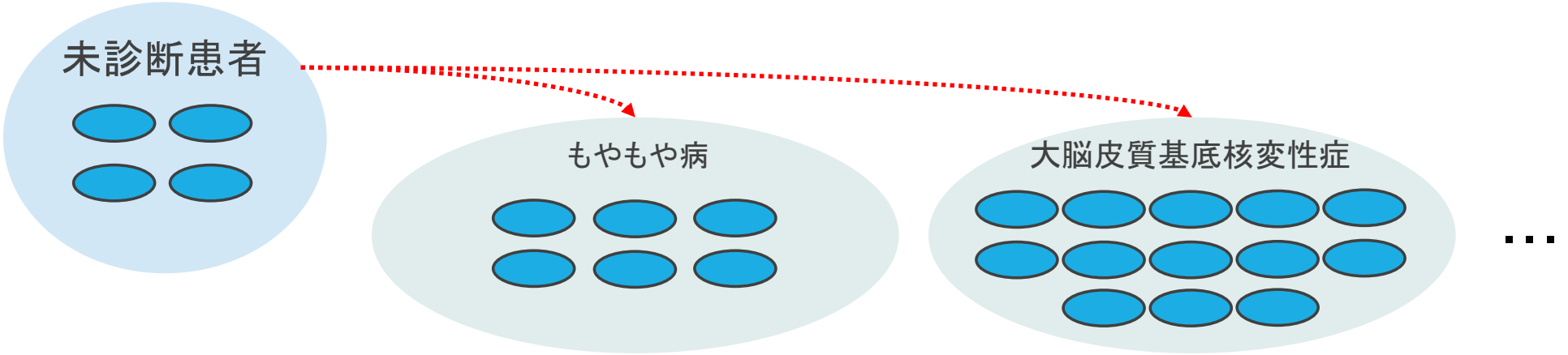
例）もやもや病（Orphanet:2573）

- 知的障害（HP:0001249）
- 脳室拡大（HP:0002119）
- 発作（HP:0001250）

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



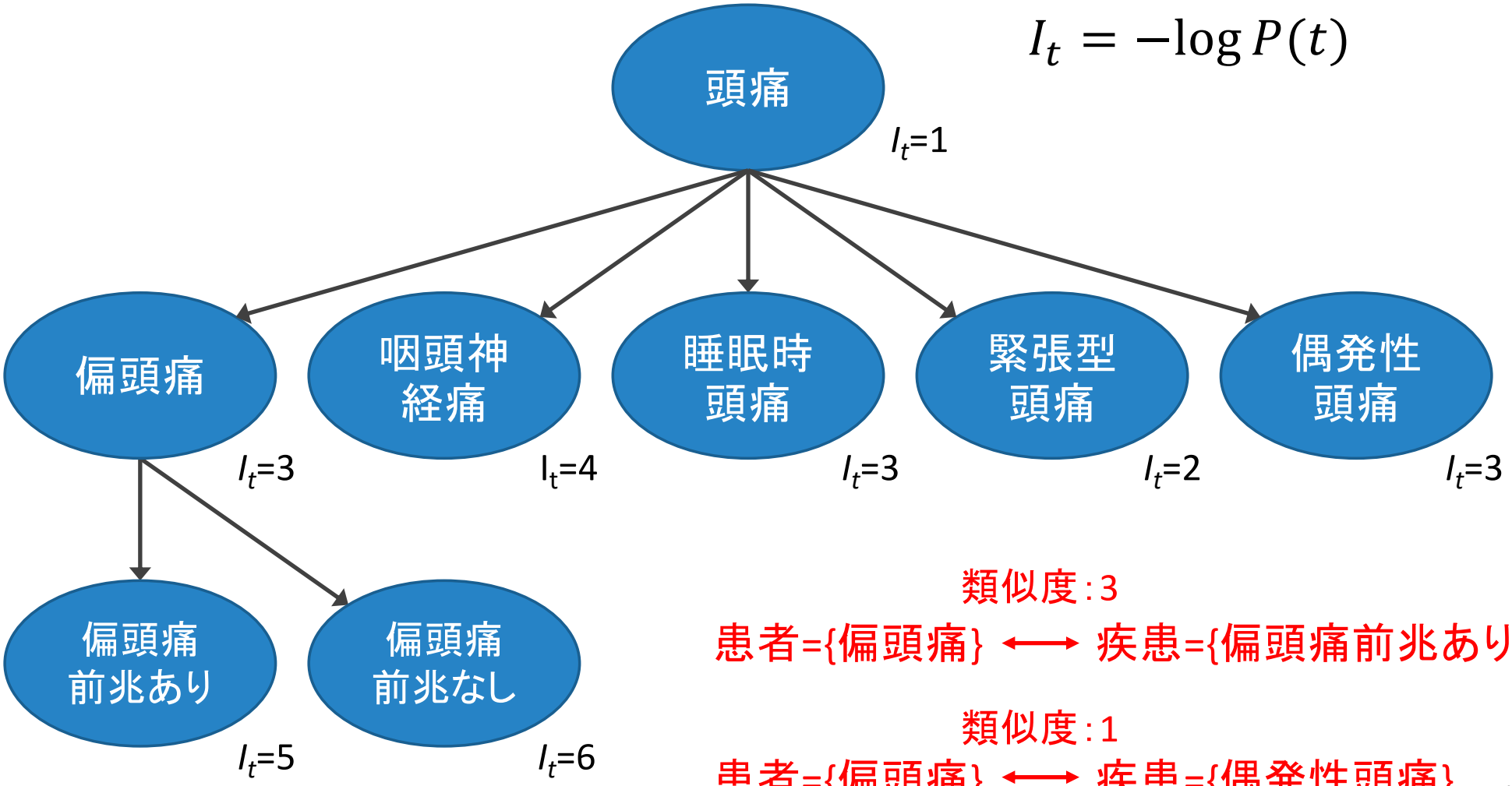
Measure	Equation	Variations	Reference
Resnik(a,b)	$\max_{t \in g^a \cap g^b} IC(t)$	Avg, Max	Rensik, 1995
Lin(a,b)	$\frac{2 * Resnik(a,b)}{IC(a) + IC(b)}$	Avg, Max	Lin, 1998
Jiang-Conrath(a,b)	$\frac{1}{IC(a) + IC(b) - 2 * Resnik(a,b) + 1}$	Avg, Max	Jiang, 1997
simGIC(P,Q)	$\frac{\sum_{t \in g^P \cap g^Q} IC(t)}{\sum_{t \in g^P \cup g^Q} IC(t)}$		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

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

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



類似度：3

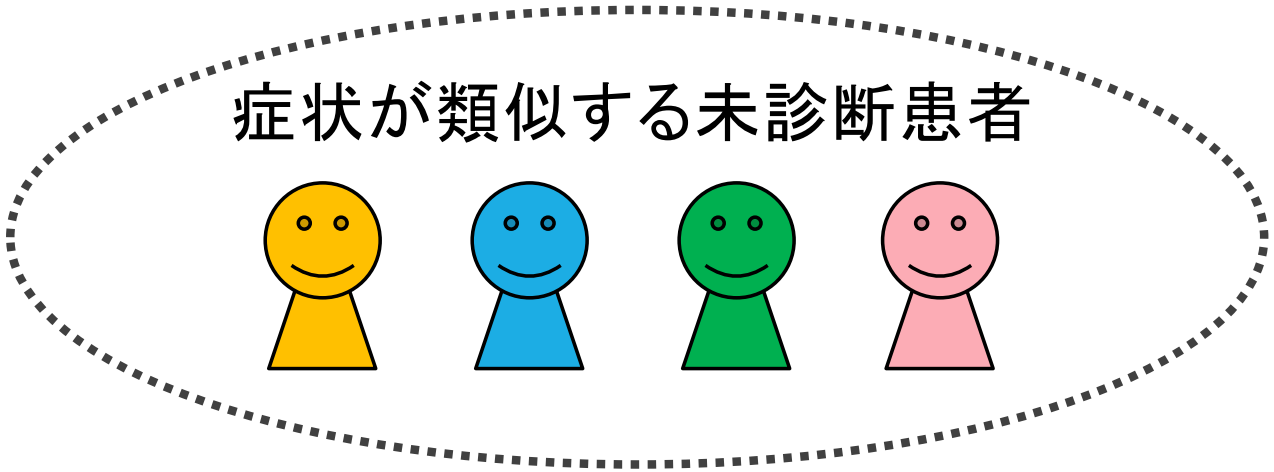
患者={偏頭痛} ⟷ 疾患={偏頭痛前兆あり}

類似度：1

患者={偏頭痛} ⟷ 疾患={偶発性頭痛}

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.humandiseasegenes.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

出所：Köhler, Sebastian, et al. "The human phenotype ontology in 2017." Nucleic acids research 45.D1 (2016): D865-D876.

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）

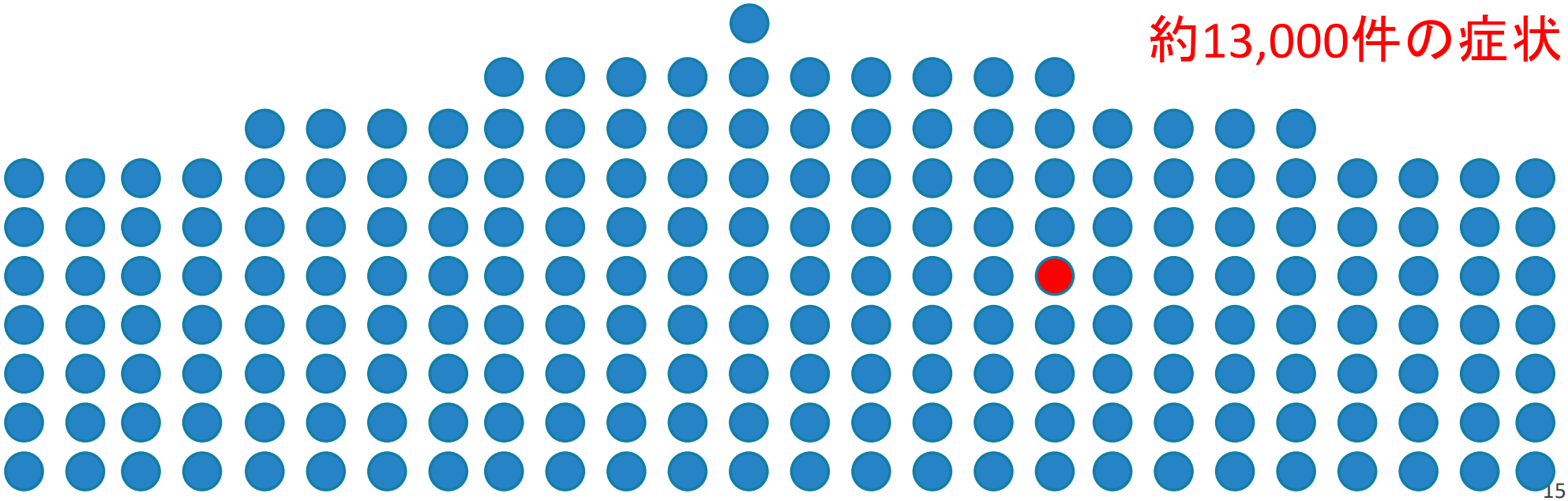
100以上の疾患原因遺伝子が同定された

問題点1：HPOの用語を探すのが困難

PubCaseFinder
Phenotype-Driven Differential-Diagnosis System

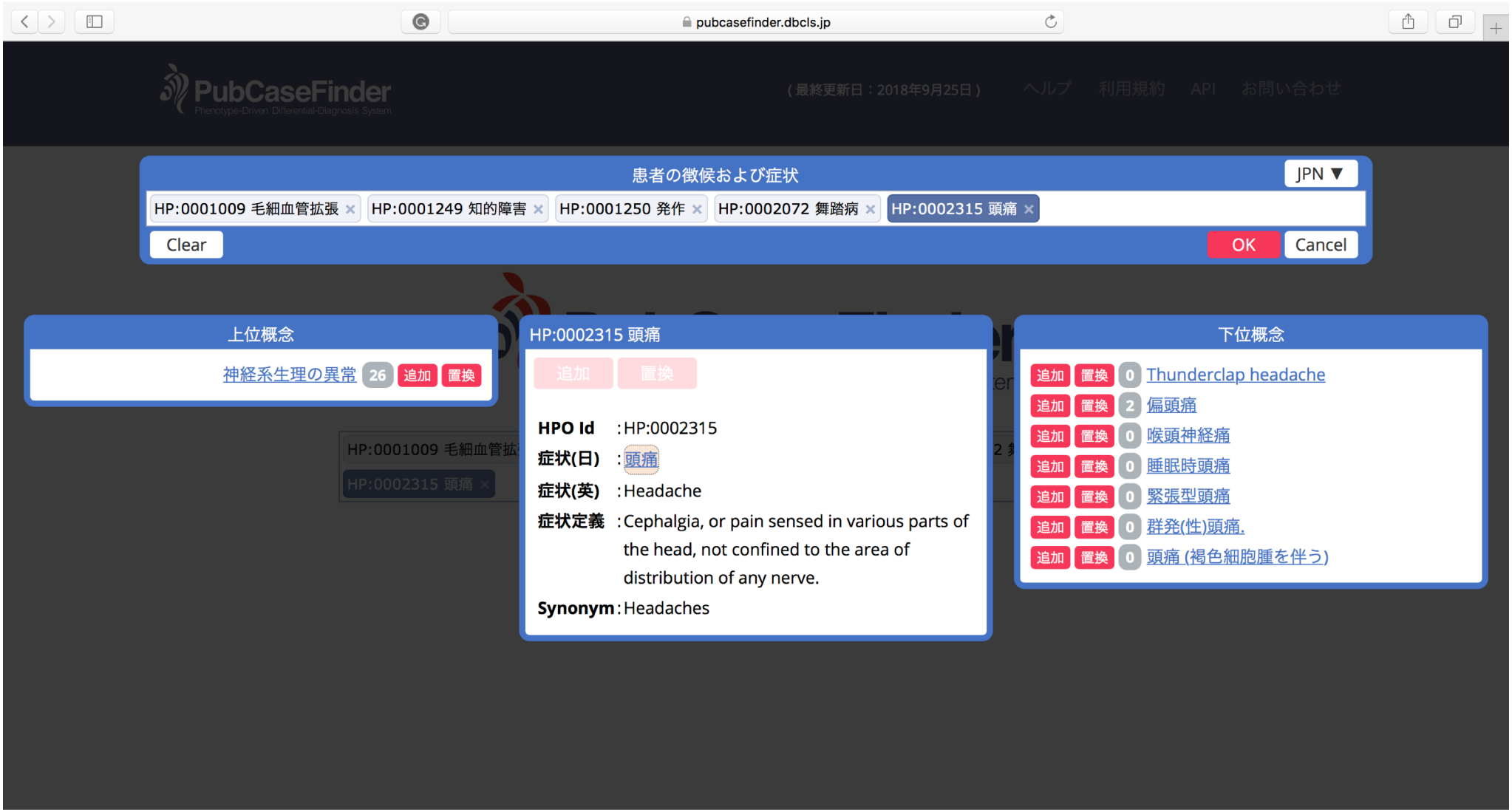
顔 ?

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



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

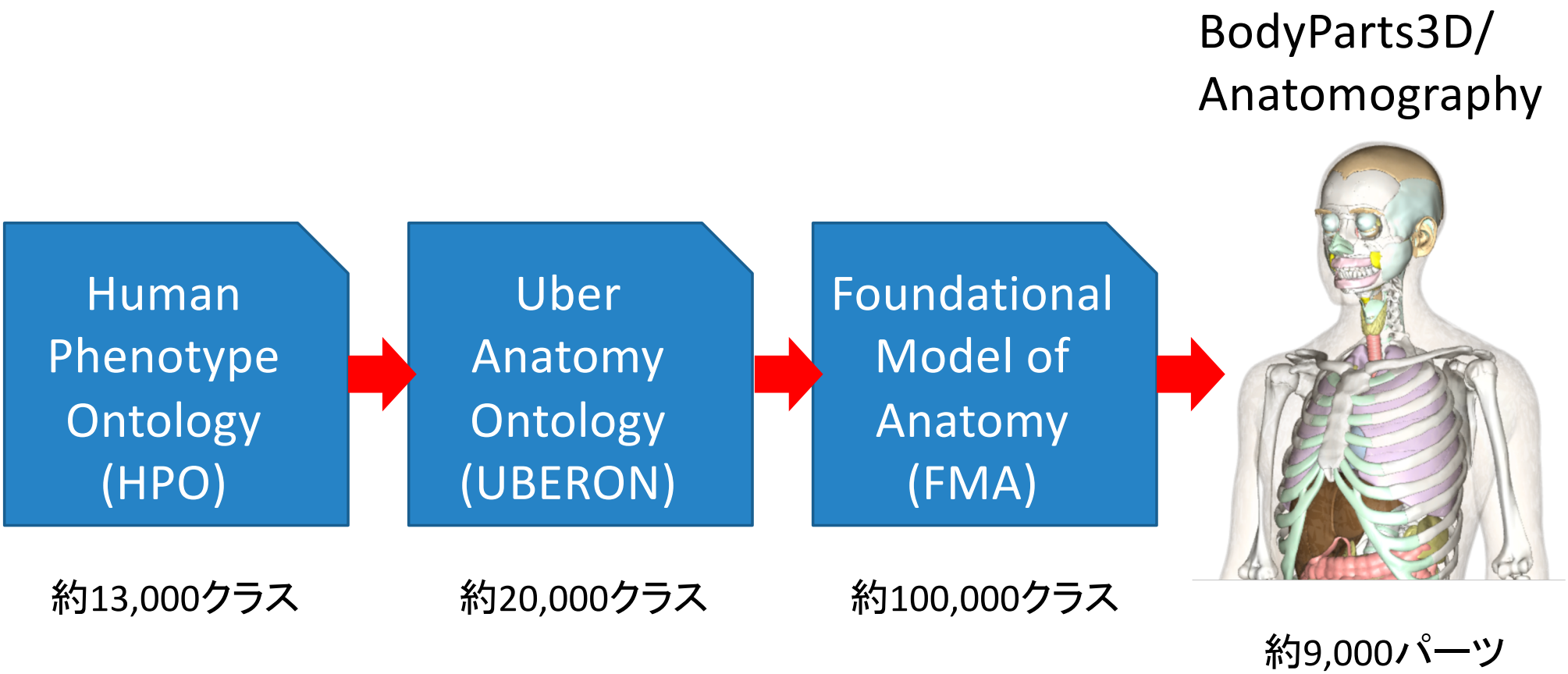


HPOの用語を探すのが困難

- 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

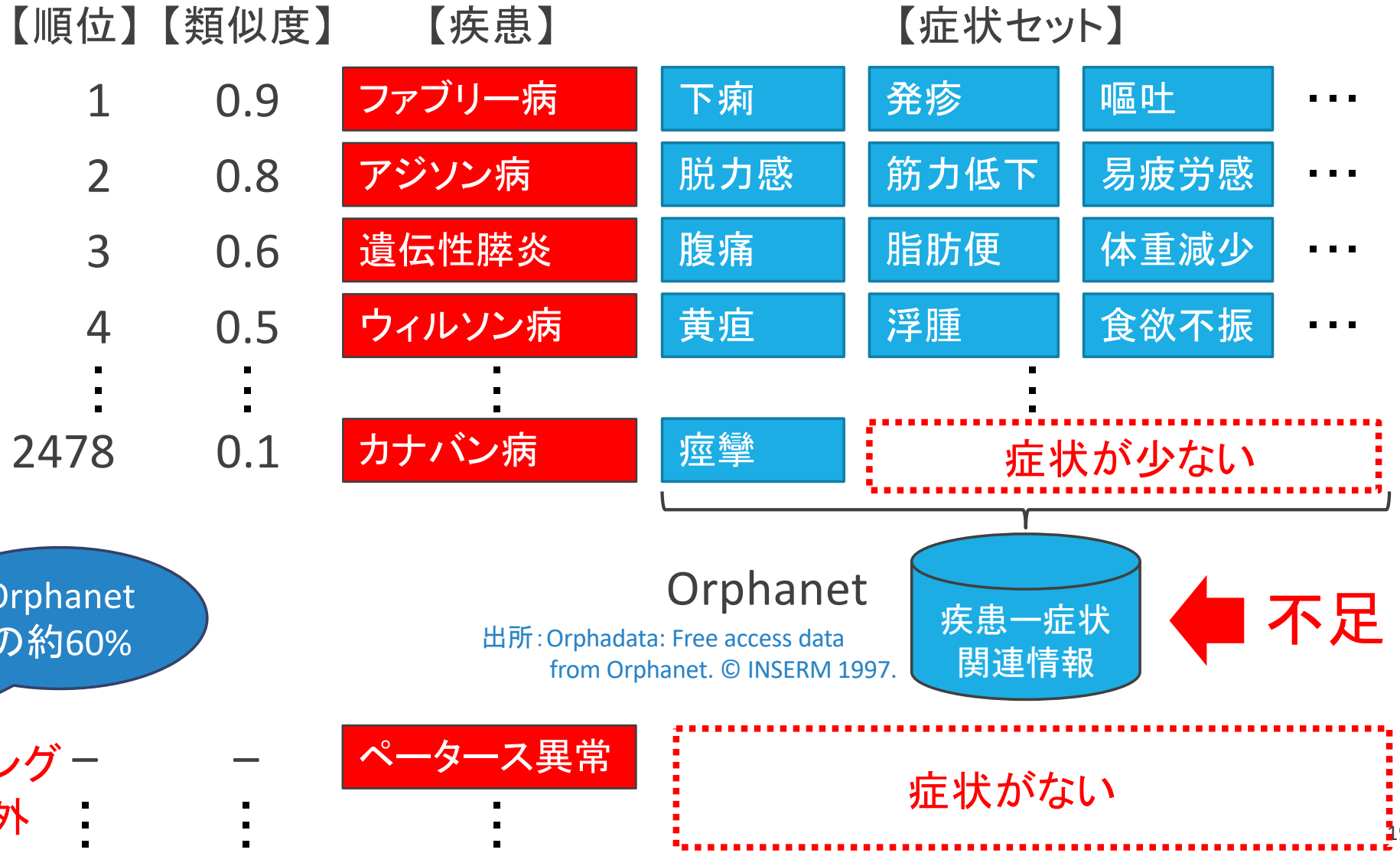
HPOの用語を探すのが困難

- 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：疾患ー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



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

文献から疾患に関連する症状を抽出

- 課題：「疾患ー症状」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

希少疾患

例） 遺伝性球状赤血球症（Orphanet:822）

- 溶血性貧血（HP:0001878）
- 黄疸（HP:0000952）
- 脾腫（HP:0001744）

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

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

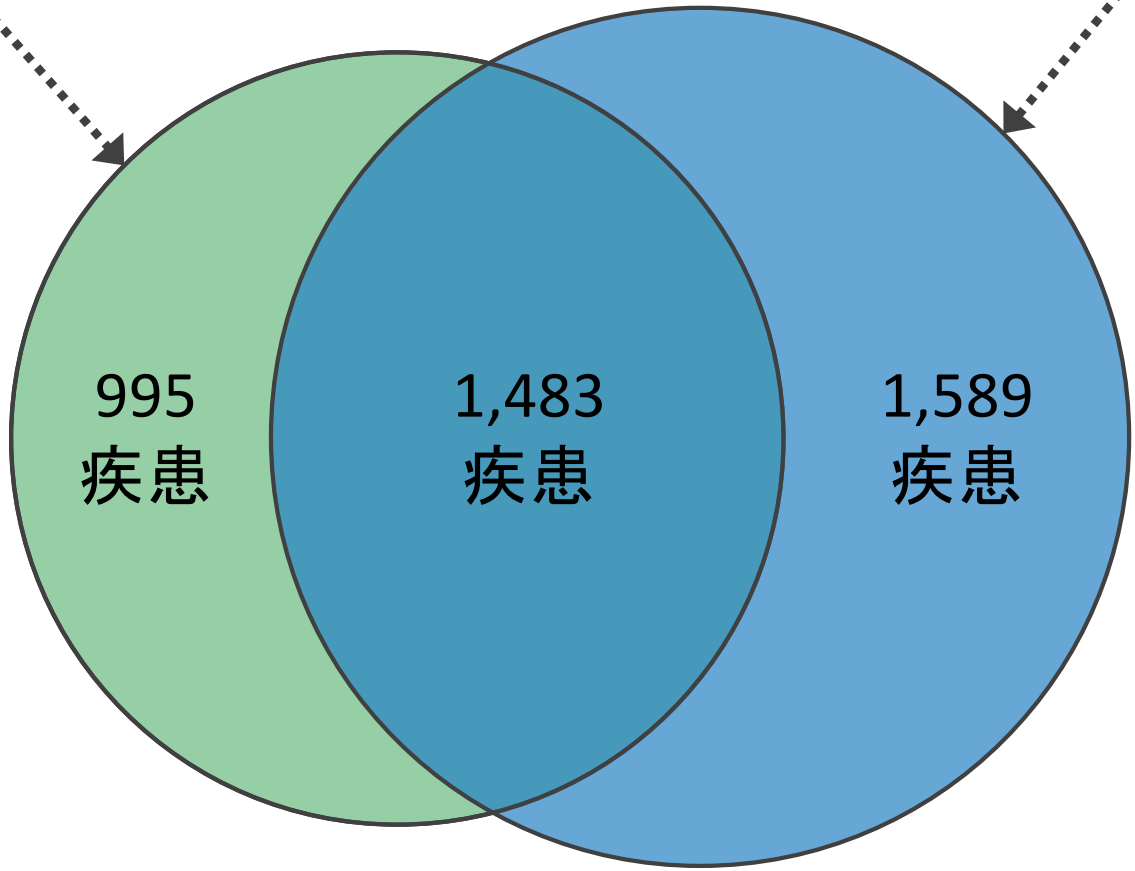
⇒ 17.7 (day)

⇒ 5.2 (hour)

疾患ー症状関連情報抽出　結果

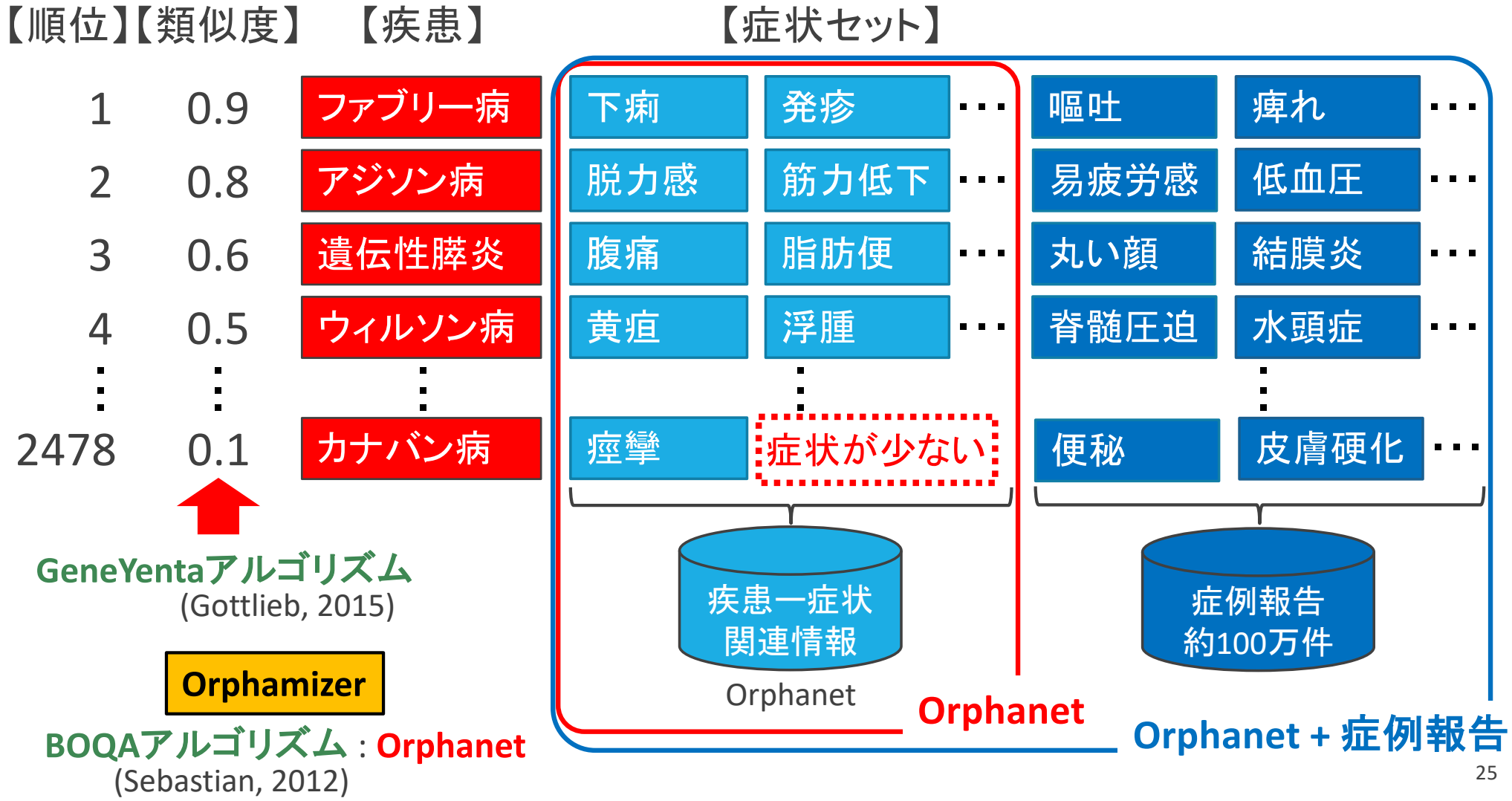
【Orphanet】
2,478疾患に症状を付与

【症例報告】
3,072疾患に症状を付与



取得した疾患ー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ー症状関連情報は、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

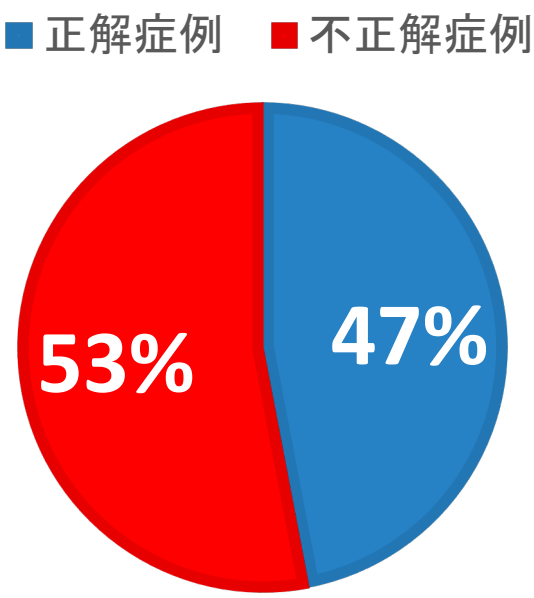
- 希少疾患の症例データセット（194件）で評価に利用
  - 診断結果（ORDO ID）、症状（HPO IDs）
- Recallで疾患ランキング精度を評価

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

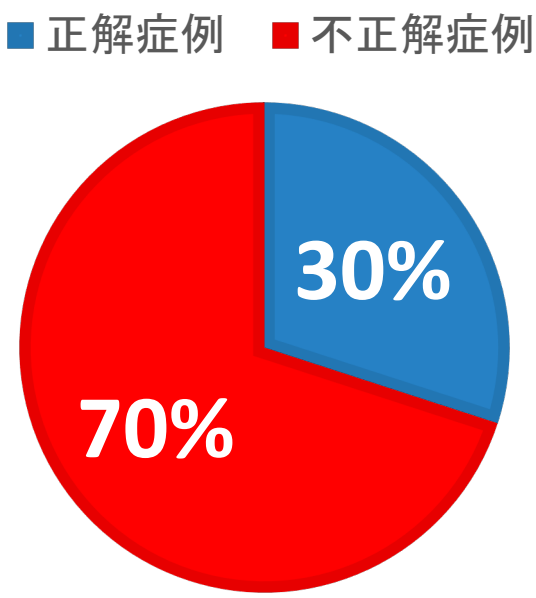
疾患ランキング精度の評価

- 評価用の135症例を利用して、トップ5のRecallを比較

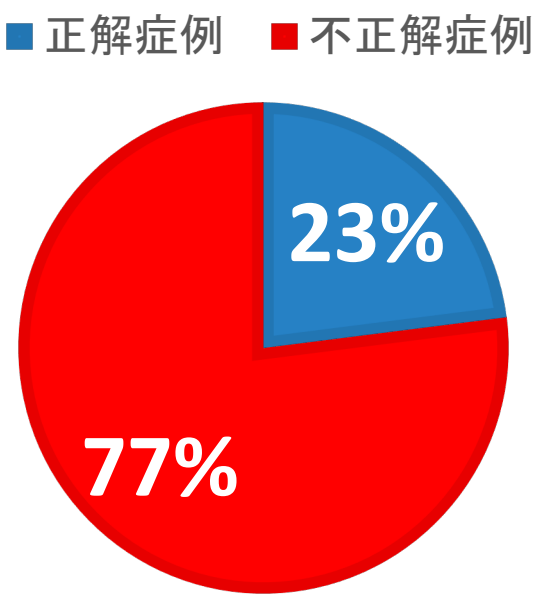
PubCaseFidner
（Orphanet + 症例報告）



PubCaseFidner
（Orphanet）

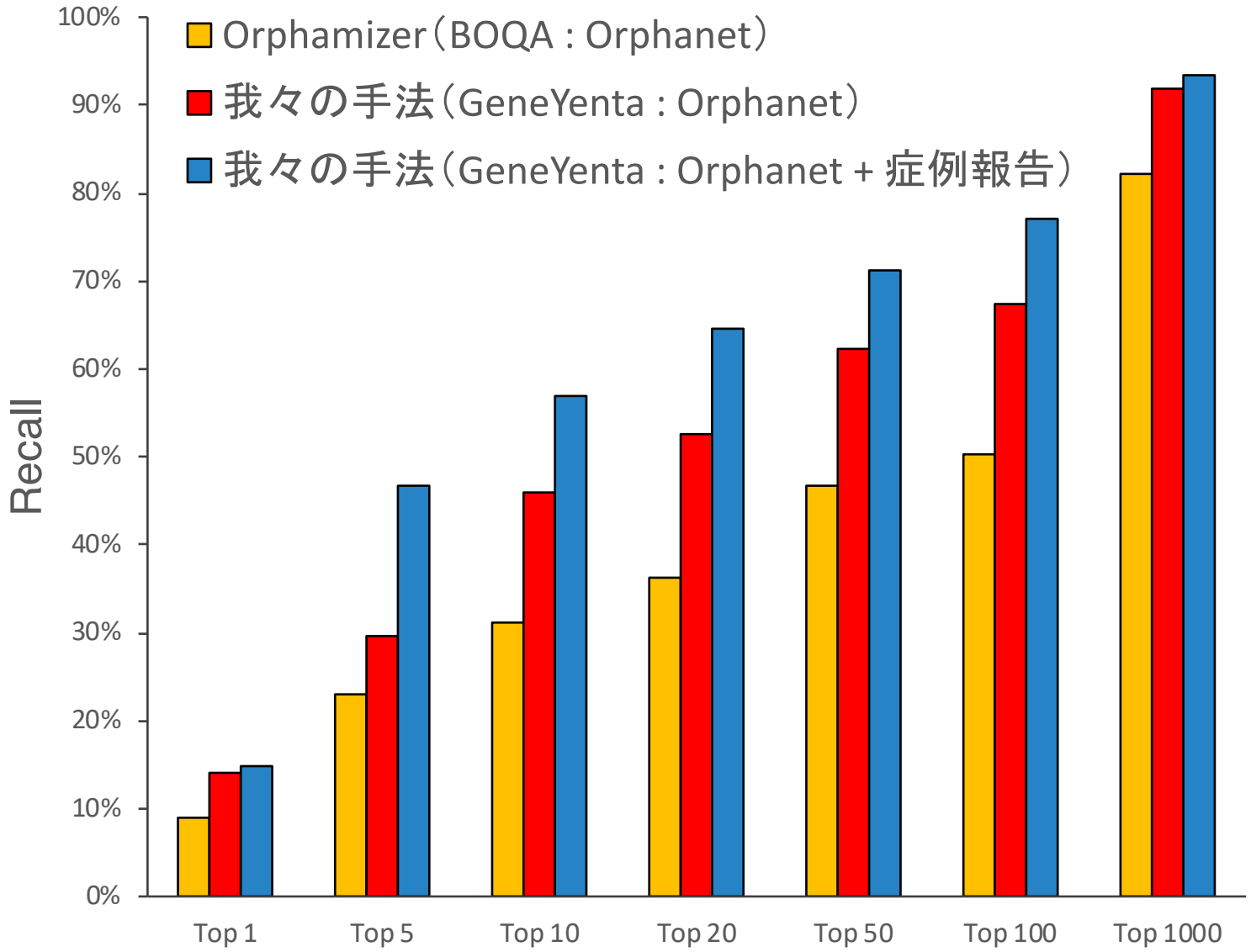


Orphamizer
（Orphanet）

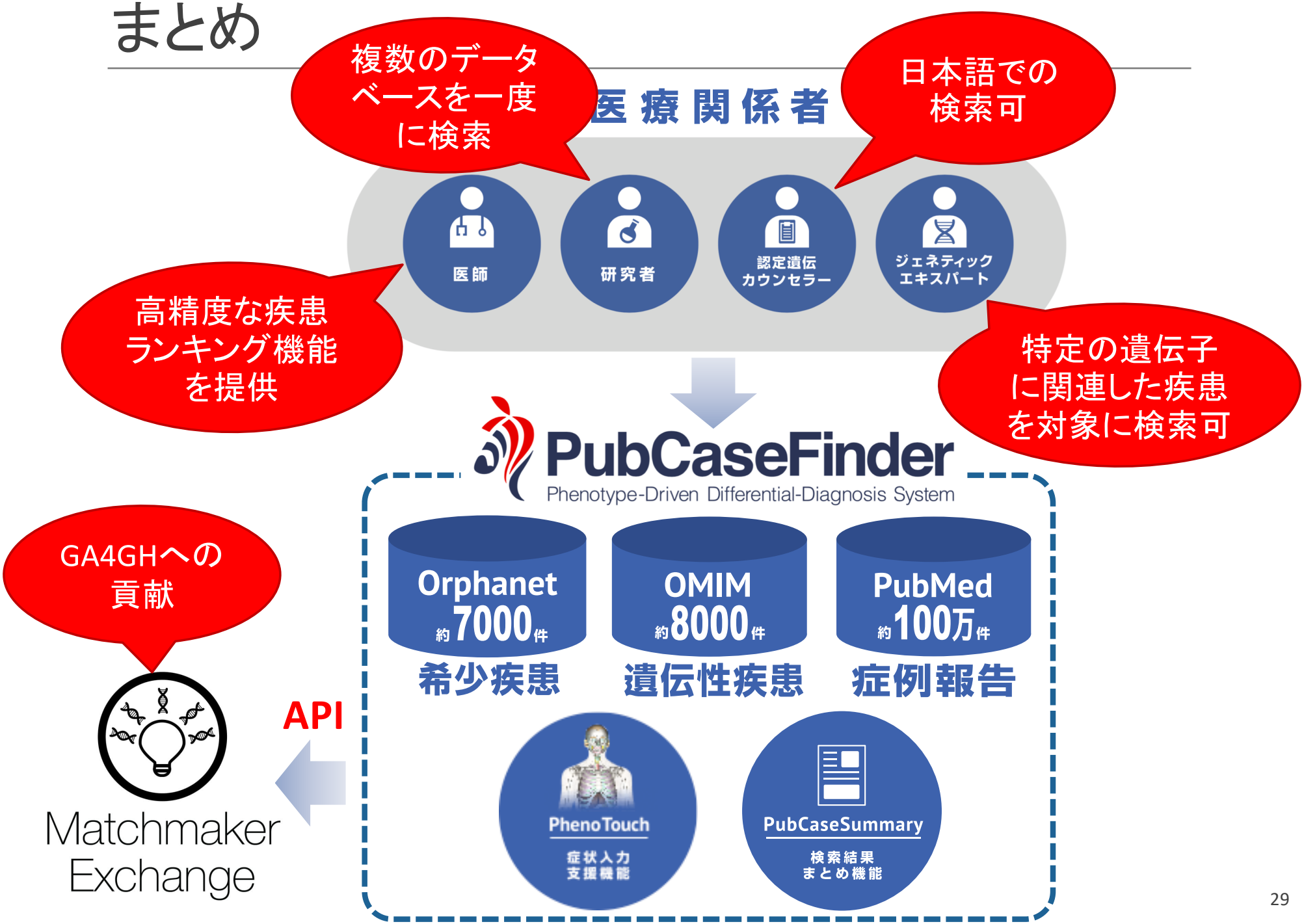


疾患ランキング精度の評価

- 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


複数のデータベースを一度に検索
医療関係者
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
API
GA4GHへの貢献
Matchmaker Exchange

謝辞

DBCLS

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

日立製作所

熊谷禎洋

BITS

武藤勇
山本利也

Gene42

Orion Buske

Genome.One

Tudor Groza

Care4Rare Canada Consortium

ご清聴ありがとうございました