Published online 6 November 2017 *Nucleic Acids Research, 2018, Vol. 46, Database issue* ***D977–D983***
doi: 10.1093/nar/gkx1049

PedAM: a database for Pediatric Disease Annotation and Medicine

Jinmeng Jia[1], Zhongxin An[1], Yue Ming[1], Yongli Guo[2], Wei Li[3], Xin Li[1], Yunxiang Liang[1], Dongming Guo[1], Jun Tai[2], Geng Chen[1], Yaqiong Jin[2], Zhimei Liu[2], Xin Ni[1,*] and Tieliu Shi[1,*]

[1]The Center for Bioinformatics and Computational Biology, Shanghai Key Laboratory of Regulatory Biology, Institute of Biomedical Sciences and School of Life Sciences, East China Normal University, Shanghai 200241, China, [2]Beijing Key Laboratory for Pediatric Diseases of Otolaryngology, Head and Neck Surgery, the Ministry of Education Key Laboratory of Major Diseases in Children, Beijing Pediatric Research Institute, Beijing Children's Hospital, Capital Medical University, National Center for Children's Health, Beijing 100045, China and [3]Beijing Key Laboratory for Genetics of Birth Defects, The Ministry of Education Key Laboratory of Major Diseases in Children, Center for Medical Genetics, Beijing Pediatric Research Institute, Beijing Children's Hospital, Capital Medical University, National Center for Children's Health, Beijing 100045, China

Received August 15, 2017; Revised October 04, 2017; Editorial Decision October 17, 2017; Accepted October 24, 2017

ABSTRACT

There is a significant number of children around the world suffering from the consequence of the misdiagnosis and ineffective treatment for various diseases. To facilitate the precision medicine in pediatrics, a database namely the Pediatric Disease Annotations & Medicines (PedAM) has been built to standardize and classify pediatric diseases. The PedAM integrates both biomedical resources and clinical data from Electronic Medical Records to support the development of computational tools, by which enables robust data analysis and integration. It also uses disease-manifestation (D-M) integrated from existing biomedical ontologies as prior knowledge to automatically recognize text-mined, D-M-specific syntactic patterns from 774 514 full-text articles and 8 848 796 abstracts in MEDLINE. Additionally, disease connections based on phenotypes or genes can be visualized on the web page of PedAM. Currently, the PedAM contains standardized 8528 pediatric disease terms (4542 unique disease concepts and 3986 synonyms) with eight annotation fields for each disease, including definition synonyms, gene, symptom, cross-reference (Xref), human phenotypes and its corresponding phenotypes in the mouse. The database PedAM is freely accessible at http://www.unimd.org/pedam/.

INTRODUCTION

Each year, there is a significant number of children worldwide suffering from the misdiagnosis and ineffective treatment for various diseases. Pediatric diseases have unique characteristics with a large portion of genetic diseases. Pediatric patients at their early ages are unable to describe their feelings clearly, which hurdles clinicians to make correct diagnoses. Many new techniques available for the diagnosis of adult diseases are difficult to be applied in pediatric patients. On the other hand, incorrect and delayed diagnoses frequently occur among pediatric patients. Therefore, a standardized description and diagnosis for pediatric diseases, and community-wide sharing clinical standards are essential to improve effectiveness of clinical applications.

The main challenge in systematic investigations of diseases is to translate information achieved from unstandardized clinical records to standardized and well-structured data (1). Over years, with increasingly in-depth studies in human diseases, new approaches for pediatric disease investigations have been developed. For example, researches have shown that the comparison study between pediatrics and adult diseases is one of the efficient ways to uncover the pathology or new subtypes of a specific disease, especially in genetic diseases and cancers (2,3) which has drawn a widely public attention in pediatric diseases. Meanwhile, along with the increasing awareness of pediatric diseases, much effort has been devoted to conduct preclinical and clinical researches in pediatrics. Although pediatric patients could gain significant benefits from these studies, the progress in pediatric disease research still significantly remains behind the studies in adult diseases. Efforts are needed in integrat-

[*]To whom correspondence should be addressed. Tel: +86 2154345020; Fax: +86 2154345020; Email: tlshi@bio.ecnu.edu.cn
Correspondence may also be addressed to Xin Ni. Email: nixin@bch.com.cn

ing, annotating and sharing of information in pediatric diseases (4).

Recently, precision medicine has opened a new door in disease treatment, but its practices mainly depend on two factors: the precisely personalized medicine trial and a close tracking of clinical manifestations, including symptoms and phenotypes (5,6). To apply precision medicine in clinical treatment, it is necessary to combine clinical-pathological data with state-of-art molecular profiling, and use these resources to precisely produce individuals' diagnostic, prognostic and therapeutic strategies. Although multiple electronic and administrative databases have been built (7,8), the heterogeneity of terms and the incompleteness of contents are generally existing in these sources, which significantly hinders the data sharing and exchange. Thus, a standardized pediatric disease database is urgently needed since it could offer a platform to integrate clinical and molecular data for clinicians to make a diagnosis and a treatment plan more efficiently. To achieve this goal, we have extensively collected information of pediatric diseases from published medical literature and clinical data, and integratively annotated enriched clinical and molecular data for most pediatric diseases, and finally standardized and classified all the information via different patterns and approaches to generate a pediatric disease annotation system, namely the Pediatric Disease Annotations and Medicines (PedAM). This database should be a valuable resource for researchers and clinicians to conduct studies and practice in pediatrics.

DISEASE DEFINITION

Disease name unification and cross-linkages

Standardization of disease names in pediatrics is the first step for the data integration, sharing and exchange, which can facilitate better communication and diagnosis of pediatric diseases (9). Considering the ubiquity of lexical heterogeneity in the realm of pediatric disease, a well-structured, completed lexicon of pediatric diseases is necessary (10–12). However, there is no a standard or a standardized disease category for pediatrics currently. We integrated the disease list from the following sources: (i) Medscape (13), including both pediatric disease records/conditions and articles; (ii) Dermatology Online Atlas (DermIS) (14), the atlas of pediatric dermatology; (iii) Case reports from both PubMed and PMC articles in MEDLINE (published from 2010 to 2015), with the reported ages ranging from 0 to 16 years old (15,16). We then mapped the disease names together with their alias strings to the Unified Medical Language System (UMLS) (17) to complete the textual unification. As a result, 8528 disease terms were added to the PedAM, containing 4542 unique disease concepts (Figure 1C) and 3986 alias strings. Considering the term usage variations in disease names and their identifiers (IDs), we mapped all the disease terms (including synonyms) among the currently controlled vocabularies and databases, including Online Mendelian Inheritance in Man (OMIM) (18), Disease Ontology (DO), international classification of diseases—version 10 (ICD10), UMLS, Medical Subject Headings (MeSH) and Systematized Nomenclature of Medicine—Clinical Terms (SNOMED-CT) (Figure 1B). Different pediatric disease identifiers mapped from the above databases were added as cross-reference (Xref) annotations.

Currently, the PedAM provides users with two ways for disease querying—precise disease search and fuzzy disease search. When querying diseases through precise search engine, users can search diseases by their name strings or IDs in UMLS, Orphanet and OMIM.

DISEASE ANNOTATION

Each disease term in the PedAM is annotated by eight aspects, including descriptions, synonyms, symptoms, genes, genotypes, Xref, human phenotypes and its corresponding phenotypes in the mouse (MPO).

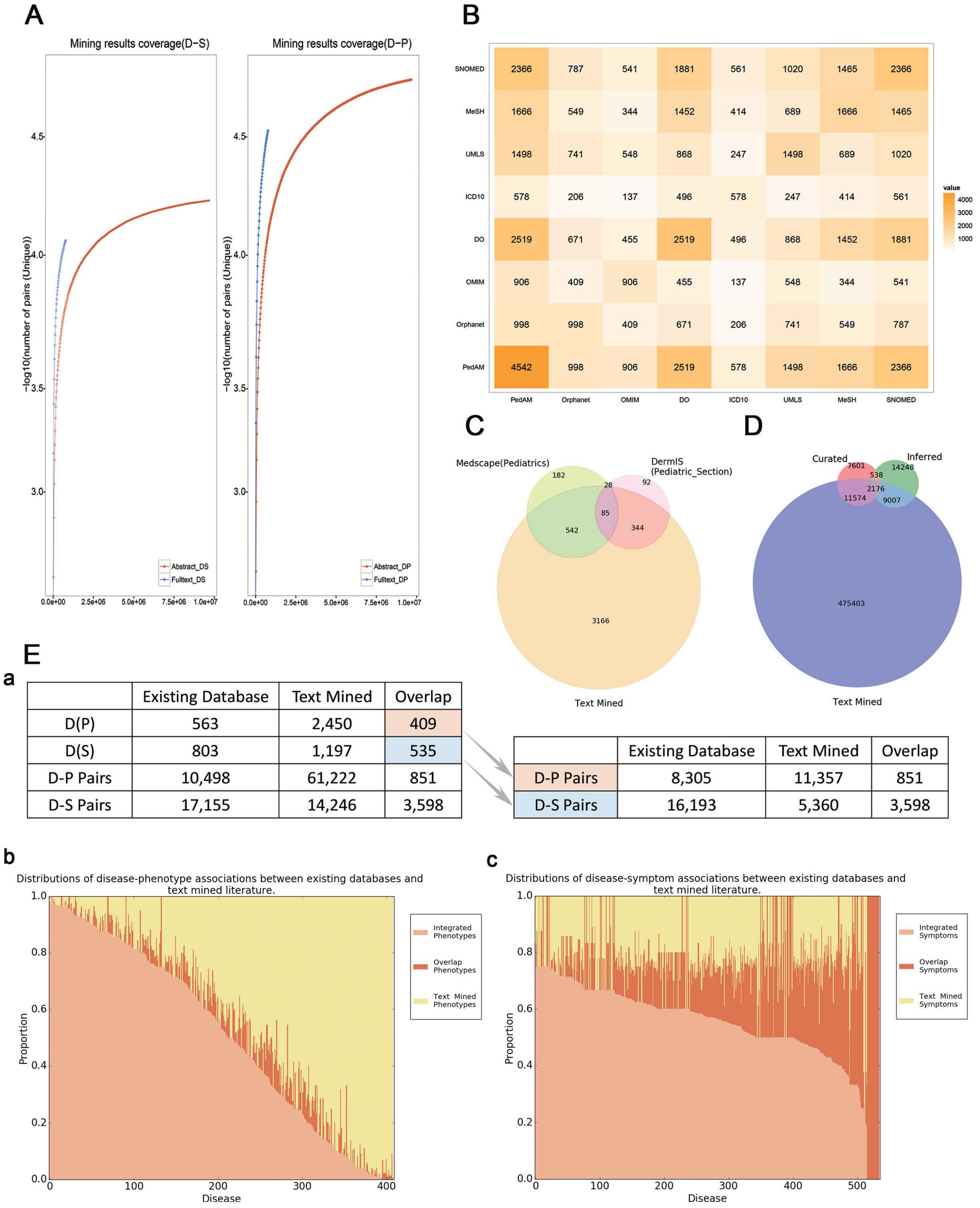
Disease descriptions

To disambiguate the meaning through different disease terms, a definition/short description for each disease is provided. For maximizing the description coverage of all the diseases in the PedAM, we extracted description/definition information from MRDEF.RRF file in UMLS (2017AA), OMIM, DO (19) and DermIS. Up to now, 3916 out of the total 4542 unique disease concepts in PedAM have their corresponding descriptions.

Disease symptoms and phenotypes

A phenotype, the expression of an organism's genotypes, is a collection of observable manifestations of genotypes. Accurate disease manifestations (phenotypes and symptoms) and adequate clinical records are critical for classifying, uncovering pathogenesis and finding new therapy for diseases (3). Given that phenotypes in HPO can realize a large-scale computational analysis of the human phenome (20), we defined terms from Human Phenotype Ontology (HPO) as 'Phenotype', while other manifestations as 'Symptom'. The PedAM obtained symptom and phenotype information from the following sources: (i) HPO (version 2017) (20); (ii) DO symptoms, using the 'has symptom' relationship; and (iii) MRREL.RRF in UMLS (2017AA). To ensure the high precision and integrity of results, a pattern based text mining approach that leverages external knowledge and limits the amount of human effort is adopted (21). We text mined the abstracts and full-text articles (from year 2010 to 2015) in MEDLINE. In total, 388 732 sentences together with 114 542 unique D-M (containing disease-phenotype and disease-symptom) annotations were generated. All the text mined results were curated manually. To evaluate the coverage of the text-mined D-M pairs, we calculated the number of unique D-M pairs extracted from articles with different sizes to observe the saturation trend. The extracted pairs from both abstracts and full-text articles showed a high coverage of phenotypic annotations of pediatric diseases, while full-text articles contained more phenotypic and disease information than abstracts (Figure 1A). All the Disease-Phenotype (D-P) and Disease-Symptom (D-M) corresponding sentences together with their PubMed identifiers (PMIDs) for each pediatric disease were provided in the PedAM.

Phenotype and symptom annotations in the PedAM are represented separately. Annotations generated by HPO



	PedAM	Orphanet	OMIM	DO	ICD10	UMLS	MeSH	SNOMED
SNOMED	2366	787	541	1881	561	1020	1465	2366
MeSH	1666	549	344	1452	414	689	1666	1465
UMLS	1498	741	548	868	247	1498	689	1020
ICD10	578	206	137	496	578	247	414	561
DO	2519	671	455	2519	496	868	1452	1881
OMIM	906	409	906	455	137	548	344	541
Orphanet	998	998	409	671	206	741	549	787
PedAM	4542	998	906	2519	578	1498	1666	2366

	Existing Database	Text Mined	Overlap
D(P)	563	2,450	409
D(S)	803	1,197	535
D-P Pairs	10,498	61,222	851
D-S Pairs	17,155	14,246	3,598

	Existing Database	Text Mined	Overlap
D-P Pairs	8,305	11,357	851
D-S Pairs	16,193	5,360	3,598

Figure 1. (**A**) Text-mined result coverage. Left panel shows the comparison of text-mined sentences containing symptoms from abstracts and full-text articles. Right panel shows the comparison of text-mined sentences containing phenotypes from abstracts and full-text articles. The *y*-axis represents the number of sentences containing disease-related phenotypes for each disease. DS stands for Disease-Symptom pairs while DP stands for Disease-Phenotype pairs. (**B**) Overlaps among different disease resources. The symmetric matrix shows the number of overlapping diseases between all pairs of primary name sources according to PedAM mapping. Colors and numerals represent the overlapping degree in disease counts. Source abbreviations: DO—Disease Ontology. (**C**) Overlaps among the major disease sources. Venn diagram for the major sources of disease names in PedAM. (**D**) Overlaps among the major gene sources. Venn diagram for the major sources of disease related genes in PedAM. (E) (a) Statistics of diseases and D-M pairs which have phenotypic annotation between existing databases and text mined literature. (b) Distributions of disease-phenotype associations between existing databases and text-mined literature. The *x*-axis represents diseases which phenotypic annotation comes from both existing databases and literature, while *y*-axis represents the proportion of disease corresponding phenotypes. (c) Distributions of disease-symptom associations between existing databases and text-mined literature. The *x*-axis represents diseases which phenotypic annotation comes from both existing databases and literature, while *y*-axis represents the proportion of disease corresponding phenotypes.

terms are shown in 'Phenotype', while others are shown in 'Symptom'. Records in both phenotype and symptom contain information integrated from currently existing databases as previously mentioned and relevant text mining, resulting in the maximum phenotypic information for pediatric diseases (Figure 1E).

To show possible molecular mechanisms of a disease in a more intuitive way, we display the phenotype and gene based disease connections for each disease. The PedAM provides the top 10 closely connected diseases for each retrieved disease based on shared phenotypes or genes. In addition, users can also check the connections between any two diseases of interest. The systematic collection of disease symptoms and phenotypes can help clinicians to distinguish diseases with similar symptoms in clinical practice.

Disease genes and genotypes

The associations between genes and novel phenotypes identified by NGS necessitate the re-annotation of human disease-causing genes (22). In this study, we collected disease-gene associations from several existing databases (Table 1). To make a better classification, we divided all these associations into three categories: curated (data manually curated by experts or validated by experiments), text mined and inferred disease-gene associations. In total, the PedAM contains 571 450 disease-gene association records currently, including 20 003 genes and 3133 diseases.

The genotype information was archived from following sources: (i) publicly available databases: DisGeNET, GWASdb (23), LOVD (24) and PharmGKB (25); and (ii) datasets from Beijing Children's hospital. In total, the PedAM contains 169 841 gene variants records.

Drug

To annotate drugs for each disease, we categorized two sets of data to extract disease-drug associations: the drug dataset was generated to archive chemical names from the DrugBank (26) and U.S. Food and Drug Administration (FDA) (https://www.fda.gov) drug database, while disease dataset was made from all the unique diseases in PedAM and their synonyms. We then extracted disease-drug pairs from Comparative Toxicogenomics Database (27). To keep the results precisely, we only included reserving marker/mechanism and therapeutic drugs in PedAM. Drugs and diseases were standardized by FDA chemical names and UMLS disease names. PedAM provides 10 063 disease-drug associations, including 1004 diseases and 483 drugs (chemicals).

Given the fact that drug dosage information is crucial in pediatric disease therapy and efforts have been made to distinguish differences of drug effectiveness and side effects (SE) between pediatric and adult patients (28–31), we also collected drug dosage and efficiency information from drug labels among FDA approved drugs for pediatrics. In total, 746 FDA approved drugs together with 397 drug dosage and efficiency information involved in 239 diseases were added to PedAM. Users can view this information through clicking the 'related drugs' button after retrieving the disease of interest.

DISEASE CONNECTIONS

Connecting diseases with similar pathological mechanisms can advance the understanding of the disease landscape and facilitate the mechanism-based exploring of new drug treatments (32). Usually, disease pairs sharing more affected genes or phenotypes are more likely to have similar pathological mechanisms (33,34). Thus, we connected disease pairs in PedAM by two approaches: (i) phenotype-based approaches, by which we calculated the phenotypic vector similarity between every two diseases; and (ii) gene-based approaches, by which we calculated the associations of each disease pairs by adding up the uniqueness of their entire shared gene(s). Phenotypes and genes are generated from the existing databases or text mined data from MEDLINE. The calculation details were shown as below.

Calculation of phenotype-based disease similarity

We computed phenotype-based disease similarity by calculating the phenotypic vector similarity between every two diseases as described in the previous research (35). We describe every disease ***j*** by a vector of phenotypes d_j

$$d_j = \left(w_{1,j},\ w_{2,j}, \cdots, w_{n,j}\right), \tag{1}$$

where, $w_{i,j}$ represents the quantitative strength of association between phenotype *i* and disease *j*:

$$w_{i,j} = W_{i,j} \log \frac{N}{n_i} \tag{2}$$

where, $W_{i,j}$ denotes the co-occurrence strength of association between phenotype *i* and disease *j*, and *N* denotes the number of the diseases that have phenotypic annotations and n_i stands for the number of diseases for which phenotype *i* (phenotype annotated to at least one disease) appears. Therefore, for every disease *j,* it has a vector of d_j length *i.* Then we calculated the cosine similarity value between the vectors d_m and d_n of two diseases *m* and *n* as follows:

$$\cos\left(d_i, d_j\right) = \frac{\sum_i d_{m,i} d_{n,i}}{\sqrt{\sum_i d_{m,i}^2}\sqrt{\sum_i d_{n,i}^2}} \tag{3}$$

The cosine similarity ranges from 0 (no shared phenotype) to 1 (identical phenotypes).

Calculation of gene-based disease similarity

We adapted the equation for the calculation of the gene uniqueness-based disease similarity to calculate the gene-based disease similarity (36). The uniqueness of each gene *i* was calculated as follows:

$$u_i = 1 - \sqrt{\frac{d_i}{d_n}} \tag{4}$$

where, d_i is described as the number of diseases associated with each gene *i* and d_n is the number of unique disease concepts in PedAM.

We then created an $N \times N$ matrix to represent the similarity between every two diseases. For each pair of diseases, we calculated their similarity score:

$$d_{ij} = u_{s_1} + u_{s_2} + \cdots + u_{s_n} \tag{5}$$

Table 1. Disease-gene association resources

Genotype resource	Access linkage	Counts
Curated Information—data manually curated by experts or validated by experiments		
Orphanet	http://www.orpha.net/	2257
OMIM (Online Mendelian Inheritance in Man)	http://omim.org/	123
UniProtKB (53) (UniProt Knowledgebase)	http://www.uniprot.org/	477
ClinVar (54)	https://www.ncbi.nlm.nih.gov/clinvar/	938
GHR (55) (Genetics Home Reference)	https://ghr.nlm.nih.gov/gene/GHR	4398
DisGeNET (56)	http://www.disgenet.org/	10 255
Text mined information		
DISEASES (57)	http://diseases.jensenlab.org/	515 378
Inferred information—predicted by protein interaction and phenotype network		
CIPHER (58) (Correlating protein interaction network and phenotype network to predict disease genes)	http://bioinfo.au.tsinghua.edu.cn/jianglab/cipher/	25 969
In total—total number of disease-Gene associations		520 537

Counts stand for the pediatric disease and gene associations extracted from designated resources.

where, d_{ij} represents any one of the disease pairs and u_{s_n} represents the uniqueness value of each gene shared between d_{ij}.

Users can retrieve top 10 similar diseases based on either phenotype-based or gene-based similarity for each disease in PedAM. Moreover, disease connections using integrated data or all data (containing integrated data and text mined data) are all provided. In addition, to ensure the integrity of the gene-based disease network, we combined disease-gene associations from all three categories (curated, text mined and inferred), and obtained a gene-disease matrix connecting 20 003 genes with 3133 disease entries. The resulted associations extracted from the inferred and text-mined categories largely intensify the disease network, suggesting much complicated relationships among diseases.

Above two approaches built up connections among diseases in PedAM. Notably, when studying mechanism-based disease connectivity in pediatric disease, connections to other diseases are also very informative. Thus, we used the former disease definition method to integrate disease concepts from PedAM, DO and OMIM, which obtained 171 938 disease terms (including synonyms). We adopted the pattern recognition approach described above to mine the literature from MEDLINE database. In total, 8 488 796 abstracts and 774 514 full-text articles were text-mined respectively, resulting in 180 350 unique disease comorbidity pairs involving in 586 743 sentences. The text mining results were then manually curated by experts in Beijing Children's hospital. All the text mined sentences as well as their PMID are recorded in the PedAM.

DISCUSSION

Early diagnosis and treatment can obviously improve prognosis and life quality of patients. To facilitate the early diagnosis and treatment in pediatrics, and systematical investigation of disease pathologies or molecular mechanisms, standardized 'omics' data and understanding the connections among 'omics' data are always crucial (37–39). Precision medicine has been proposed to be the most effective strategy in current clinical applications, and big data based applications with artificial intelligence technology in clinical fields have also been considered as a great potential for accurate disease diagnoses. However, these promising applications greatly rely on data mining, sharing and exchange. Currently, one of the major challenges for the application of big data is the lack of uniformly structured data in clinical practice, because different Healthy Information Systems and different Electronic Medical Records are applied by different hospitals, and different vocabularies are used to describe clinical observations and treatment strategies by different healthcare providers. The high disparity in clinical data among hospitals and healthcare providers is a significant obstacle for the data sharing and downstream analysis. In addition, as there is a considerable amount of information stored in published articles and Electronic Health Records (EHRs), transforming them from unstructured descriptions to well-structured clinical data will efficiently benefit data analysis and application in clinical practice (40). Thus, comprehensively integrating multi-omics data from different resources is the first step to accomplish the goal (41,42).

As a comprehensive platform for pediatric disease research and diagnoses, the PedAM provides enriched clinical and molecular annotations for 4542 pediatric diseases, containing 4893 human disease-related phenotypic terms, 26 332 mammalian phenotypic terms and 9840 symptoms from UMLS, in addition to 20 003 genes and 169 841 genotype records. Currently, each disease term in PedAM is annotated in eight aspects, including definition, synonyms, symptoms, genes, genotypes, Xref, human phenotypes and its corresponding phenotypes in the mouse (MPO).

To promote clinical data sharing and exchange for better diagnosis and treatment of pediatric diseases, we will continue to extract and standardize clinically related information regarding symptoms, phenotypes and case reports for pediatric diseases from published literature to further enrich our PedAM database. On the other hand, we have built up a cooperation relationship with *Science China Life Sciences* and *Pediatric Investigation* journals to promote data sharing. The database PedAM is designated as the data repository to host all the pediatrics related clinical data and research results published on both journals. For example, in the recently published thematic issue of rare pediatric diseases by *Science China Life Sciences* journal, all patient de-identified clinical data and molecular results have been stored in our PedAM (43–52). We believe that with more standardized clinical data and information accumulated in the PedAM, this platform will help medical professionals in this community to gain more knowledge, and make more accurate diagnoses for pediatric diseases. The database PedAM should greatly facilitate a better understanding of underlying mechanisms for complex pediatric diseases, which will ultimately benefit pediatric patients and their parents.

ACKNOWLEDGEMENTS

We are grateful to Dr Baitang Ning from National Center of Toxicological Research of the FDA for his great help, Clinicians from Beijing Children's Hospital for their helps of manual curation of the text-mined results.

FUNDING

China Human Proteomics Project [2014DFB30010, 2014DFB30030]; National High Technology Research and Development Program of China [2015AA020108]; National Natural Science Foundation of China [31671377]; Shanghai 111 Project [B14019]; Beijing Municipal Administration of Hospitals Clinical Medicine Development of Special Funding Support [ZYLX201508]; Beijing Municipal Science and Technology Project [D131100005313014]. Funding for open access charge: China Human Proteomics Project [2014DFB30010] and National High Technology Research and Development Program of China [2015AA020108].
Conflict of interest statement. None declared.

REFERENCES

1. Rodriguez-Galindo,C., Friedrich,P., Alcasabas,P., Antillon,F., Banavali,S., Castillo,L., Israels,T., Jeha,S., Harif,M., Sullivan,M.J. *et al.* (2015) Toward the Cure of All Children With Cancer Through Collaborative Efforts: Pediatric Oncology As a Global Challenge. *J. Clin. Oncol.*, **33**, 3065–3073.
2. Chen,L., Wang,M., Fan,H., Hu,F. and Liu,T. (2017) Comparison of pediatric and adult lymphomas involving the mediastinum characterized by distinctive clinicopathological and radiological features. *Sci. Rep.*, **7**, 2577.
3. Patel,M.D., Mohan,J., Schneider,C., Bajpai,G., Purevjav,E., Canter,C.E., Towbin,J., Bredemeyer,A. and Lavine,K.J. (2017) Pediatric and adult dilated cardiomyopathy represent distinct pathological entities. *JCI Insight*, **2**, e94382.
4. Hay,W.W. Jr, Gitterman,D.P., Williams,D.A., Dover,G.J., Sectish,T.C. and Schleiss,M.R. (2010) Child health research funding and policy: imperatives and investments for a healthier world. *Pediatrics*, **125**, 1259–1265.
5. Arnedos,M., Vicier,C., Loi,S., Lefebvre,C., Michiels,S., Bonnefoi,H. and Andre,F. (2015) Precision medicine for metastatic breast cancer—limitations and solutions. *Nat. Rev. Clin. Oncol.*, **12**, 693–704.
6. Mirnezami,R., Nicholson,J. and Darzi,A. (2012) Preparing for precision medicine. *N. Engl. J. Med.*, **366**, 489–491.
7. Gipson,D.S., Kirkendall,E.S., Gumbs-Petty,B., Quinn,T., Steen,A., Hicks,A., McMahon,A., Nicholas,S., Zhao-Wong,A., Taylor-Zapata,P. *et al.* (2017) Development of a Pediatric Adverse Events Terminology. *Pediatrics*, **139**, e20160985.
8. Wolbrink,T.A., Kissoon,N., Mirza,N. and Burns,J.P. (2017) Building a global, online community of practice: the OPENPediatrics World Shared Practices Video Series. *Acad. Med.*, **92**, 676–679.
9. Mascalzoni,D., Knoppers,B.M., Ayme,S., Macilotti,M., Dawkins,H., Woods,S. and Hansson,M.G. (2013) Rare diseases and now rare data? *Nat. Rev. Genet.*, **14**, 372.
10. Rappaport,N., Twik,M., Plaschkes,I., Nudel,R., Iny Stein,T., Levitt,J., Gershoni,M., Morrey,C.P., Safran,M. and Lancet,D. (2017) MalaCards: an amalgamated human disease compendium with diverse clinical and genetic annotation and structured search. *Nucleic Acids Res.*, **45**, D877–D887.
11. Nambot,S., Gavrilov,D., Thevenon,J., Bruel,A.L., Bainbridge,M., Rio,M., Goizet,C., Rotig,A., Jaeken,J., Niu,N. *et al.* (2017) Further delineation of a rare recessive encephalomyopathy linked to mutations in GFER thanks to data sharing of whole exome sequencing data. *Clin. Genet.*, **92**, 188–198.
12. Trama,A., Marcos-Gragera,R., Sanchez Perez,M.J., van der Zwan,J.M., Ardanaz,E., Bouchardy,C., Melchor,J.M., Martinez,C., Capocaccia,R., Vicentini,M. *et al.* (2017) Data quality in rare cancers registration: the report of the RARECARE data quality study. *Tumori*, **103**, 22–32.
13. Byrne,G. (2015) MedScape. *Nurs. Stand.*, **29**, 29.
14. Diepgen,T.L. and Eysenbach,G. (1998) Digital images in dermatology and the Dermatology Online Atlas on the World Wide Web. *J. Dermatol.*, **25**, 782–787.
15. Lozano,P., Henrikson,N.B., Morrison,C.C., Dunn,J., Nguyen,M., Blasi,P. and Whitlock,E.P. (2016) *Lipid Screening in Childhood for Detection of Multifactorial Dyslipidemia: a Systematic Evidence Review for the U.S. Preventive Services Task Force*, Rockville.
16. Rose,K. and Van den Anker,J.N. (2010) *Guide to Paediatric Drug Development and Clinical Research*. Karger, Basel.
17. Lindberg,C. (1990) The Unified Medical Language System (UMLS) of the National Library of Medicine. *J. Am. Med. Rec. Assoc.*, **61**, 40–42.
18. Amberger,J.S., Bocchini,C.A., Schiettecatte,F., Scott,A.F. and Hamosh,A. (2015) OMIM.org: Online Mendelian Inheritance in Man (OMIM(R)), an online catalog of human genes and genetic disorders. *Nucleic Acids Res.*, **43**, D789–D798.
19. Kibbe,W.A., Arze,C., Felix,V., Mitraka,E., Bolton,E., Fu,G., Mungall,C.J., Binder,J.X., Malone,J., Vasant,D. *et al.* (2015) Disease Ontology 2015 update: an expanded and updated database of human diseases for linking biomedical knowledge through disease data. *Nucleic Acids Res.*, **43**, D1071–D1078.
20. Kohler,S., Vasilevsky,N.A., Engelstad,M., Foster,E., McMurry,J., Ayme,S., Baynam,G., Bello,S.M., Boerkoel,C.F., Boycott,K.M. *et al.* (2017) The Human Phenotype Ontology in 2017. *Nucleic Acids Res.*, **45**, D865–D876.
21. Xu,R., Li,L. and Wang,Q. (2013) Towards building a disease-phenotype knowledge base: extracting disease-manifestation relationship from literature. *Bioinformatics*, **29**, 2186–2194.
22. Boycott,K.M., Vanstone,M.R., Bulman,D.E. and MacKenzie,A.E. (2013) Rare-disease genetics in the era of next-generation sequencing: discovery to translation. *Nat. Rev. Genet.*, **14**, 681–691.
23. Li,M.J., Liu,Z., Wang,P., Wong,M.P., Nelson,M.R., Kocher,J.P., Yeager,M., Sham,P.C., Chanock,S.J., Xia,Z. *et al.* (2016) GWASdb v2: an update database for human genetic variants identified by genome-wide association studies. *Nucleic Acids Res.*, **44**, D869–D876.
24. Pan,M., Cong,P., Wang,Y., Lin,C., Yuan,Y., Dong,J., Banerjee,S., Zhang,T., Chen,Y., Zhang,T. *et al.* (2011) Novel LOVD databases for hereditary breast cancer and colorectal cancer genes in the Chinese population. *Hum. Mutat.*, **32**, 1335–1340.
25. Thorn,C.F., Klein,T.E. and Altman,R.B. (2013) PharmGKB: the Pharmacogenomics Knowledge Base. *Methods Mol. Biol.*, **1015**, 311–320.
26. Wishart,D.S., Knox,C., Guo,A.C., Shrivastava,S., Hassanali,M., Stothard,P., Chang,Z. and Woolsey,J. (2006) DrugBank: a comprehensive resource for in silico drug discovery and exploration. *Nucleic Acids Res.*, **34**, D668–D672.
27. Davis,A.P., Grondin,C.J., Johnson,R.J., Sciaky,D., King,B.L., McMorran,R., Wiegers,J., Wiegers,T.C. and Mattingly,C.J. (2017) The Comparative Toxicogenomics Database: update 2017. *Nucleic Acids Res.*, **45**, D972–D978.

28. Ward,R.M. (2001) Children, drugs, and the Food and Drug Administration: studies of pediatric drugs are beginning to catch up. *Pediatr. Ann.*, **30**, 189–194.
29. Zimmerman,K., Putera,M., Hornik,C.P., Brian Smith,P., Benjamin,D.K. Jr, Mulugeta,Y., Burckart,G.J., Cohen-Wolkowiez,M. and Gonzalez,D. (2016) Exposure matching of pediatric anti-infective drugs: review of drugs submitted to the Food and Drug Administration for pediatric approval. *Clin. Therapeut.*, **38**, 1995–2005.
30. Buck,M.L. (2000) Impact of new regulations for pediatric labeling by the Food and Drug Administration. *Pediatr. Nurs.*, **26**, 95–96.
31. Maxey,D.M., Ivy,D.D., Ogawa,M.T. and Feinstein,J.A. (2013) Food and Drug Administration (FDA) postmarket reported side effects and adverse events associated with pulmonary hypertension therapy in pediatric patients. *Pediatr. Cardiol.*, **34**, 1628–1636.
32. Liu,C.C., Tseng,Y.T., Li,W., Wu,C.Y., Mayzus,I., Rzhetsky,A., Sun,F., Waterman,M., Chen,J.J., Chaudhary,P.M. *et al.* (2014) DiseaseConnect: a comprehensive web server for mechanism-based disease-disease connections. *Nucleic Acids Res.*, **42**, W137–W146.
33. Pinero,J., Berenstein,A., Gonzalez-Perez,A., Chernomoretz,A. and Furlong,L.I. (2016) Uncovering disease mechanisms through network biology in the era of Next Generation Sequencing. *Sci. Rep.*, **6**, 24570.
34. Nabhan,A.R. and Sarkar,I.N. (2014) Structural network analysis of biological networks for assessment of potential disease model organisms. *J. Biomed. Inform.*, **47**, 178–191.
35. Zhou,X., Menche,J., Barabasi,A.L. and Sharma,A. (2014) Human symptoms-disease network. *Nat. Commun.*, **5**, 4212.
36. Carson,M.B., Liu,C., Lu,Y., Jia,C. and Lu,H. (2017) A disease similarity matrix based on the uniqueness of shared genes. *BMC Med. Genomics*, **10**, 26.
37. Panattoni,L., Hurlimann,L., Wilson,C., Durbin,M. and Tai-Seale,M. (2017) Workflow standardization of a novel team care model to improve chronic care: a quasi-experimental study. *BMC Health Serv. Res.*, **17**, 286.
38. Lusk,K. (2015) A decade of standardization: data integrity as a foundation for trustworthiness of clinical information. *J. AHIMA*, **86**, 54–57.
39. Wilkins,S.A., Shannon,C.N., Brown,S.T., Vance,E.H., Ferguson,D., Gran,K., Crowther,M., Wellons,J.C. 3rd and Johnston,J.M. Jr (2014) Establishment of a multidisciplinary concussion program: impact of standardization on patient care and resource utilization. *J. Neurosurg. Pediatr.*, **13**, 82–89.
40. Pathak,J., Bailey,K.R., Beebe,C.E., Bethard,S., Carrell,D.C., Chen,P.J., Dligach,D., Endle,C.M., Hart,L.A., Haug,P.J. *et al.* (2013) Normalization and standardization of electronic health records for high-throughput phenotyping: the SHARPn consortium. *J. Am. Med. Inform. Assoc.*, **20**, e341–e348.
41. Lu,P., Chen,X., Feng,Y., Zeng,Q., Jiang,C., Zhu,X., Fan,G. and Xue,Z. (2016) Integrated transcriptome analysis of human iPS cells derived from a fragile X syndrome patient during neuronal differentiation. *Sci. China. Life Sci.*, **59**, 1093–1105.
42. Yang,L., Mei,T., Lin,X., Tang,H., Wu,Y., Wang,R., Liu,J., Shah,Z. and Liu,X. (2016) Current approaches to reduce or eliminate mitochondrial DNA mutations. *Sci. China Life Sci.*, **59**, 532–535.
43. Wu,D., Gong,C. and Su,C. (2017) Genome-wide analysis of differential DNA methylation in Silver-Russell syndrome. *Sci. China Life Sci.*, **60**, 692–699.
44. Wang,Y., Gong,C., Wang,X. and Qin,M. (2017) AR mutations in 28 patients with androgen insensitivity syndrome (Prader grade 0–3). *Sci. China Life Sci.*, **60**, 700–706.
45. Cai,S., Wang,X., Zhao,W., Fu,L., Ma,X. and Peng,X. (2017) DICER1 mutations in twelve Chinese patients with pleuropulmonary blastoma. *Sci. China Life Sci.*, **60**, 714–720.
46. Fu,L., Jin,Y., Jia,C., Zhang,J., Tai,J., Li,H., Chen,F., Shi,J., Guo,Y., Ni,X. *et al.* (2017) Detection of FOXO1 break-apart status by fluorescence in situ hybridization in atypical alveolar rhabdomyosarcoma. *Sci. China Life Sci.*, **60**, 721–728.
47. Geng,J., Wang,H., Liu,Y., Tai,J., Jin,Y., Zhang,J., He,L., Fu,L., Qin,H., Song,Y. *et al.* (2017) Correlation between BRAF V600E mutation and clinicopathological features in pediatric papillary thyroid carcinoma. *Sci. China Life Sci.*, **60**, 729–738.
48. Qi,Z., Shen,Y., Fu,Q., Li,W., Yang,W., Xu,W., Chu,P., Zhang,Y. and Wang,H. (2017) Whole-exome sequencing identified compound heterozygous variants in MMKS in a Chinese pedigree with Bardet-Biedl syndrome. *Sci. China Life Sci.*, **60**, 739–745.
49. Fang,F., Liu,Z., Fang,H., Wu,J., Shen,D., Sun,S., Ding,C., Han,T., Wu,Y., Lv,J. *et al.* (2017) The clinical and genetic characteristics in children with mitochondrial disease in China. *Sci. China Life Sci.*, **60**, 746–757.
50. Bai,D., Shi,W., Qi,Z., Li,W., Wei,A., Cui,Y., Li,C. and Li,L. (2017) Clinical feature and waveform in infantile nystagmus syndrome in children with FRMD7 gene mutations. *Sci. China Life Sci.*, **60**, 707–713.
51. Bai,D., Zhao,J., Li,L., Gao,J. and Wang,X. (2017) Analysis of genotypes and phenotypes in Chinese children with tuberous sclerosis complex. *Sci. China Life Sci.*, **60**, 763–771.
52. Li,C., Zhang,J., Li,S., Han,T., Kuang,W., Zhou,Y., Deng,J. and Tan,X. (2017) Gene mutations and clinical phenotypes in Chinese children with Blau syndrome. *Sci. China Life Sci.*, **60**, 758–762.
53. The UniProt, C. (2017) UniProt: the universal protein knowledgebase. *Nucleic Acids Res.*, **45**, D158–D169.
54. Landrum,M.J., Lee,J.M., Benson,M., Brown,G., Chao,C., Chitipiralla,S., Gu,B., Hart,J., Hoffman,D., Hoover,J. *et al.* (2016) ClinVar: public archive of interpretations of clinically relevant variants. *Nucleic Acids Res.*, **44**, D862–D868.
55. Mitchell,J.A. and McCray,A.T. (2003) The Genetics Home Reference: a new NLM consumer health resource. *AMIA Annu. Symp. Proc.*, **2003**, 936.
56. Pinero,J., Bravo,A., Queralt-Rosinach,N., Gutierrez-Sacristan,A., Deu-Pons,J., Centeno,E., Garcia-Garcia,J., Sanz,F. and Furlong,L.I. (2017) DisGeNET: a comprehensive platform integrating information on human disease-associated genes and variants. *Nucleic Acids Res.*, **45**, D833–D839.
57. Pletscher-Frankild,S., Palleja,A., Tsafou,K., Binder,J.X. and Jensen,L.J. (2015) DISEASES: text mining and data integration of disease-gene associations. *Methods*, **74**, 83–89.
58. Wu,X., Jiang,R., Zhang,M.Q. and Li,S. (2008) Network-based global inference of human disease genes. *Mol. Syst. Biol.*, **4**, 189.