Correspondence

DOI: 10.5582/irdr.2024.01043

PytheasDB: An open-access graphical database of clinical data on rare pediatric digestive diseases

Alice Percheron[1], Paul Guerry[2], Alexandre Fabre[1,3,*]

[1] APHM, Timone Children's Hospital, Department of Multidisciplinary Pediatrics, Marseille, France;
[2] Green Grow Scientific, Marseille, France;
[3] Aix Marseille University, INSERM, MMG, Marseille, France.

SUMMARY Advances in genetic testing over the past decades are driving a continuing increase in the diagnosis and reporting of rare genetic diseases, but no tool has yet been developed to aggregate published molecular and phenotypic data, a task that is nevertheless essential to optimize patient care. In this article, we present PytheasDB, an online database of published clinical data from patients with rare digestive diseases. At the time of writing (August 2024), the database contains data from 833 patients with progressive familial intrahepatic cholestasis or trichohepatoenteric syndrome, collected from 172 articles. Users can compare the phenotypic profiles, sex ratios, survival curves, ages at first symptoms, and consanguinity rates of the included diseases. PytheasDB is the first ever online resource providing access to aggregated clinical data from case reports of rare digestive diseases in the literature. The database is currently being expanded to cover ultra-rare pediatric digestive diseases with regular updates to optimize the study and treatment of these diseases.

Keywords PytheasDB, database, rare diseases, PFIC, THE

1. Introduction

The diagnosis of many rare digestive diseases in pediatrics has been revolutionized in recent years by advances in genetic analysis techniques, with next-generation sequencing (NGS) allowing the parallel testing of a panel of genes in a matter of hours. In progressive familial hepatocellular cholestasis (PFIC) for instance, the diagnostic yield of NGS is about 30% (*1*), compared with just a few percent for traditional Sanger sequencing. PFIC is a group of rare disorders, caused by defects of bile secretion or of primary bile acid synthesis. A wide range of genes have been implicated in PFIC since the first genetic cause was identified in *ATP8B1*, in 1998 (*2*), and the NGS panel currently used in clinical practice contains about 50 genes.

As well as being required to establish genotype-phenotype correlations, molecular diagnosis can also inform prognosis and allow personalized treatment. In some cases however, the pathogenicity of identified variants is ambiguous, notably for missense variants, complicating clinical decision making (*3*). The challenge then for clinicians is to analyze molecular results alongside available phenotypic data (clinical, laboratory, radiographic and histological) to make a precise diagnosis and adapt treatments.

This often involves a painstaking literature review because no aggregate resource has yet been developed for these diseases. While the Online Mendelian Inheritance in Man (OMIM) website catalogs published articles by disease, it does not provide classified patient-level clinical data and is not regularly updated. For trichohepatoenteric syndrome (THES) type 1 for instance, the latest of the 13 listed references (as of September 2024) dates from 2018, whereas more than 20 articles have since been published on this topic.

Our recent study of microvillus inclusion disease (*4*) (MVID; *MYO5B* and *STX3* mutations) highlights the value of regrouping published data on rare diseases. Our literature review of published cases (323 patients in 86 articles), clarified the natural history of the disease and revealed a previously overlooked association between *MYO5B* variants and preterm birth, leading to the recommendation that preterm birth and the associated risks should be considered in patient management and may help to better estimate prognosis.

Our aim in developing PytheasDB was therefore to gather, classify and provide access to relevant and regularly updated clinical data from published cases of rare digestive diseases. This resource should save time in clinical practice, facilitate comparisons between diseases and the study of phenotypic patterns and associations,

improving research and patient care.

2. Studied diseases

2.1. Included diseases

This initial version of PytheasDB was built by gathering published data on patients with PFIC or THES, associated with variants in *ATB8B1* (PFIC1), *ABCB11* (PFIC2), *ABCB4* (PFIC3), *TJP2* (PFIC4), *NR1H4* (PFIC5), *SLC51A* (PFIC6), *USP53* (PFIC7), *KIF12* (PFIC8), *ZFYVE19* (PFIC9), *MYO5B* (PFIC10), *SEMA7A* (PFIC11), *SKIC3* (THES1) and/or *SKIC2* (THES2). Searches were conducted in the PubMed database for articles published in English before 1 January 2023, using the gene and associated disease as separate search terms, *i.e.* for PFIC1, either "PFIC 1" or "ATP8B1". Articles that did not report clinical data, literature reviews, and animal studies were excluded. Care was taken to identify patients described in several articles and gather all relevant data while avoiding duplication.

2.2. Studied variables

The studied variables are listed in Supplemental Table S1 (*https://www.irdrjournal.com/action/getSupplementalData.php?ID=215*). These were chosen based on our previous work on rare digestive diseases (*4*), and describe patients' genetic (associated variants) and clinical (symptoms, weight and height growth) characteristics.

3. Website

PytheasDB is accessible at *www.pytheasdb.com*. The website is a client-side application, written in Vue JS, with tables and graphics created on the fly from patient data stored in JSON files. Users can analyze patient data for the included diseases/genes, in the form of tables and figures, or compare data between any number of diseases. The graphical representations include bar and pie charts of reported symptoms at different HPO (Human Phenotype Ontology) (*5*) branch levels and Kaplan–Meier survival curves.

Among the 491 articles identified using search terms in PubMed, 172 matched the inclusion criteria, from which the data for 833 patients were extracted, classified and added to PytheasDB (Supplemental Figure S1, *https://www.irdrjournal.com/action/getSupplementalData.php?ID=215*).

The first version of the website only provides information on a subset of accessible variables, namely the phenotypic profile (reported symptoms), sex ratio, age at first symptoms, survival, and consanguinity. Supplemental Table S2 (*https://www.irdrjournal.com/action/getSupplementalData.php?ID=215*) lists the number of patients and data points included for each disease, along with a completeness score (the proportion of considered characteristics reported in the articles for each group of patients), devised as a semi-quantitative measure of the completeness of patient descriptions in the literature for each gene/disease.

The number of included patients varies from just 1 for PFIC11 (*SEMA7A*) to 159 for PFIC2 (*ABCB11*). Roughly twice as many patients were included for THES1 (*SKIC3*) as for THES2 (*SKIC2*). The number of included data ranges from 5 for PFIC11 to 622 for PFIC11. The completeness of the datasets range from 47.2% for PFIC4 to 100% for PFIC5–9 and PFIC11, indicating that some reports do not include all the basic characteristics (sex, consanguinity, symptoms, age at first symptoms, survival status) considered here.

Examples of PytheasDB outputs are presented in Figures 1-2 and Supplemental Figures S2-S3 (*https://www.irdrjournal.com/action/getSupplementalData.php?ID=215*). Figure 1 presents the list of symptoms reported for PFIC10 patients (*MYO5B*) analyzed at HPO (*5*) branch level 3 (user adjustable) as bar and pie charts, color-coded by HPO branch level 2 category. Figure S2 compares the sex ratio of patients with PFIC1–5 and Figure S3, the corresponding consanguinity rates. Note that the bars are color-coded by disease (and gene), and the outputs are ordered by the chosen category to facilitate comparisons.

Figure 2 compares the Kaplan–Meier survival curves of patients with THES1 and THES2. The survival rates of the 106 THES1 patients are 71% and 68%, at 5 and 10 years respectively, and for the 54 THES2 patients 84%, both at 5 and 10 years, with a most deaths occurring in the first three years of life.

4. Discussion

PytheasDB is the first ever open-access resource providing aggregate information from the literature on patients with rare pediatric digestive diseases. The database continues to grow and will be updated with data from relevant publications. For these rare but increasingly diagnosed diseases indeed, the number of articles in the literature has been growing regularly (Supplemental Figure S4, *https://www.irdrjournal.com/action/getSupplementalData.php?ID=215*). Fabre *et al.*'s 2018 review of THES (*6*), based on roughly 50 patients, is now obsolete, since PytheasDB already contains data on 179 patients. The aim is to provide easy access to exhaustive up-to-date patient information to clinicians, researchers and the general public, simplifying the utilization of medical literature data. Our results for the survival outcomes for THES1 and THES2 patients are similar to those reported by Caralli *et al.* in their 2021 review of THES (*4*).

The main limitation of this approach is that it relies completely on the quality and completeness of patient

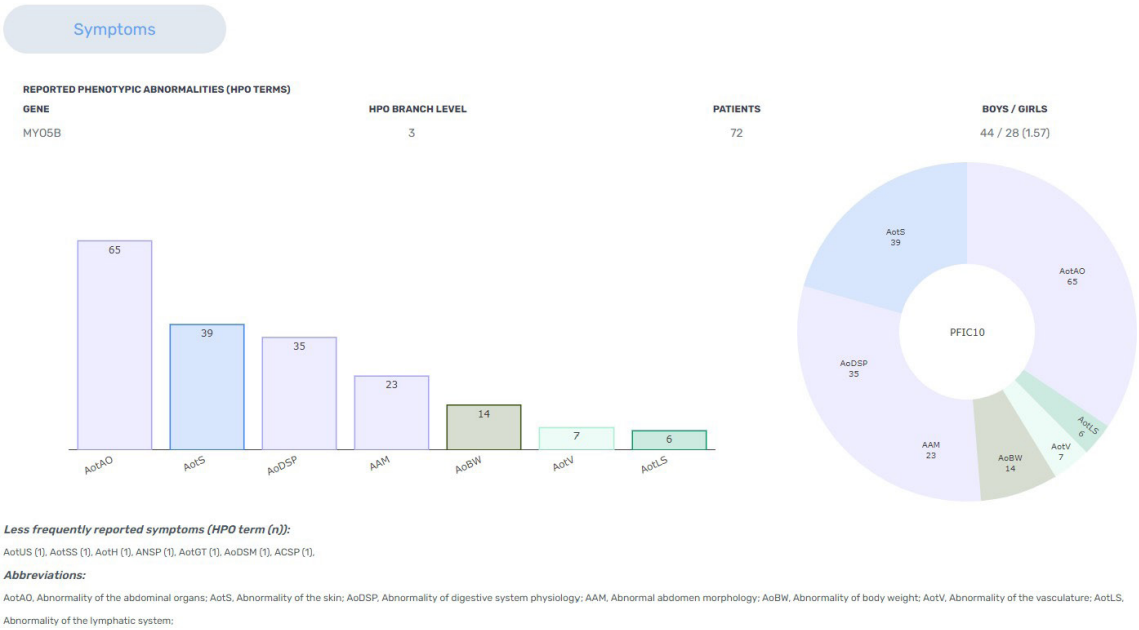

Figure 1. Screenshot of PytheasDB output for PFIC10 symptoms (phenotypic abnormalities reported for PFIC10 patients), classified at HPO (*5*) branch level 3 (user adjustable from 2 to 6). The bars and pie segments are color coded according to the HPO branch level 2 category of the symptom: abnormalities of the digestive system in mauve, abnormalities of the integument in blue, growth abnormalities in olive green; abnormalities of the cardiovascular system in light green, and abnormalities of the immune system in forest green.

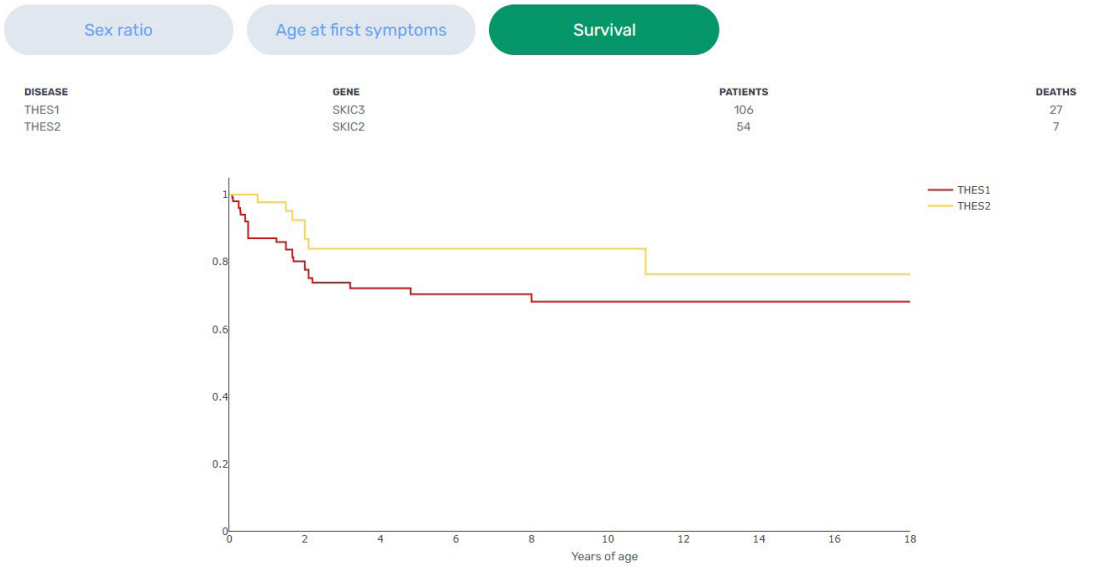

Figure 2. Screenshot of PytheasDB output for the comparison of survival rates between patients with THES1 (red line) and THES2 (yellow).

descriptions in the literature. For the "oldest" diseases for instance, PFIC1–3, first described in the late 1990s, far fewer patients were included than initially expected, as only 79 of the 350 eligible articles were found to report individualized patient characteristics. The absence of individual patient data, as in the results of the ongoing NAPPED study (*7*), reduces the phenotypic resolution of the data. Furthermore, while it is sometimes possible to interpret the non-reporting of a characteristic as absence of this characteristic, this does not generally hold true. In this regard, it is encouraging that according to our estimates of the completeness of data reporting for the different diseases (Supplemental Table S2, *https://www.irdrjournal.com/action/getSupplementalData.php?ID=215*), those that have been described more recently tend to be described in greater detail in the literature. It remains to be seen however whether this reflects an improvement in reporting standards or simply greater diligence in initial case reports.

Another limitation lies in the uniform weighting of patient entries regardless of the level of detail of the descriptions in the original article (*e.g.* subject of an entire paragraph compared with a line in a supplemental table). A more nuanced approach would perhaps be to weight data points according to the number of words devoted to the corresponding patient, thereby assigning greater importance to data from more detailed articles.

Our approach is also limited by the difficulty of distinguishing phenotypes associated with variants in the same gene. For example, mutations in *ATP8B1* and *ABCB11* (PFIC1 and PFIC2) can also cause benign recurrent intrahepatic cholestasis types 1 and 2 (BRIC1 and BRIC2), autosomal recessive disorders at the opposite end of a clinical spectrum ranging from intermittent cholestatic episodes to chronic, progressive cholestasis. It may therefore be interesting to include information on clinical progression to offer a more subtle classification of patients' phenotypes.

In conclusion, this first version of PytheasDB shows that it is possible to aggregate and present clinical data on rare digestive diseases in a way that enriches our understanding of the phenotypic profile and natural history of these disorders and will hopefully contribute to improving patient care. The public availability and regular updating of the data should also stimulate further research. Work is ongoing to expand the database to include additional clinical characteristics, including laboratory findings and treatment types and efficacy, and the long-term aim is to include a range of genes associated with ultra-rare pediatric digestive diseases.

Funding: This work was supported by a grant from MIRUM pharma to complete data collection and create the website. MIRUM pharma was not involved at any stage of the study.

Conflict of Interest: The authors have no conflicts of interest to disclose.

References

1. Almes M, Spraul A, Ruiz M, *et al.* Targeted-capture next-generation sequencing in diagnosis approach of pediatric cholestasis. Diagnostics (Basel). 2022; 12:1169.
2. Bull LN, van Eijk MJ, Pawlikowska L, DeYoung JA, Juijn JA, Liao M, Klomp LW, Lomri N, Berger R, Scharschmidt BF, Knisely AS, Houwen RH, Freimer NB. A gene encoding a P-type ATPase mutated in two forms of hereditary cholestasis. Nat Genet. 1998; 18:219-224.
3. Chen HL, Li HY, Wu JF, Wu SH, Chen HL, Yang YH, Hsu YH, Liou BY, Chang MH, Ni YH. Panel-based next-generation sequencing for the diagnosis of cholestatic genetic liver diseases: Clinical utility and challenges. J Pediatr. 2019; 205:153-159.e6.
4. Caralli M, Roman C, Coste ME, Roquelaure B, Buffat C, Bourgeois P, Badens C, Fabre A. Genetic enteropathies linked to epithelial structural abnormalities and enteroendocrine deficiency: A systematic review. J Pediatr Gastroenterol Nutr. 2021; 72:826-832.
5. Gargano MA, Matentzoglu N, Coleman B, *et al.* The human phenotype ontology in 2024: Phenotypes around the world. Nucleic Acids Res. 2024; 52:D1333-D1346.
6. Fabre A, Bourgeois P, Chaix C, Bertaux K, Goulet O, Badens C. Trichohepatoenteric Syndrome. In: Adam MP, Feldman J, Mirzaa GM, Pagon RA, Wallace SE, Bean LJH, Gripp KW, Amemiya A, editors. GeneReviews®. Seattle (WA): University of Washington, Seattle, 1993-2024.
7. van Wessel DBE, Gonzales E, Hansen BE, Verkade HJ. Defining the natural history of rare genetic liver diseases: Lessons learned from the NAPPED initiative. Eur J Med Genet. 2021; 64:104245.

Received August 20, 2024; Revised September 23, 2024; Accepted October 1, 2024.

**Address correspondence to:*
Alexandre Fabre, Timone Children's Hospital, Multidisciplinary Pediatrics Department, 264 Saint-Pierre Street, 13005 Marseille, France.
E-mail: alexandre.fabre@ap-hm.fr

Released online in J-STAGE as advance publication November 8, 2024.