Curation and Expansion of Human Phenotype Ontology for Defined Groups of Inborn Errors of Immunity

Matthias Haimel PhD [1-3]*, Julia Pazmandi MSc [1-3]*, Raúl Jiménez Heredia MSc [1-3], Jasmin Dmytrus MSc [1-3], Sevgi Köstel Bal M.D.,PhD [1-3], Samaneh Zoghi PhD [1-3], Paul van Daele M.D.[4], Tracy A. Briggs PhD [5,6], Carine Wouters M.D.[7,8], Brigitte Bader-Meunier M.D.[9,10], Florence A. Aeschlimann M.D.[9,10], Roberta Caorsi M.D.[11], Despina Eleftheriou M.D.[12,13], Esther Hoppenreijs M.D.[14], Elisabeth Salzer M.D.,PhD [1-3], Shahrzad Bakhtiar M.D.[15], Beata Derfalvi M.D.[16], Francesco Saettini M.D.[17], Maaike A. A. Kusters M.D.,PhD [12,13], Reem Elfeky M.D.[12,13], Johannes Trück M.D.,Phil [18], Jacques G. Rivière M.D.[19,20], Mirjam van der Burg PhD [21,22], Marco Gattorno M.D.[11], Markus G. Seidel M.D.[23], Siobhan Burns M.D.[24], Klaus Warnatz M.D.[25,26], Fabian Hauck M.D.,PhD [27,28], Paul Brogan M.D.[12,13], Kimberly C. Gilmour PhD [13], Catharina Schuetz M.D.[29], Anna Simon M.D.,PhD [30], Christoph Bock PhD [1,3,31], Sophie Hambleton PhD [32], Esther de Vries M.D., PhD [33,34], Peter Robinson M.D.[35], Marielle van Gijn PhD [36]†#, Kaan Boztug M.D.[1-3,37]†#

* and †, these authors contributed equally

to whom correspondence should be addressed:

Kaan Boztug, Ludwig Boltzmann Institute for Rare and Undiagnosed Diseases (LBI-RUD) and St. Anna Children's Cancer Research Institute (CCRI), Zimmermannplatz 10, A-1090 Vienna, kaan.boztug@rud.lbg.ac.at, Phone: +43 1-40470-4080, Fax: +43-1-40170-7280

Marielle Van Gijn, Department of Genetics, University Medical Center Groningen, Antonius Deusinglaan 1, 9713AV Groningen, Netherlands, m.e.van.gijn@umcg.nl, +31-55256416

[1]Ludwig Boltzmann Institute for Rare and Undiagnosed Diseases, Vienna, Austria

[2]St. Anna Children's Cancer Research Institute, Vienna, Austria

[3]CeMM Research Center for Molecular Medicine of the Austrian Academy of Sciences, Vienna, Austria

[4]Department of Clinical Immunology, Erasmus University Medical Center, Rotterdam, The Netherlands

[5]NW Genomic Laboratory Hub, Manchester Centre for Genomic Medicine, St Mary's Hospital, Manchester University NHS Foundation Trust, Manchester, United Kingdom

[6]Division of Evolution and Genomic Sciences, School of Biological Sciences, University of Manchester, United Kingdom.

[7]Department of Microbiology and Immunology, Immunobiology, KU Leuven, Leuven, Belgium

[8]Department of Pediatrics, Division of Pediatric Rheumatology, University Hospitals Leuven, Leuven, Belgium

[9]Pediatric Immuno-Hematology and Rheumatology Unit, Necker Hospital for Sick Children - AP-HP, Paris, France, EU.

[10]Reference Center for Rheumatic, Autoimmune and Systemic Diseases in Children (RAISE), Paris, France

[11]Center for Autoinflammatory diseases and Immunodeficiency, IRCCS Istituto Giannina Gaslini, Genova, Italy

[12]University College London Great Ormond Street Institute of Child Health, London, United Kingdom

[13]Department of immunology, Great Ormond Street (GOS) Hospital for Children NHS Foundation Trust, London, United Kingdom

[14]Department of Paediatric Rheumatology, Radboud University Medical Centre, Nijmegen, The Netherlands

[15]Department for Children and Adolescents, Division for Stem Cell Transplantation, Immunology and Intensive Care Unit,Goethe University, Frankfurt, Germany

[16]Deptment of Pediatrics, Division of Immunology, Dalhousie University/IWK Health Centre Halifax, Nova Scotia, Canada

[17]Pediatric Hematology Department, Fondazione MBBM, University of Milano Bicocca, via Pergolesi 33, 20900, Monza, Italy

[18]Division of Immunology, University Children's Hospital Zurich, Switzerland

[19]Pediatric Infectious Diseases and Immunodeficiencies Unit, Vall d'Hebron Research Institute, Hospital Universitari Vall d'Hebron, Universitat Autònoma de Barcelona, Barcelona, Spain

[20]Jeffrey Model Foundation Excellence Center, Barcelona, Spain

[21]Department of Immunology, University Medical Center Rotterdam, Rotterdam, The Netherlands

[22]Laboratory for Pediatric Immunology, Department of Pediatrics, Leiden University Medical Center, Leiden, The Netherlands

[23]Research Unit for Pediatric Hematology and Immunology, Division of Pediatric Hemato-Oncology, Department of Pediatrics and Adolescent Medicine, Medical University Graz, Graz, Austria

[24]Department Immunology, UCL Institute of Immunity & Transplantation, Department of immunology, Royal Free Hospital NHS Foundation Trust, Pond Street, London, NW3 2QG, UK

[25]Division of Immunodeficiency, Department of Rheumatology and Clinical Immunology, Medical Center - University of Freiburg, Faculty of Medicine, University of Freiburg, Freiburg, Germany

[26]Center for Chronic Immunodeficiency (CCI), Medical Center - University of Freiburg, Faculty of Medicine, University of Freiburg, Freiburg, Germany

[27]Department of Pediatrics, Dr. von Hauner Children's Hospital, University Hospital, Ludwig-Maximilians-Universität München, Munich, Germany

[28]Munich Centre for Rare Diseases (M-ZSE^{LMU}), University Hospital, Ludwig-Maximilians-Universität München, Munich, Germany

[29]Department of Pediatrics, Medizinische Fakultät Carl Gustav Carus, Technische Universität Dresden, Germany

[30]Radboudumc Expertise Centre for Immunodeficiency and Autoinflammation (REIA), department of Internal Medicine, Radboud University Nijmegen Medical Centre, Nijmegen, The Netherlands

[31]Institute of Artificial Intelligence and Decision Support, Center for Medical Statistics, Informatics, and Intelligent Systems, Medical University of Vienna, Vienna, Austria

[32]Immunity and Inflammation Theme, Translational and Clinical Research Institute, Newcastle University, Newcastle upon Tyne, United Kingdom

[33]Tranzo, Tilburg University, Tilburg, The Netherlands

[34]Laboratory for Medical Microbiology and Immunology, Elisabeth-Tweesteden Hospital, Tilburg, The Netherlands

[35]The Jackson Laboratory for Genomic Medicine, 10 Discovery Drive, Farmington, CT 06032, USA

[36]Department of Genetics,University Medical Center Groningen, Groningen, The Netherlands

[37]Department of Pediatrics and Adolescent Medicine, Medical University of Vienna, Vienna, Austria

Competing interests: The authors declare no conflict of interests.

Funding: The work was supported by the European Research Council (ERC Consolidator Grant 820074 "iDysChart" to K.B. Additional financial support for the workshops was granted by the Ludwig Boltzmann Institute for Rare and Undiagnosed Diseases (LBI-RUD), the European Reference Network on Rare Primary Immunodeficiency, Autoinflammatory and Autoimmune diseases (ERN-RITA), and the European Society for Immunodeficiencies (ESID).

Author contributions: MH, JP, MVG, KB: study design and manuscript writing. MH, JP: data acquisition, coordination of working groups. MH, JP analysis and interpretation of data. SH, clinical cohort data extraction. KB, MVG: Study supervision. All co-authors participated in the meetings and revision of terms. The manuscript was reviewed, edited and approved by all co-authors.

Abstract

BACKGROUND: Accurate, detailed and standardized phenotypic descriptions are essential to support diagnostic interpretation of genetic variants and to discover new diseases. The Human Phenotype Ontology (HPO), extensively used in rare disease research, provides a rich collection of vocabulary with standardized phenotypic descriptions in a hierarchical structure. However, to date the use of HPO has not yet been widely implemented in the field of inborn errors of immunity (IEIs), mainly due to a lack of comprehensive IEI-related terms.

OBJECTIVES: We sought to systematically review available terms in HPO for the depiction of IEIs, to expand HPO yielding more comprehensive sets of terms, and to reannotate IEIs with HPO terms to provide accurate, standardized phenotypic descriptions.

METHODS: We initiated a collaboration involving expert clinicians, geneticists, researchers working on IEIs and bioinformaticians. Multiple branches of the HPO tree were restructured and extended based on expert review. Our ontology-guided machine learning coupled with a two-tier expert review was applied to reannotate defined subgroups of IEIs.

RESULTS: We revised and expanded four main branches of the HPO tree. Here, we reannotated 73 diseases from four IUIS-defined IEI disease subgroups with HPO terms. We achieved a 4.7-fold increase in number of phenotypic terms per disease. Given the new HPO annotations, we demonstrated improved ability to computationally match selected IEI cases to their known diagnosis, and improved phenotype-driven disease classification.

CONCLUSION: Our targeted expansion and reannotation presents enhanced precision of disease annotation, will enable superior HPO-based IEI characterization and hence benefit both IEI diagnostic and research activities.

Key message

HPO is a robust resource for supporting IEI diagnostics and genetics with adequate ontology breadth and disease annotation depth.

Capsule Summary

Our newly formed expert consortium systematically reviewed and expanded existing HPO terms of IEIs and reannotated IEIs with HPO terms. This will support diagnostic pipelines and analysis of variants from next-generation sequencing.

Key words

HPO; ontology; phenotype; rare diseases; inborn errors of immunity; immune deficiencies; disease classification; diagnostic support; patient matching; genetic analysis

Abbreviations

ALPS - Autoimmune Lymphoproliferative Syndrome

CVID – Common Variable Immunodeficiency Disorders

EBV – Epstein-Barr Virus

EHR - Electronic Health Record

ERN-RITA - European Reference Network on Rare Primary Immunodeficiency; Autoinflammatory and Autoimmune diseases

ESID - European Society for Immunodeficiencies

HLH - Hemophagocytic Lymphohistiocytosis

HPO - Human Phenotype Ontology

IEI - Inborn Errors of Immunity

IUIS – International Union of Immunological Societies

ISSAID - International Society of Systemic Autoinflammatory Diseases

LBI-RUD - Ludwig Boltzmann Institute for Rare and Undiagnosed Diseases

OMIM – Online Mendelian Inheritance in Man

PAD – Primary Antibody Deficiencies

SCID – Severe Combined Immunodeficiency

TRAPS - Tumor necrosis factor receptor-associated periodic syndrome

UDNI - Undiagnosed Diseases Network International

UDP and UDN - Undiagnosed Disease Program and Network

Introduction

Rare and undiagnosed diseases pose challenges for affected patients, clinicians and researchers working to improve diagnostic and therapeutic approaches. Because of the rarity, clinicians often only see a few patients with specific rare phenotypes throughout their careers, leading to considerable diagnostic delay (*1*). Genetic research on rare diseases often relies on single pedigrees or a few patients, leaving many patients undiagnosed (*1*). Compiling a cohort of patients - so-called patient matching - is often crucial to gain insight into the phenotypic spectrum, natural/clinical history of the disease, and adequate monitoring and treatment strategies. The rare disease community has recognized these challenges and established tools enabling efficient data sharing across institutions and borders, including genetic data exchange through the Matchmaker Exchange platform (*2*) to solve undiagnosed exomes and genomes (*3*). These platforms however are highly dependent on accurately phenotyped and categorized patients and standardized disease classifications.

To date, several nomenclatures and reference systems for diseases have been developed (*4,5*). In parallel, ontologies were established to provide a more systematic, hierarchical classification of diseases (*6,7*). However, these nomenclatures group patients by disease label and do not describe the underlying phenotypic features. Consequently, clinical features, laboratory measurements, anatomical and functional phenotypes of patients are often described with variable quality and specificity, which hampers patient matching, diagnostic efficiency, genetic variant prioritization in diagnostic pipelines and global data exchange.

Given these challenges and the need for accurate, standardized phenotyping, the Human Phenotype Ontology (HPO) system was conceptualized and published with initial terminology in 2008 (*8,9*). To date, HPO provides the most comprehensive deep phenotyping resource for rare diseases for clinicians, researchers, bioinformaticians and electronic health record (EHR)

196 systems in the world. HPO is used in many projects including the 100,000 Genomes Project,
197 the NIH Undiagnosed Disease Program and Network (UDP and UDN), the Undiagnosed
198 Diseases Network International (UDNI), RD-CONNECT, and SOLVE-RD (*1,10-13*). HPO is
199 a community-based tool and is increasingly adapted as the standard to describe phenotypic
200 abnormalities for everyday use (*14*). Each term in HPO describes a distinct phenotypic feature
201 (e.g. lymphadenopathy, HP:0002716) and the HPO tree structure allows similarity measures
202 between patient phenotypes. HPO contains over 200,000 phenotypic annotations for hereditary
203 diseases, of which 2,120 are considered rare diseases. Inborn errors of immunity (IEIs) form a
204 subgroup of these rare diseases. Clinical experts in IEI agree that a major barrier to the adoption
205 of HPO terminology has not been used widely for IEIs partly due to the lack of disease specific
206 HPO terms for IEI patients *(15)*. Adequate depiction of the complex clinical and
207 immunological phenotypes of IEI disease entities with HPO terms would allow discrimination
208 between heterogeneous groups of IEIs. Illustrating the lack of terms, in 2017 HPO contained
209 more than 11,000 terms, out of which 5,000 terms have been applied to the musculoskeletal
210 system, with only 1,000 terms related to IEIs (*9,15*). In addition, the phenotypic annotation of
211 IEIs often includes results of specific immunological assays, which pose a challenge to
212 accurately reflect in HPO terms (*15*). Because of the lack of specific HPO terms depicting
213 results of laboratory assays, often a non-specific broader term is used for the annotation of IEIs.
214 Therefore, HPOs are currently not specific enough to be used for genetic analysis and
215 diagnostic aid for IEIs. In a study addressing the clinical efficacy of genetic testing in IEI,
216 bioinformatics tools using existing HPO terms missed the disease causing gene in 37% of the
217 patients with known monogenic disorders *(16)*. In this study, we set out to improve HPO
218 terminology for IEIs by applying established bioinformatic methodologies coupled with expert
219 review. The aims of this project were therefore to i) systematically review existing HPO terms
220 for IEIs, ii) revise ontology structures, to iii) add missing terms, as well as iv) reannotate

existing IEIs with HPO terms, to collectively enable systematic use of HPO by the IEI-community.

Materials and Methods

Spearheaded by the European Reference Network on Rare Primary Immunodeficiency, Autoinflammatory and Autoimmune diseases (ERN-RITA) and the European Society for Immunodeficiencies (ESID), we set up working groups comprising members of the participating immunodeficiency societies to revise and expand HPO terms for IEIs. Three workshops, numerous teleconferences and joint task forces took place over the span of 2 years, with over 30 participants including expert clinicians, geneticists, researchers working on IEIs and bioinformaticians. All participating clinicians and geneticists identified through ERN-RITA, ESID, and the International Society of Systemic Autoinflammatory Diseases (ISSAID) are established experts in their fields from different European countries and North America. Additional scientific support provided the indispensable bioinformatics expertise.

Establishment of working structure

A remote working structure (detailed in the Supplementary Methods) was launched to address gaps in the HPO tree and in the annotation of IEI diseases.

Expansion and restructuring of disease-related branches of the HPO tree

Disease-specific HPO restructuring was discussed within four working groups. Each group focused on a different HPO branch; the suggested changes were agreed on among all participants. Differences between centers and countries in the use of terms and definitions were highlighted during the face-to-face workshops. The results were summarized electronically in Excel documents or pictures and flipchart drawings by the main coordinators before being submitted to HPO. The full list of restructured tree elements is detailed in the Supplementary Document 1. New submitted HPO terms can be found in Supplementary Document 2. Additionally, missing terms describing pulmonary and gastro-intestinal complications of

248 primary antibody deficiency (PAD) were discussed during teleconferences and thereafter
249 submitted to update the HPO ontology.
250
251 *Standardized reannotation of rare, genetically diagnosed diseases*
252 A four-step process was developed for a standardized reannotation effort across working
253 groups and to consistently annotate IEIs (spanning over 300 different diseases in Online
254 Mendelian Inheritance in Man (OMIM)) with HPO terms (Fig 1). As IEIs represent a large and
255 heterogenous group of rare diseases, we here decided to selectively focus on defined subgroups
256 of IEI to test the feasibility and usefulness of such an endeavor. First, publications were
257 collected by experts for each disease within the subgroups (minimum of two articles per
258 disease), representing key phenotypic presentation(s) of the specific disease. In the second step,
259 HPO terms were extracted from the provided publications for each disease using machine
260 learning ((*17*), explained in detail in Supplementary Materials and Methods) and summarized
261 into Excel documents. Third, a two-tier expert review evaluated the text mined terms,
262 suggested additional terms if required and the responsible working group agreed (defined as at
263 least 80% agreement amongst group experts) on the final HPO annotations for each disease.
264 Fourth, the validated terms were submitted to HPO. Supplementary Document 2 contains the
265 reannotated diseases and the list of reannotated terms for each disease is available in
266 Supplementary Document 3.
267
268 *Standardized reannotation of genetically undiagnosed diseases*
269 The methods above were specifically designed for application in (very) rare diseases, where
270 the number of patients and therefore the described phenotypic spectrum and clinical
271 presentation is sparse. In case of diseases and disease groups where an adequate amount of
272 patient and phenotype data was available, in addition to a True/False annotation, the frequency

of each phenotypic item was assessed. The frequencies correspond to the following representation in patients: common = Frequent (79-30%); sometimes = Occasional (29-5%); rare = Very rare (<4-1%).

Patient cohort

We randomly selected 30 patients that harbored a genetic diagnosis in one of the reannotated diseases from a large pediatric referral center research database. Clinical summaries of these patients prior to genetic diagnosis were retrieved by an expert clinician. The clinical summaries were parsed and HPO terms were extracted using machine learning as in the Supplementary Methods.

HPO information content measures, and disease patient similarity measures

Information content of all HPO terms was assessed with the *R* package ontologyIndex v2.5 (*18*). The phenotypic similarity of diseases and patients before and after reannotation was compared using the *R* package ontologySimilarity v2.3 (*18*). The Euclidean distances between the diseases were computed based on similarity measures, clustered with hierarchical clustering and visualized with ggtree using the *R* packages ggtree (*19*) and ape v5.2 (*20*).

A detailed description including the data processing pipeline and tools are available in the Supplementary Materials and Methods.

Supplementary Materials

Supplementary Materials and Methods

Supplementary Document 1: HPO tree restructuring and list of new terms

Supplementary Document 2: Summary of diseases reannotated

Supplementary Document 3: List of all terms per disease after reannotation

Supplementary Document 4: List of cases used for phenotype to diagnosis matching

Results

Systematic evaluation and expansion of the HPO structure and terms relevant to IEIs

Our approach has resulted in the restructuring of four main branches of the HPO tree, namely: i) abnormality of the immune system (HP:0002715) ii) abnormality of metabolism/homeostasis (HP:0001939) iii) abnormality of the integument (HP:0001574) and iv) abnormality of the cardiovascular system. (Fig 2A, Supplementary Document 1). Together, this revision prompted the replacement/restructuring of 67 terms, and the addition of 57 new terms to the HPO tree, among them "recurrent fever", "unusual infections", "IgG levels in blood" (Fig 2B, comprehensive list in Supplementary Documents 1 and 2).

Directed expansion of primary antibody deficiency (PAD) terms

Overall, the PAD working group focused on replacing broad and non-specific terms with terms that describe phenotypes in more detail and accuracy (example: 'partially absent total IgG/IgA/IgM in blood' and '(near) absent total IgG/IgA/IgM in blood' instead of 'hypogammaglobulinemia') Fig 2B. In addition, we proposed that the full detailed spectrum of specific antibody as well as IgG-subclass deficiencies was described by separate HPO terms. For example, we described individual terms related to 'decreased specific antibody response to vaccination in blood' divided according to the response to different types of vaccination (protein, protein-conjugated polysaccharide and unconjugated polysaccharide).

Standardized reannotation of rare, genetically diagnosed IEIs

We started by a systematic review of four disease categories of the IUIS classification of IEIs, as proof of concept: diseases affecting cellular and humoral immunity (IUIS Table 1), diseases of immune dysregulation (IUIS Table 4), autoinflammatory disorders (IUIS Table 7) and

325 genetically undiagnosed predominantly antibody deficiencies (IUIS Table 3), detailed in Table
326 1 and Supplementary Document 3. As a first step, we assessed the already available HPO
327 annotation for each disease in the 2019-06-03 HPO release. We found that 15% of diseases
328 considered (11 of 73 diseases in total) did not have any associated HPO terms (Fig 3A). Overall,
329 we found that on average 13.3 phenotype terms were available per disease (Fig 3B), later
330 referred to as “existing terms”.
331 The text mining and evaluation process was separated into four steps shown in Fig 3C. We
332 have first focused the reannotation of 72 genetically diagnosed IEIs, and genetically
333 undiagnosed PADs. For genetically diagnosed IEIs, text mining was based on 162 expert-
334 curated articles, on average 2.57 articles per disease (Fig 3D). This resulted in 4,517 extracted
335 phenotype terms, 66.42 terms per disease (Fig 3E). Of these terms, 3,242 - or 71% per disease
336 (47.67 out of 66.42) - were accepted as correctly attributed terms by the expert reviewers (Fig
337 3F). Expert suggestions added up to 529 additional HPO terms, in addition to the existing and
338 text mined terms.
339 After reannotation, a mean of 63.1 terms were available for each disease, resulting in a 4.7-fold
340 gain in the number of available annotations (Fig 3G). The mean information content as
341 measured by the overall frequency of terms in each disease’s annotations has increased from
342 6.17 to 8.3 (Fig 3H) after reannotation.
343
344 The new annotation of diseases consisted mainly of text mined terms (70.6%) (Fig 3I),
345 followed by already existing terms (9.3%) and additional suggestions by experts (9.3%, adding
346 a further 5.2 additional terms per disease) (Supplementary Document 3).
347

Standardized reannotation of genetically undiagnosed primary antibody deficiencies (PADs)

PADs form a heterogeneous group, and the majority of PADs do not (as yet) have a genetic diagnosis. We collected articles describing the heterogeneous PADs related to common variable immunodeficiency disorders (CVID), agammaglobulinemia, selective IgM deficiency, selective IgA deficiency, IgG-subclass deficiency, specific antibody deficiency and unclassified antibody deficiency subgroups. In total, 541 terms were text mined from these articles, many of these in more than one PAD subgroup, and 245 of these terms (45.2%) were annotated as correctly associated to the respective PAD subgroup by the expert reviewers (Fig 3J). Of these 245 terms, the experts annotated 16.3% as commonly found in PAD diseases, 48.97% as sometimes associated (albeit less commonly), and 34.7% as rarely associated with PAD (Fig 3K).

Patient-disease matching

We set out to showcase the efficacy of our reannotation effort by highlighting the potential diagnostic impact of optimized disease annotation. To do this, we have selected 30 clinical cases from a large immunology referral center research database (Supplementary Document 4). HPO terms were matched to patient phenotypes by experts from the clinical synopsis and the phenotypic similarity to all HPO-annotated diseases was calculated based on these selected patient HPO terms (Fig 4A), as illustrated by one concrete clinical example of a patient with Tumor Necrosis Factor Receptor Associated Periodic Syndrome (TRAPS, Fig 4B). Overall, we show a significant 47% improvement in the specificity of patient phenotype matching to correct diagnosis (from 0.49 to 0.72, p value = 1.8e-07, Fig 4C), and a significantly better ranking of the correct clinical diagnosis across all possible diseases after reannotation: in the majority of cases, the correct diagnosis was in the top 10 of matched diseases (Fig 4D) after reannotation, and the rank of the correct diagnosis for individual patients was highly

significantly improved, from a mean of 285 to 19 (14.9 fold improvement, p value = 9.1e-07, Fig 4E).

Phenotype-driven disease classification

We tested the efficacy of our approach in selecting biologically and clinically meaningful phenotypes by assessing the HPO-ontology based phenotypic similarity of diseases before and after reannotation. In particular, we assessed whether the similarity was greater within or between IUIS clinically defined groups. We found that the phenotype-driven disease classification after reannotation has resulted in a clustering more in concordance with the IUIS-based clinical classification (Fig 5A-B).

Discussion

Unified data standards, consistent classification and robustly verified clinical data are vital pillars supporting diagnostic pipelines and data-driven research. Although databases and vocabularies that aim to provide accurate phenotypic descriptions exist (*5-9*), there are still major gaps in the depiction of IEIs in these datasets. Here we used a cross-community collaboration to review, expand and improve the depiction of IEIs in HPO, and reannotate IEIs with HPO terms. We reviewed four separate branches of the HPO tree and submitted 57 new and expanded HPO terms, the majority of which are now included in the official HPO dataset. We introduced a semi-automated reannotation pipeline, that combines ontology-guided machine learning and a two-tier expert review to reannotate four main categories of IEIs. The basis of the ontology-guided machine learning was the expert curated list of articles (162 in total), that was submitted to the PanelApp (*21*) to serve as a public resource. The text mined phenotypes were subjected to expert review to confer face validity or refute the putative new HPO terms. IEIs and their current HPO terms covered by the working groups were scrutinized in-depth, resulting in high-quality annotations. Overall, we have achieved a 4.7-fold gain in number of HPO terms annotating each disease. These annotations included unspecific (frequently annotated) as well as specific (less frequently annotated) HPO terms holding less and more information content respectively. Combined, the mean information content increased from 6.17 to 8.3.

Each reannotated disease showed an increase in information content and a quantitative gain in the number of available HPO terms. Through patient-disease matching and disease-similarity examples we illustrated that these gains and increases translated to significant qualitative improvement in patient-disease matching in an independent cohort of IEI patients (Figure 4), and phenotype-driven classification of IEIs that more closely resembles clinical consensus

(Figure 5). Although neither of these measures are systematic assessments of global patient-disease matching and disease similarity comparisons, they highlight that there is considerable benefit by the revision of specific subclasses of diseases. Once a near complete HPO phenotype reannotation of almost all IEIs is available, it will be intriguing to assess how well patients with genetic diagnoses match reannotated OMIM diseases in a clinical setting, how patient matching to genetic diagnosis is transformed, and if these changes ultimately lead to an earlier diagnosis. Finally, once a detailed and accurate phenotypic description is available for all IEIs, identification phenotype-driven patient subgroups will be common practice, and a more objective entirely phenotype-driven classification and ontology of IEIs can become a reality.

Accurate phenotypic description of patients holds promise for diagnostic utility and for the discovery of novel diseases. Phenotype-driven genetic diagnostic tools now exist, but their full clinical potential is hampered by the lack of complete phenotypic descriptions for most types of IEIs. Phenotips (*22*) is a free and open source software for collecting and analyzing phenotypic information of patients with genetic disorders that is widely used in the rare disease community. Tools such as Exomiser use HPO terms to annotate and to prioritize potentially casual variants (*23*). New integrative ‘omics approaches and the analysis of large-scale data with artificial intelligence will allow us to go from a one-size-fits-all to a more personalized medicine, including in IEIs. We see the potential to integrate the richer phenotyping of previously undiagnosed groups of IEI patients with available sequencing data to accelerate disease gene discovery and at the same time increase the diagnostic rate in new patients (*24*). Novel disease-gene or phenotype associations depends on sufficient numbers of cases as well as a control cohort of comparable quality. Cross-institute and cross-country collaborations for cohorts of undiagnosed, but well-phenotyped patients could shed light on novel disease-causing genes not only of the immune system. Trusted and accepted data and information

sharing platforms are already being developed (*13, 22*) to provide robust and sufficiently granular HPO terms as a standardized way of phenotyping patients. Electronic health records (EHR) (*25*) could facilitate the transfer of HPO terms by integrating with available sharing platforms. Capturing HPO annotations of novel rare diseases or cases is an ongoing challenge for a complete disease representation. Thus it is important that alongside of updating the official IUIS classification, HPO descriptions of disorders are curated once every several years. We suggest a community effort for such regular reviews of HPO regarding IEIs, such as a team of experts, part of big international groups of clinicians such as ESID or ERN RITA, the Clinical Immunology Society (CIS) or other similar organizations. Publication standards that require the submission of HPO annotations up-front would greatly improve this process.

Once phenotyped patients are available, robust and global approaches are accessible (*2*) to find phenotypic similar cases. These comparisons are performed by advanced machine learning algorithms. However, machine learning can also be a very powerful tool to automate the identification of relevant phenotype information in publications or clinical notes. We applied an ontology-guided machine learning tool to support the annotation of diseases and explored the full spectrum of terms – from very relevant to not relevant at all. The same process can be applied to unstructured clinical notes to accelerate in-depth annotation of patients. For patients with EHR (*25*), abnormal clinical values can automatically be translated into HPO codes (*26*) for a more precise diagnostic application and integrated with sharing platforms as mentioned before. The foundation of these comparisons is an ontology with a comprehensive set of term, which is widely used.

As there is currently no gold-standard on how to perform an expert-based review of ontologies. guidance on annotating diseases with HPO phenotypes can vary between diseases, disease

458 classes and centers. IEIs are rare diseases, and often there are only a few patients described
459 (sometimes only one kindred in case of ultra-rare diseases). Therefore, the depth of currently
460 available published phenotypes is at times limited. The low number of patients and insufficient
461 depth of available phenotypes brings up a question as to which diseases to include in
462 phenotyping exercises of this nature. On the one hand, focusing on IEIs that are commonly
463 accepted, with multiple patients diagnosed and well described by multiple researchers can
464 increase the depth of phenotyping. However, this approach excludes at least 10% of IEIs (the
465 ultra-rare diseases). On the other hand, an all-inclusive approach including every disease
466 systematically means that we rely on sparsely phenotyped patients and perhaps insufficient
467 data for ultra-rare disorders. A warning of accuracy by indicating the frequency of each
468 phenotype for diseases could soon be possible, with the addition of phenotype frequency to the
469 HPO dataset, an expansion that is currently work in progress. This implies the need for a
470 responsive system, capable of assimilating new phenotypic information as the pool of
471 confidently diagnosed patients increases.
472
473 Our ongoing approach aims to address these gaps for IEIs and to provide an ontology that is
474 practical, useful and as complete as possible. However, the existence of a well-built ontology
475 and the awareness of clinicians and researchers itself does not guarantee a shift in the
476 community to fully adapt a standardized phenotyping approach. Our approach raised awareness
477 regarding the concept and importance of HPO amongst the IEI community. Moreover, the
478 process made the participating clinicians aware of the available terms and highlighted where
479 these were lacking. Moving forward, it is very important that official entities adopt HPO terms
480 as the unified means of patient phenotyping. We hypothesize that as soon as the widely used
481 registries such as the Undiagnosed Disease Network (*11*) or the IUIS (*27*) use HPO to refer to
482 phenotypic annotation, this will propel the IEI field towards adopting HPO as the main

483 nomenclature for phenotyping IEI patients. One promising move in this direction is the recent
484 expansion of the ESID registry working definitions for the clinical diagnosis of IEIs (*28*), which
485 derives HPO terms from OrphaNet using the ORDO Ontological Module (HOOM) platform
486 (*29*), prompted by our HPO initiative.
487
488 In summary, our work reviewed and expanded the phenotypic depiction of multiple subclasses
489 of IEIs, and to our knowledge, this initiative is the first endeavor of its kind with the aim of
490 standardizing IEI phenotypes. Our semi-automated annotation-based approach is scalable to
491 include all IEIs as illustrated herein. We propose our reannotation approach as a blueprint for
492 systematic HPO (re)annotation for additional immunological and non-immunological diseases.
493
494
495

Fig 1: Pipeline for of standardized reannotation of IEI diseases. First, scientific publications were collected by experts for each disease within the subgroups. Second, HPO terms were extracted from the provided publications for each disease using machine learning and summarized into Excel documents. Third, a two-tier expert review evaluated the text mined terms, suggested additional terms if required and the responsible working group agreed on the final HPO annotations for each disease. Fourth, data were collated, and the agreed terms were submitted to HPO.

Fig 2: Revision and expansion of the HPO tree. A) Schematic representation of the restructuring of the HPO tree. Main branches of the HPO tree where restructuring was performed are marked with light green. B) "Abnormality of temperature", "Abnormality of immunoglobulin level" and "Unusual infections" as examples of revised branches of the HPO tree. New additions to the tree are marked with green, repositioned terms are marked with yellow.

Fig 3: Result of disease reannotation. A) HPO annotation availability in the subset of 72 diseases. B) Distribution of number of available HPO terms per disease. C) Distribution of the number of articles used per disease for the reannotation pipeline. D) Number of mined terms per disease. Each dot represents a disease. E) All mined vs all accepted terms. F) Number of available terms per disease before and after reannotation. Each dot represents a disease. G) Mean information content available per disease before and after reannotation. H) The aggregate mean annotation per disease after reannotation. I) All text mined terms from PAD publications J) Frequency distribution of different PAD terms according to the experts. HPO: Human Phenotype Ontology; PAD: Primary Antibody Deficiencies.

Fig 4: Patient-disease matching. A) Schematic overview of the different steps of patient-to-disease matching. First, the phenotypes were identified in a patient's clinical history. Second, these phenotypes were translated to HPO terms. Finally, patient phenotype to disease matching was measured by Resnik similarity. B) Matching patient 1 to a diagnosis. C) Similarity of patients in patient cohort to genetic diagnosis before and after reannotation. D) The rank of correct clinical diagnosis more often is in the top 10 of matched diseases after reannotation. E) Improvement of ranks of clinical diagnosis before and after reannotation. Significance was assessed by Student t-test.

Fig 5: Phenotypic similarity of diseases before and after reannotation. Diseases are annotated with the IUIS disease group (inner circle), sub-group (outer circle) and OMIM identifier. A) Clustering of diseases based on phenotypic similarity before reannotation. B) Clustering of diseases based on phenotypic similarity after reannotation. HPO: Human Phenotype Ontology; IUIS: International Union of Immunological Societies, OMIM: Online Mendelian Inheritance in Men; IEI: Inborn Errors of Immunity; EBV: Epstein-Barr Virus

References

1. Gahl WA, Markello TC, Toro C, Fajardo KF, Sincan M, Gill F, et al. The National Institutes of Health Undiagnosed Diseases Program: insights into rare diseases. *Genet Med*. 2012 Jan;14(1):51–9.
2. Philippakis AA, Azzariti DR, Beltran S, Brookes AJ, Brownstein CA, Brudno M, et al. The Matchmaker Exchange: a platform for rare disease gene discovery. *Hum Mutat*. 2015 Oct;36(10):915–21.
3. Sobreira N, Schiettecatte F, Valle D, Hamosh A. GeneMatcher: a matching tool for connecting investigators with an interest in the same gene. *Hum Mutat*. 2015 Oct;36(10):928–30.
4. Hernandez-Ibarburu G, Perez-Rey D, Alonso-Oset E, Alonso-Calvo R, de Schepper K, Meloni L, et al. ICD-10-CM extension with ICD-9 diagnosis codes to support integrated access to clinical legacy data. *Int J Med Inform*. 2019;129:189–97.
5. Amberger J, Bocchini C, Hamosh A. A new face and new challenges for Online Mendelian Inheritance in Man (OMIM®). *Hum Mutat*. 2011 May;32(5):564–7.
6. Pavan S, Rommel K, Mateo Marquina ME, Höhn S, Lanneau V, Rath A. Clinical Practice Guidelines for Rare Diseases: The Orphanet Database. *PLoS ONE*. 2017;12(1):e0170365.
7. Schriml LM, Mitraka E, Munro J, Tauber B, Schor M, Nickle L, et al. Human Disease Ontology 2018 update: classification, content and workflow expansion. *Nucleic Acids Res*. 2019 08;47(D1):D955–62.
8. Robinson PN, Köhler S, Bauer S, Seelow D, Horn D, Mundlos S. The Human Phenotype Ontology: a tool for annotating and analyzing human hereditary disease. *Am J Hum Genet*. 2008 Nov;83(5):610–5.

562 9. Köhler S, Vasilevsky NA, Engelstad M, Foster E, McMurry J, Aymé S, et al. The
563 Human Phenotype Ontology in 2017. *Nucleic Acids Res*. 2017 04;45(D1):D865–76.
564 10. Ramoni RB, Mulvihill JJ, Adams DR, Allard P, Ashley EA, Bernstein JA, et al. The
565 Undiagnosed Diseases Network: Accelerating Discovery about Health and Disease.
566 *Am J Hum Genet*. 2017 02;100(2):185–92.
567 11. Taruscio D, Groft SC, Cederroth H, Melegh B, Lasko P, Kosaki K, et al. Undiagnosed
568 Diseases Network International (UDNI): White paper for global actions to meet patient
569 needs. *Mol Genet Metab*. 2015 Dec;116(4):223–5.
570 12. Gall T, Valkanas E, Bello C, Markello T, Adams C, Bone WP, et al. Defining Disease,
571 Diagnosis, and Translational Medicine within a Homeostatic Perturbation Paradigm:
572 The National Institutes of Health Undiagnosed Diseases Program Experience. *Front*
573 *Med (Lausanne)*. 2017;4:62.
574 13. Thompson R, Johnston L, Taruscio D, Monaco L, Béroud C, Gut IG, et al. RD-
575 Connect: an integrated platform connecting databases, registries, biobanks and clinical
576 bioinformatics for rare disease research. *J Gen Intern Med*. 2014 Aug;29 Suppl
577 3:S780-787.
578 14. Köhler S, Carmody L, Vasilevsky N, Jacobsen JOB, Danis D, Gourdine J-P, et al.
579 Expansion of the Human Phenotype Ontology (HPO) knowledge base and resources.
580 *Nucleic Acids Research*. 2019 Jan 8;47(D1):D1018–27.
581 15. Chinn IK, Chan AY, Chen K, Chou J, Dorsey MJ, Hajjar J, et al. Diagnostic
582 interpretation of genetic studies in patients with primary immunodeficiency diseases:
583 A working group report of the Primary Immunodeficiency Diseases Committee of the
584 American Academy of Allergy, Asthma & Immunology. *Journal of Allergy and*
585 *Clinical Immunology*. 2020 Jan;145(1):46–69.

16. Rae W, Ward D, Mattocks C, Pengelly RJ, Eren E, Patel SV, et al. Clinical efficacy of a next-generation sequencing gene panel for primary immunodeficiency diagnostics. *Clin Genet*. 2018 Mar;93(3):647–55.

17. Arbabi A, Adams DR, Fidler S, Brudno M. Identifying Clinical Terms in Medical Text Using Ontology-Guided Machine Learning. *JMIR Med Inform*. 2019 May 10;7(2):e12596.

18. Greene D, Richardson S, Turro E. ontologyX: a suite of R packages for working with ontological data. *Bioinformatics*. 2017 01;33(7):1104–6.

19. Yu G, Lam TT-Y, Zhu H, Guan Y. Two Methods for Mapping and Visualizing Associated Data on Phylogeny Using Ggtree. *Mol Biol Evol*. 2018 01;35(12):3041–3.

20. Paradis E, Schliep K. ape 5.0: an environment for modern phylogenetics and evolutionary analyses in R. *Bioinformatics*. 2019 Feb 1;35(3):526–8.

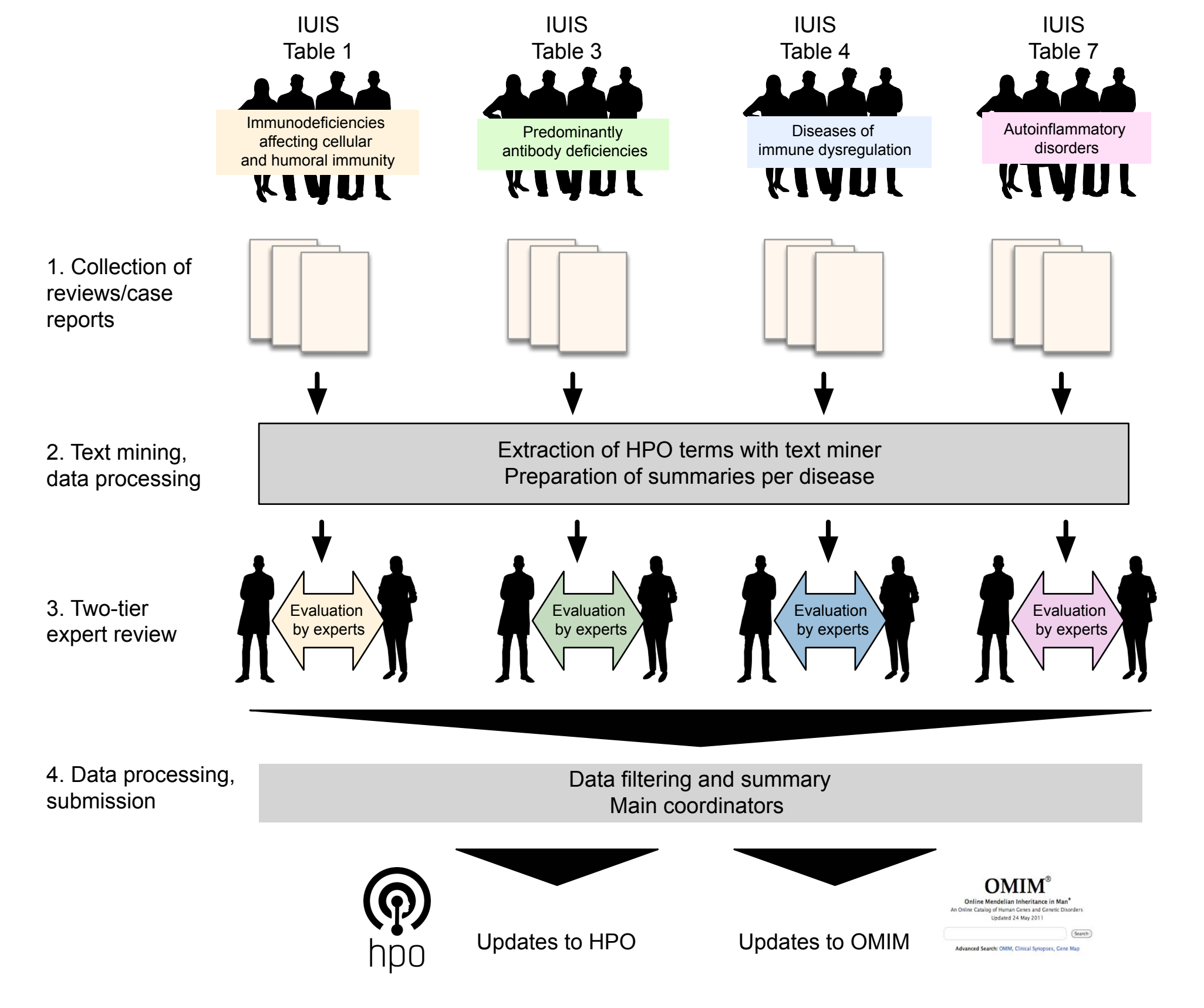
21. Martin AR, Williams E, Foulger RE, Leigh S, Daugherty LC, Niblock O, et al. PanelApp crowdsources expert knowledge to establish consensus diagnostic gene panels. *Nat Genet*. 2019;51(11):1560–5.

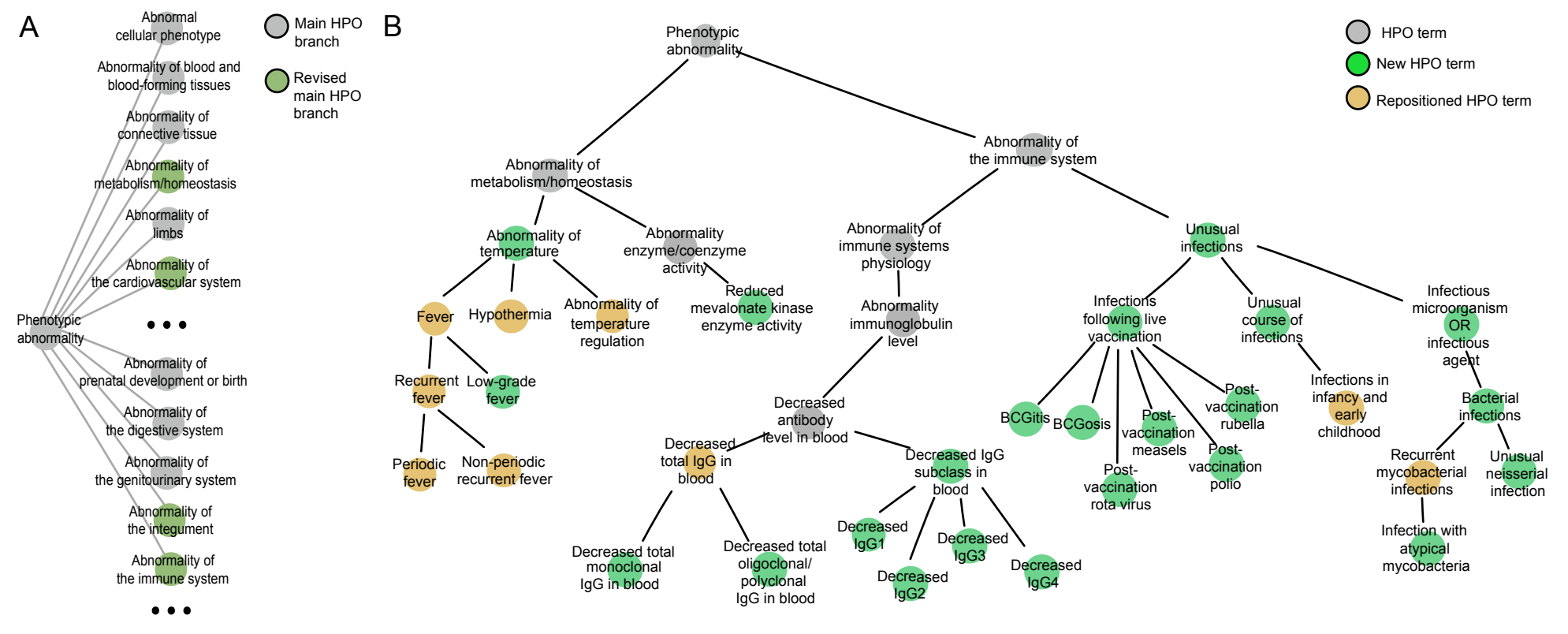
22. Girdea M, Dumitriu S, Fiume M, Bowdin S, Boycott KM, Chénier S, et al. PhenoTips: patient phenotyping software for clinical and research use. *Hum Mutat*. 2013 Aug;34(8):1057–65.

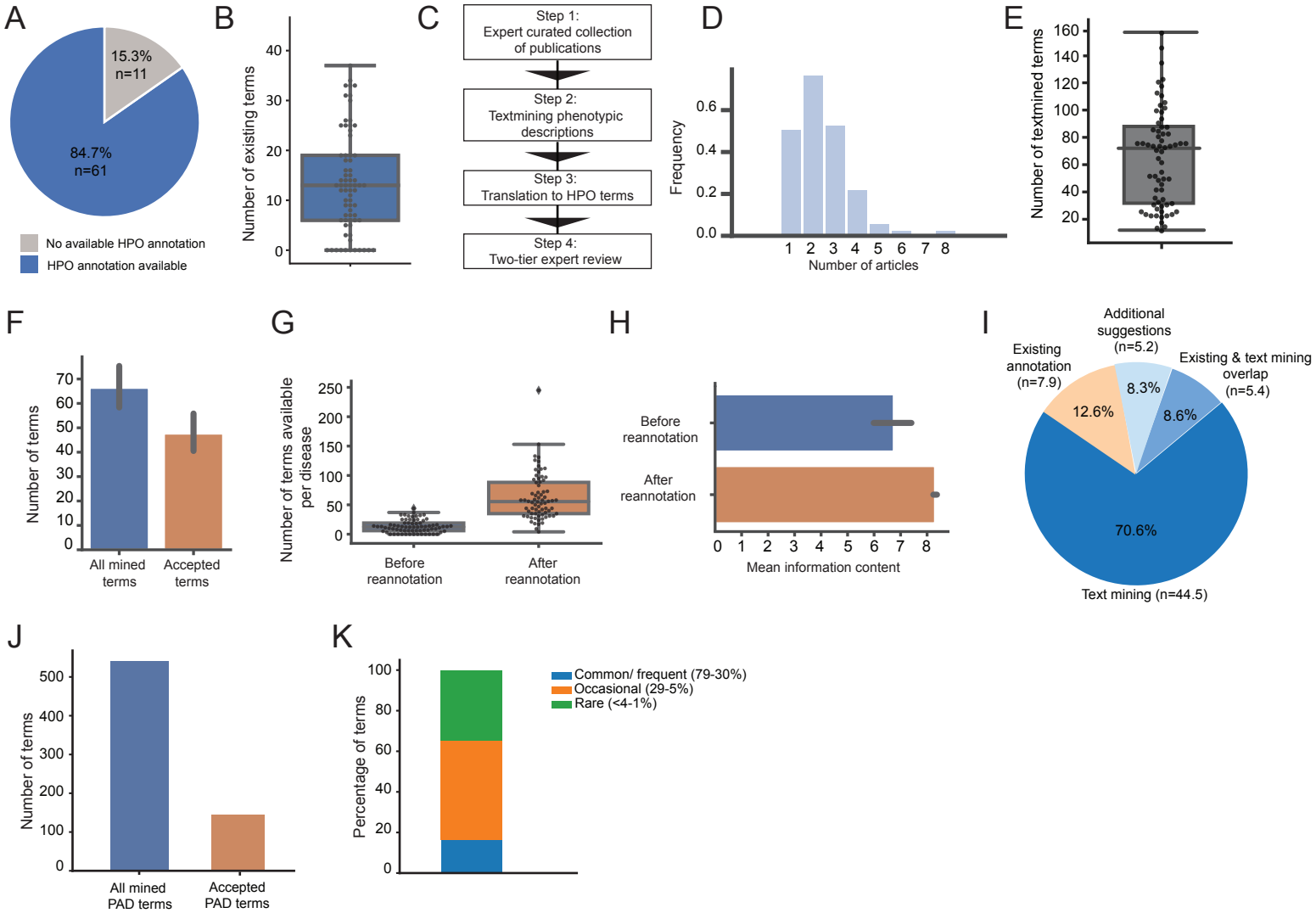
23. Smedley D, Jacobsen JOB, Jäger M, Köhler S, Holtgrewe M, Schubach M, et al. Next-generation diagnostics and disease-gene discovery with the Exomiser. *Nat Protoc*. 2015 Dec;10(12):2004–15.

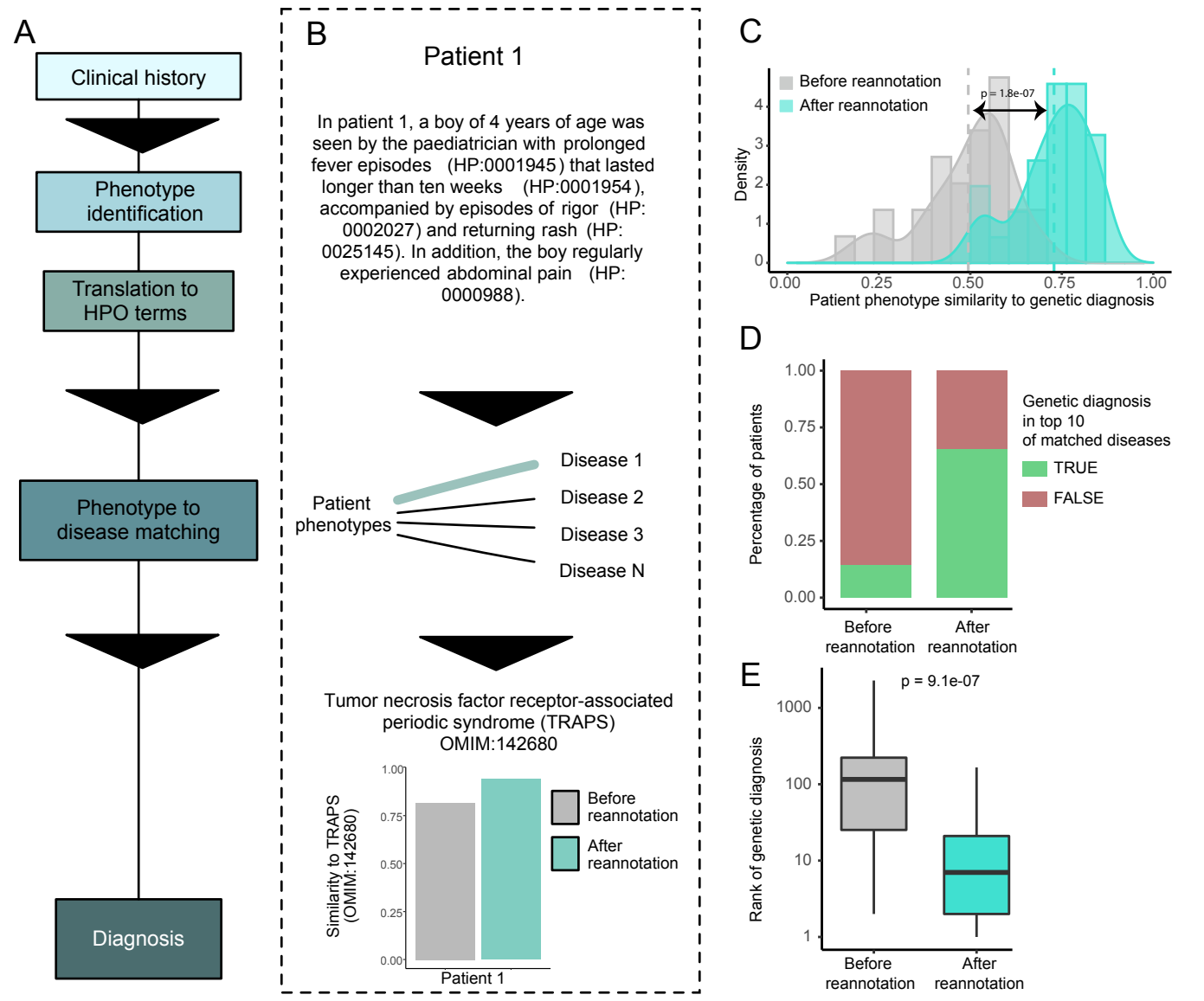
24. Westbury SK, Turro E, Greene D, Lentaigne C, Kelly AM, Bariana TK, et al. Human phenotype ontology annotation and cluster analysis to unravel genetic defects in 707 cases with unexplained bleeding and platelet disorders. *Genome Medicine*. 2015 Apr 9;7(1):36.

611 25. Lehne M, Luijten S, Vom Felde Genannt Imbusch P, Thun S. The Use of FHIR in
612 Digital Health - A Review of the Scientific Literature. *Stud Health Technol Inform*.
613 2019 Sep 3;267:52–8.
614 26. Shefchek KA, Harris NL, Gargano M, Matentzoglu N, Unni D, Brush M, et al. The
615 Monarch Initiative in 2019: an integrative data and analytic platform connecting
616 phenotypes to genotypes across species. *Nucleic Acids Res*. 2020 08;48(D1):D704–15.
617 27. Bousfiha A, Jeddane L, Picard C, Ailal F, Bobby Gaspar H, Al-Herz W, et al. The
618 2017 IUIS Phenotypic Classification for Primary Immunodeficiencies. *J Clin*
619 *Immunol*. 2018;38(1):129–43.
620 28. Seidel MG, Kindle G, Gathmann B, Quinti I, Buckland M, van Montfrans J, et al. The
621 European Society for Immunodeficiencies (ESID) Registry Working Definitions for
622 the Clinical Diagnosis of Inborn Errors of Immunity. *J Allergy Clin Immunol Pract*.
623 2019 Aug;7(6):1763–70.
624 29. Gasteiger LM, Robinson PN, Pazmandi J, Boztug K, Seppänen MRJ, Seidel MG.
625 Supplementation of the ESID registry working definitions for the clinical diagnosis of
626 inborn errors of immunity with encoded human phenotype ontology (HPO) terms. *The*
627 Journal of Allergy and Clinical Immunology: In Practice. 2020 May 1;8(5):1778.
628 30. Groza T, Köhler S, Doelken S, Collier N, Oellrich A, Smedley D, et al. Automatic
629 concept recognition using the human phenotype ontology reference and test suite
630 corpora. *Database (Oxford)*. 2015;2015.

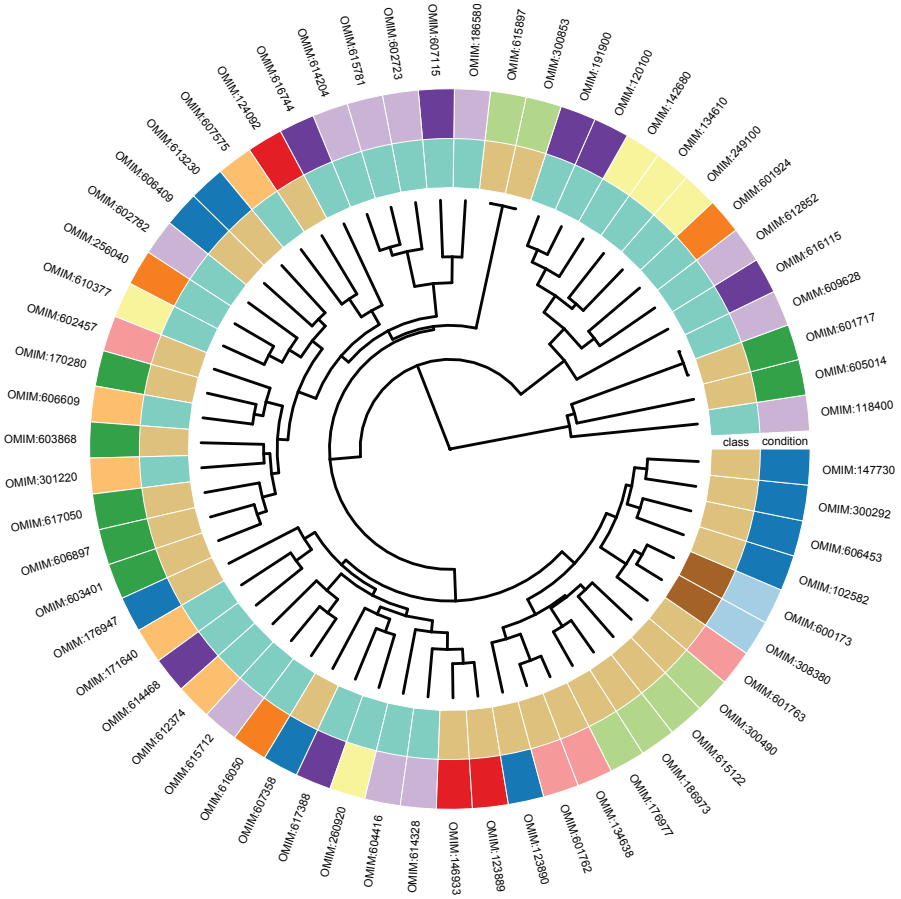

IUIS
Table 1
IUIS
Table 3
IUIS
Table 4
IUIS
Table 7
Immunodeficiencies affecting cellular and humoral immunity
Predominantly antibody deficiencies
Diseases of immune dysregulation
Autoinflammatory disorders
1. Collection of reviews/case reports
2. Text mining, data processing
Extraction of HPO terms with text miner
Preparation of summaries per disease
3. Two-tier expert review
Evaluation by experts
Evaluation by experts
Evaluation by experts
Evaluation by experts
4. Data processing, submission
Data filtering and summary
Main coordinators
hpo
Updates to HPO
Updates to OMIM
OMIM®
Online Mendelian Inheritance in Man®
An Online Catalog of Human Genes and Genetic Disorders
Updated 24 May 2011
Search
Advanced Search: OMIM, Clinical Synopses, Gene Map


A
Main HPO branch
Revised main HPO branch
Phenotypic abnormality
Abnormal cellular phenotype
Abnormality of blood and blood-forming tissues
Abnormality of connective tissue
Abnormality of metabolism/homeostasis
Abnormality of limbs
Abnormality of the cardiovascular system
Abnormality of prenatal development or birth
Abnormality of the digestive system
Abnormality of the genitourinary system
Abnormality of the integument
Abnormality of the immune system
B
HPO term
New HPO term
Repositioned HPO term
Phenotypic abnormality
Abnormality of metabolism/homeostasis
Abnormality of the immune system
Abnormality of temperature
Abnormality enzyme/coenzyme activity
Abnormality of immune systems physiology
Unusual infections
Fever
Hypothermia
Abnormality of temperature regulation
Reduced mevalonate kinase enzyme activity
Abnormality immunoglobulin level
Infections following live vaccination
Unusual course of infections
Infectious microorganism OR infectious agent
Recurrent fever
Low-grade fever
Periodic fever
Non-periodic recurrent fever
Decreased antibody level in blood
Decreased total IgG in blood
Decreased IgG subclass in blood
BCGitis
BCGosis
Post-vaccination measles
Post-vaccination rubella
Post-vaccination rota virus
Post-vaccination polio
Infections in infancy and early childhood
Bacterial infections
Recurrent mycobacterial infections
Unusual neisserial infection
Infection with atypical mycobacteria
Decreased total monoclonal IgG in blood
Decreased total oligoclonal/ polyclonal IgG in blood
Decreased IgG1
Decreased IgG2
Decreased IgG3
Decreased IgG4


A
15.3%
n=11
84.7%
n=61
No available HPO annotation
HPO annotation available
B
Number of existing terms
C
Step 1:
Expert curated collection
of publications
Step 2:
Textmining phenotypic
descriptions
Step 3:
Translation to HPO terms
Step 4:
Two-tier expert review
D
Frequency
Number of articles
E
Number of textmined terms
F
Number of terms
All mined
terms
Accepted
terms
G
Number of terms available
per disease
Before
reannotation
After
reannotation
H
Before
reannotation
After
reannotation
Mean information content
I
Additional
suggestions
(n=5.2)
Existing
annotation
(n=7.9)
Existing & text mining
overlap
(n=5.4)
8.3%
12.6%
8.6%
70.6%
Text mining (n=44.5)
J
Number of terms
All mined
PAD terms
Accepted
PAD terms
K
Percentage of terms
Common/ frequent (79-30%)
Occasional (29-5%)
Rare (<4-1%)


A
Clinical history
Phenotype identification
Translation to HPO terms
Phenotype to disease matching
Diagnosis
B
Patient 1
In patient 1, a boy of 4 years of age was seen by the paediatrician with prolonged fever episodes (HP:0001945) that lasted longer than ten weeks (HP:0001954), accompanied by episodes of rigor (HP: 0002027) and returning rash (HP: 0025145). In addition, the boy regularly experienced abdominal pain (HP: 0000988).
Patient phenotypes
Disease 1
Disease 2
Disease 3
Disease N
Tumor necrosis factor receptor-associated periodic syndrome (TRAPS) OMIM:142680
Similarity to TRAPS (OMIM:142680)
Before reannotation
After reannotation
Patient 1
C
Before reannotation
After reannotation
p = 1.8e-07
Density
Patient phenotype similarity to genetic diagnosis
D
Percentage of patients
Genetic diagnosis in top 10 of matched diseases
TRUE
FALSE
Before reannotation
After reannotation
E
p = 9.1e-07
Rank of genetic diagnosis
Before reannotation
After reannotation

A



IUIS classification

- Autoinflammatory disorders
- Diseases of immune dysregulation
- Immunodeficiencies affecting cellular and humoral immunity
- Primary antibody deficiencies

Disease subgroup

- Autoimmune Lymphoproliferative Syndrome (ALPS)
- Hemophagocytic Lymphohistiocytosis (HLH)
- Immune dysregulation with colitis
- Others
- Primary antibody deficiency
- Recurrent inflammation
- SCID T−B+
- Sterile inflammation (skin / bone / joints)
- Susceptibility to EBV
- Syndromes with autoimmunity
- Systemic inflammation with urticaria rash
- Type 1 Interferonopathies

B