Havrilla *et al.*
BMC Medical Informatics and Decision Making (2022) 22:198
https://doi.org/10.1186/s12911-022-01927-1

BMC Medical Informatics and Decision Making

RESEARCH

Open Access

PheNominal: an EHR-integrated web application for structured deep phenotyping at the point of care

James M. Havrilla[1], Anbumalar Singaravelu[2], Dennis M. Driscoll[2], Leonard Minkovsky[2], Ingo Helbig[3,4,5,6], Livija Medne[7], Kai Wang[1,5,8], Ian Krantz[7] and Bimal R. Desai[9*]

From International Conference on Intelligent Biology and Medicine (ICIBM 2021)
Philadelphia, PA, USA. 8-10 August 2021

Abstract

Background: Clinical phenotype information greatly facilitates genetic diagnostic interpretations pipelines in disease. While post-hoc extraction using natural language processing on unstructured clinical notes continues to improve, there is a need to improve point-of-care collection of patient phenotypes. Therefore, we developed "PheNominal", a point-of-care web application, embedded within Epic electronic health record (EHR) workflows, to permit capture of standardized phenotype data.

Methods: Using bi-directional web services available within commercial EHRs, we developed a lightweight web application that allows users to rapidly browse and identify relevant terms from the Human Phenotype Ontology (HPO). Selected terms are saved discretely within the patient's EHR, permitting reuse both in clinical notes as well as in downstream diagnostic and research pipelines.

Results: In the 16 months since implementation, PheNominal was used to capture discrete phenotype data for over 1500 individuals and 11,000 HPO terms during clinic and inpatient encounters for a genetic diagnostic consultation service within a quaternary-care pediatric academic medical center. An average of 7 HPO terms were captured per patient. Compared to a manual workflow, the average time to enter terms for a patient was reduced from 15 to 5 min per patient, and there were fewer annotation errors.

Conclusions: Modern EHRs support integration of external applications using application programming interfaces. We describe a practical application of these interfaces to facilitate deep phenotype capture in a discrete, structured format within a busy clinical workflow. Future versions will include a vendor-agnostic implementation using FHIR. We describe pilot efforts to integrate structured phenotyping through controlled dictionaries into diagnostic and research pipelines, reducing manual effort for phenotype documentation and reducing errors in data entry.

Keywords: EHR, EMR, Phenotype, Epic healthcare, Health record data

Background

Phenotypic data, especially in the electronic health records (EHR), is heterogeneous and sparse [1–3]. The data comes in numerous different formats that do not facilitate interoperability or direct comparison of

*Correspondence: desai@chop.edu

[9] Department of Pediatrics, University of Pennsylvania Perelman School of Medicine, Philadelphia, PA 19104, USA
Full list of author information is available at the end of the article

information [4–7]. Having a standardized, controlled phenotype vocabulary and data format is preferable to distinguish phenotypic groups and obtain a diagnosis [8]. Distinct phenotype ontology terms have already been used in several studies [9–16] to facilitate diagnosis of rare diseases and identification of causal genes. There is an acute need in clinical settings for well-structured phenotype data so that it can be combined with downstream gene panel, SNP array, or sequencing data for rapid, accurate comprehension of both common and rare diseases [17, 18]. Some tools can even use this data to rank genes without any sequencing information [18, 19], which can direct sequencing, gene panels, or downstream analyses.

To obtain distinct phenotype terms, we use data from standardized phenotype vocabularies, but the most common and widely utilized vocabulary is the Human Phenotype Ontology (HPO) [20]. The Human Phenotype Ontology was created to enable "deep phenotyping" through the capture of symptoms and phenotypic findings using a logically constructed hierarchy of phenotypic terms and enables a deep phenotyping approach wherein computable phenotypic profiles of human diseases and individual patients allow the linking of terms that are close to one another in the hierarchy and provides for a computational bridge between genome biology and clinical medicine. HPO has become the de facto standard for representing clinical phenotype data in a multitude of programs including the NIH Undiagnosed Diseases Program (UDP) [21], several NCBI databases including MedGen [22], ClinVar [23], and the Genetic Testing Registry [24], the Sanger Institute databases DDD [25] and DECIPHER [26], the rare diseases section of the UK's 100,000 genome project [27], the Genomic Matchmaking API of the Global Alliance for Genomics and Health [28], and many others. The UDP demonstrated that the use of HPO in comparison to clinical data alone in ES/GS variant analysis improves molecular diagnosis by 10–20% [29], as has several other studies [26, 30–32]. HPO provides a substantially more detailed representation of clinical phenotypes than other clinical terminologies and ontologies and is designed for computational analysis by linking to computational disease definitions and to ontologies of gene function, anatomy, biochemistry, and other biologic attributes.

Despite the improvements provided by HPO, it can be difficult to train people in the use of standardized ontologies, and natural language processing (NLP) tools can be difficult to integrate and vary greatly in consistency [18, 33, 34]. Several extracted terms are also near-synonymous and not merged depending on the ontology used [35]. Recently, point-of-care strategies have emerged to address these facts, such as improved testing quality [36], deeper phenotyping strategies in EHR data [37, 38], and improved clinical documentation [39], but these improvements are not yet enough.

Even with these improvements at the point-of-care, these automated tools are unlikely to achieve the same result as manual annotation by expert users (such as the patients' physicians, specialists, or genetic counselors), and while there are resources containing integrated HPO terms for use in manual annotation such as the Human Disease Gene website [40], they are not kept up-to-date individually and either miss new or contain outdated terms. For example, the epilepsy phenotypes within the HPO were recently updated in December 2020 [41] in alignment with the most current guidelines formulated by the International League Against Epilepsy in 2017. However, if a provider uses a framework based on a static version of the HPO from 2019 it may lack such recent information, or use terms that are no longer in use in the epilepsy community. This inability to implement the most recent release rapidly can create inconsistent patient phenotyping depending on the expert user's resource of choice.

Previously, genetic counselors and physicians at the Roberts Individualized Medical Genetics Center (RIMGC) at the Children's Hospital of Philadelphia (CHOP) used a manual process of annotating clinic encounter notes with tags from the HPO website. Providers copied and pasted HPO codes into encounter progress notes, often with errors, and while they were easy to find they were not discretely captured. We have created PheNominal to remedy this deficiency. PheNominal is a tool to assist expert users in annotation of patient notes with preexisting phenotype terms from HPO. In addition to raw patient notes, users can now extract a full set of HPO terms curated by the physicians themselves in easily parsable formats for downstream bioinformatics pipelines. In cases where genetic variant data is attained for patients, there are a variety of tools that can utilize this deeper, more normalized phenotype data to rank candidate genes for variants [19, 42–44].

With this tool, we hope to create a more consistent and standardized method for expert curation of phenotype terminology for reproducibility and dissemination to bioinformatics pipelines. In this study, we describe the tool development process and how to use the tool at point-of-care. We demonstrate how the tool has improved the accuracy, speed and willingness of physicians in phenotyping their patients at a major genetic testing center for pediatric patients. Finally, we provide a realistic example of how discrete, structured phenotyping of patients can lead to a genetic diagnosis in disease.

Materials

Development

Using the Agile software methodology and a series of development sprints [45], the Emerging Technology and Transformation Team actively involved customers in the product development process and continually implemented their feedback and rapidly tested each feature. The goal of the design was to create an interface that was easy to use for clinicians with no specialized technical background, particularly those less familiar with various ontological databases. The development team worked with providers at the RIMGC at CHOP to design and validate the interface. Autocompletion was added to deal with common typographic errors that led to incomplete or inaccurate annotations and increased note taking time. Physician suggestions, such as adding comments to the HPO terms for negation and gene updates, were critical. Providers were given biweekly demonstrations to assess progress, perform tests and work with the product; provider feedback was clear and concise as requirements were clearly specified for feedback, and every new feature was tested immediately in production by physicians and counselors.

Typical use case of PheNominal

The PheNominal app is smoothly integrated into the Epic electronic health record (Epic Systems Corporation, Verona, WI), and currently available as a tab labeled "HPO URL" in the Epic Hyperspace once the patient chart is opened. The application works akin to a shopping cart for an online website (Fig. 1). Training in using the tool involves a brief live demo for prospective users on how they can search and enter a HPO term for the patient and some troubleshooting guidelines. Users have access to the entire HPO vocabulary, which is kept up-to-date: as of January 2021 it has over 13,000 terms and 156,000 annotations. The user can browse the full scope of HPO simply by typing the initial letters of a term, using the "autocomplete" feature as an accelerator, selecting the term, and clicking "Add HPO" to confirm selection. Users can browser the HPO hierarchy to find the correct level of precision by clicking "Details," which reveals the related synonymous, superclass, and subclass terms, as well as the description of the term and gene annotation data for the term from Entrez [46] and HGNC [47] (Fig. 1, Part 3).

As the user adds terms to the "shopping cart," they can add free-text annotations to each term, such as linked

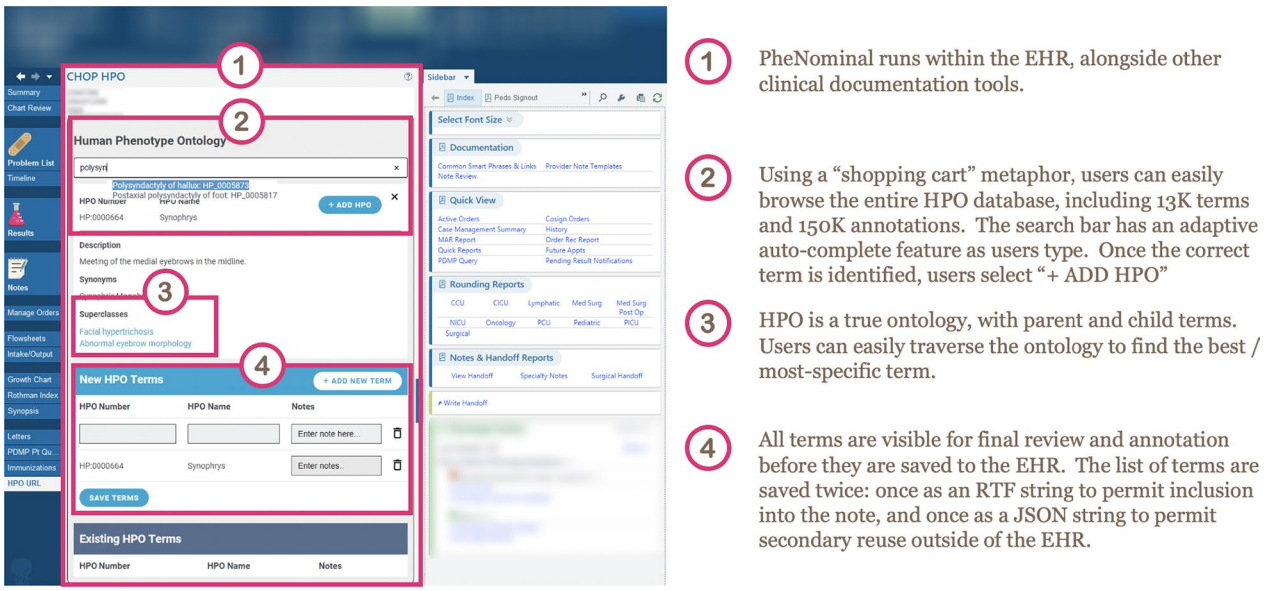

Fig. 1 PheNominal basic use case. In part 2, the user is considering the addition of 2 discrete HPO terms, synophrys (unibrow) and polysyndactyly of hallux (extra big toe) for their patient. In part 3, the superclass and subclass terms are highlighted, as in some cases the user may prefer to be more or less specific in their term choice. After adding terms in part 2, in part 4 the user may choose to save the terms to the EHR, which makes them immediately available for inclusion in patient notes as a pre-formatted Rich Text Format (RTF) table, as well as available for downstream use in bioinformatics or analytics pipelines

genes, or if the term is a negated term (Fig. 1, Part 4). Clicking "Save Terms" saves the information to "Existing HPO Terms." The full set of HPO terms and annotations are formatted as a single JavaScript Object Notation (JSON) string, which is then stored in the Epic Chronicles database as an Epic SmartData Element (SDE). An SDE is a vendor-specific technique for storing discrete key-value pairs in Epic such that important phenotype information can be stored consistently, reproducibly, and discretely. New, modified, or removed terms are saved to the patient's medical record after updating terms as well (Additional File 1, demo at: PheNominal Demonstration and Tutorial). Because the Epic EHR has standard web service methods to access and write SDEs as well as methods to manipulate SDE data within a clinical note using "SmartLinks," we are able to format and present an RTF formatted tabular view of the JSON data for use in Epic notes, retrieve discrete HPO terms for use in downstream bioinformatics and pipelines, and query patient records for specific terms in the enterprise data warehouse (Fig. 2, Additional file 1).

App architecture and design

The app was developed in JavaScript and CSS using Node.js. It communicates in real-time with Epic Caché as well as the Bioontology API from BioPortal, which gives us the benefit of access to the very latest version of the HPO ontology and its terms. We also maintain a local version of the HPO ontology for performance that is also kept up-to-date via mailing lists and automated server queries. Users can benefit from access to fully updated terms as well as outdated terms if necessary, as all saved terms are version-tagged and permanently stored in the electronic health record unless updated by an authorized user.

PheNominal uses two vendor-specific web services (getSmartDataValues and setSmartDataValues) to access and write HPO terms as SDEs (Fig. 3). When the end user launches PheNominal, the tool retrieves and parses the most current HPO term set from NCBO BioPortal via the Bioontology API [48]. As the user manipulates HPO terms in the interface, changes are relayed to the application server to send and receive relevant data. The web services communicate via Epic Interconnect to generate two SDEs, a JSON payload and a pre-formatted RTF table for inclusion in clinical notes. If the user decides to save the term, this is updated downstream by adding the SDE to our Epic Clarity database for the patient. Finally, as in Fig. 2, Epic Smartlink allows for easy insertion of SDEs into patient notes.

Results

We collaborated with the Emerging Technology and Transformation team of the IT department and the clinicians and counselors of the RIMGC at CHOP to assess the improvement of PheNominal over manual entry of phenotype terms in number, accuracy, and speed. The IT team members were only the designers of the tool, and the RIMGC clinicians and counselors were the users. Before PheNominal, RIMGC clinicians had to navigate to the HPO web page manually, copy HPO terms and IDs, return to the Epic Hyperspace to paste them into patient notes without correcting them, or even type them out manually. This had an unacceptably high error rate for reliable inclusion in bioinformatics pipelines. During an effort to manually convert free text HPO terms from manually entered, historical notes into discrete HPO terms, the development team found thousands of HPO terms that had incorrect annotations, missing data, or typographic errors. Forty of these records were so corrupted they could not be migrated programmatically. The need for a tool like PheNominal became abundantly clear early on in the migration process.

In over almost 5 years of the legacy system of manual SmartPhrase input, only 1175 individuals' records were annotated with HPO terms. But in the 1 year and 4 months since PheNominal's implementation in the system, over 1500 patients' records were annotated (Table 1). When the Legacy system was in place, HPO terms were only assigned to patients undergoing exome sequencing tests. Since the implementation of PheNominal, HPO terms have been assigned to all patients undergoing any clinical evaluation as well as those undergoing any genetic testing through the RIMGC program. The ease of use of the PheNominal app allowed it to be applied to the entire patient population evaluated by the RIMGC clinic, which is the primary reason for the notable increase in the number of patients served. In a little over a year after PheNominal's implementation, 1000 more HPO terms were saved into Epic accurately than in nearly 5 years of manual input, with 3 times the average speed at 5 min per patient. The other benefit of having an automated app with autocompletion is it ensures all terms are entered correctly, and thus discretely, for future pipelines and downstream analyses. In addition all terms must be entered and viewed on an encounter-by-encounter basis for each patient in the legacy system, but with PheNominal, all HPO terms can be edited and viewed simultaneously for convenience and speed. It is important to note that there was no difference between the patient populations served by the Legacy system and PheNominal, nor in the clinical teams doing the HPO annotation: for each case the annotation was performed by a genetic counselor and reviewed by a physician geneticist.

After PheNominal was implemented around the end of June 2019, the number of patient records annotated per month only went up slightly over time, but the HPO

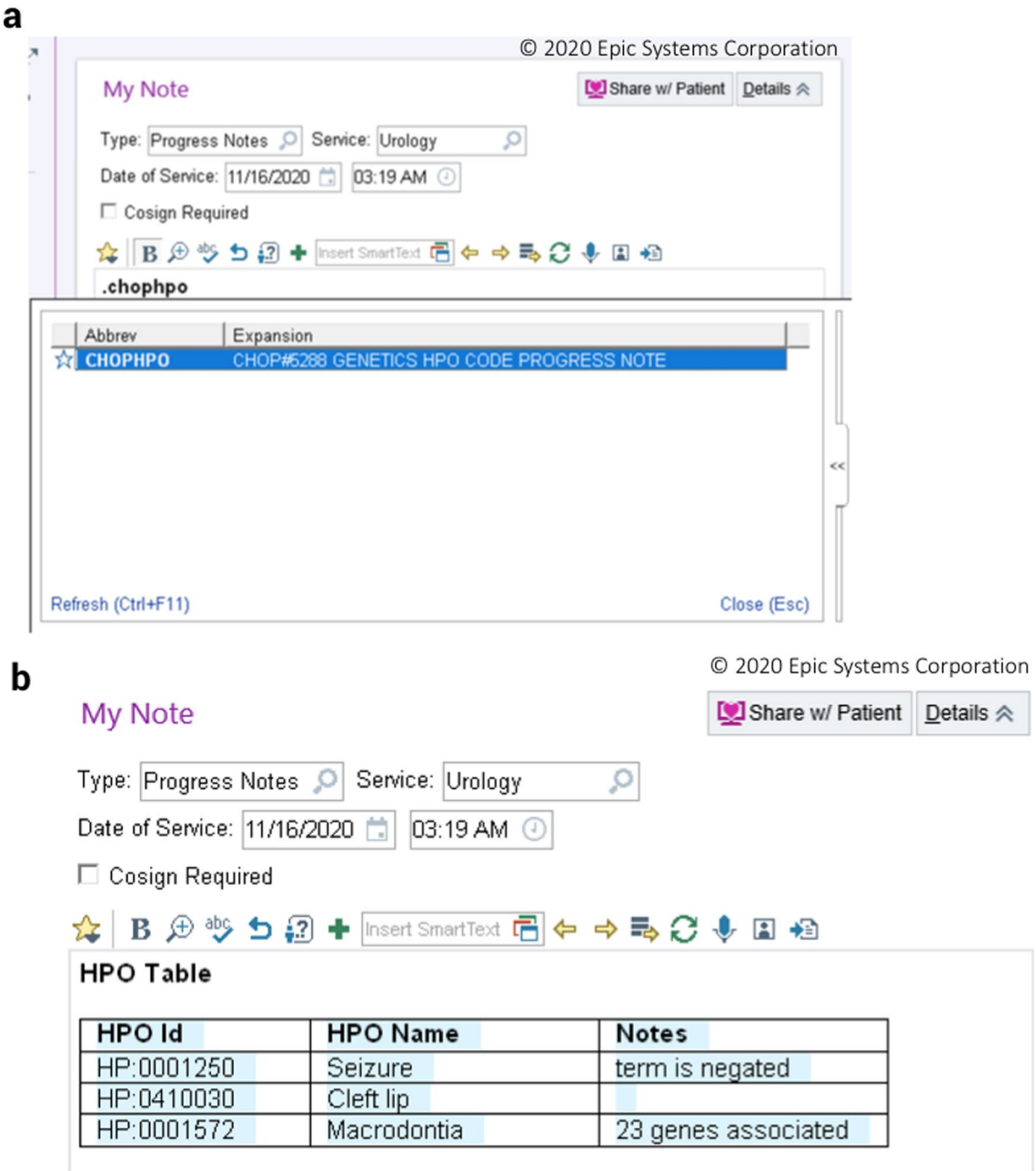

Fig. 2 Integrating HPO terms from PheNominal into patient notes as RTF tables. **a** HPO terms are stored as two discrete SDEs: one, a JSON-formatted payload containing date and version information, and the other, **b** a simplified RTF table intended for use in clinical notes. Including the SmartLink ".chophpo" in a clinical note pulls the preformatted RTF table with annotations into the note itself

terms annotated per patient substantially increased (Fig. 4). This is partially due to recruitment of more patients outside of simple clinical testing over time but primarily because there was a marked increase in comfort by physicians and counselors using the tool to find the most specific and descriptive terms as the users familiarized themselves with the ontology. We believe, while difficult to quantify, that this factor contributed strongly to the large gradual increase in HPO term count per month over time, which is not explained by the far less appreciable increase in patient count per month.

By virtue of PheNominal, not only are there less errors due to the discrete choices provided by the app, but now users can traverse parent and child HPO

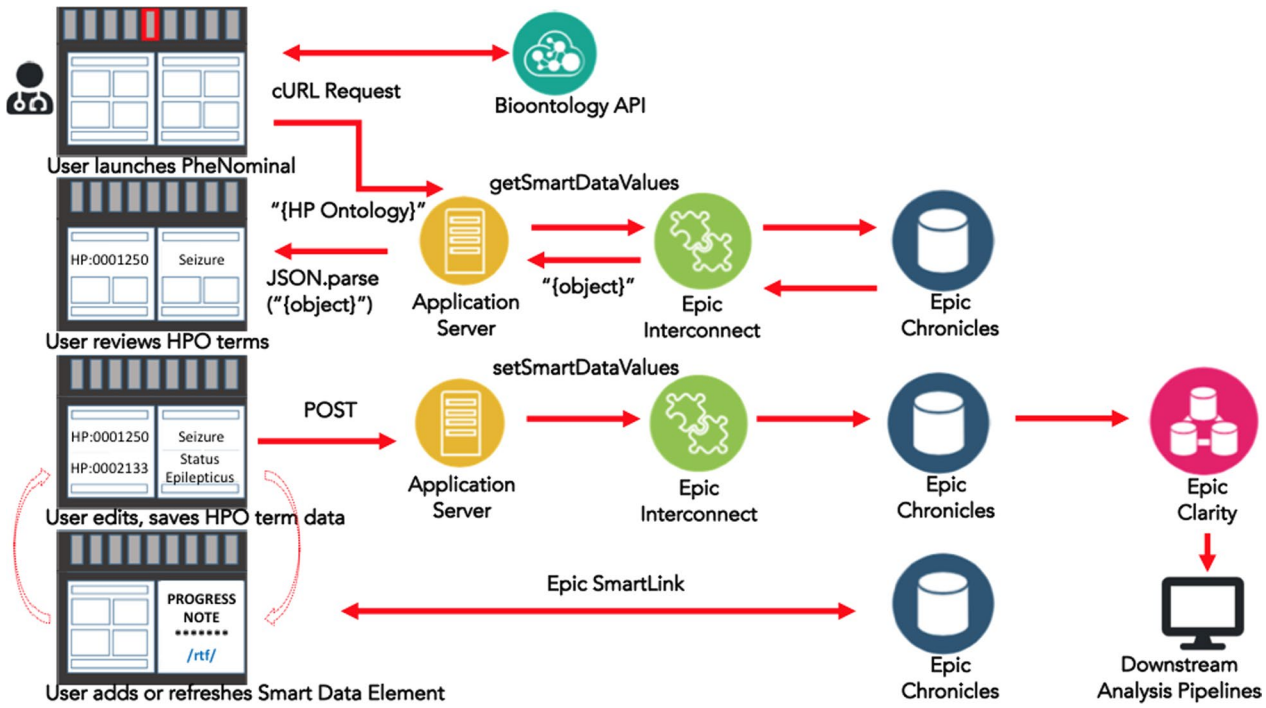

Fig. 3 PheNominal system architecture and workflow. Users can search, review, add and delete HPO terms by having the Progress Notes open side by side for clinical analysis. The search is based on an external API built by NCBO BioPortal, which gets the HPO data from the HPO site. We store a copy of the full HPO data locally into our local application server and use it to populate the HPO details such as HPO name, description, synonyms, superclasses, subclasses and HPO gene association. Terms are discretely stored in the EHR, and pushed to Epic Chronicles, and daily to Epic Clarity and our proprietary CHOP data warehouse, and can be imported into the clinical note directly for use in other downstream pipelines, or downloaded from Clarity by direct SQL query by authorized researchers

Table 1 Comparison between legacy system (manual entry of SmartPhrases) and PheNominal on patient encounters. This compares the total measurement period for each system, number of patients served, the number of HPO terms entered correctly, how long each method takes on average, and how terms are viewed and if they are truly discrete in all cases

	Legacy (manual entry)	PheNominal
Time	4 years, 9.5 months (09/01/2014–06/19/2019)	1 year, 7.5 months (06/19/2019–1/31/2021)
Patients served	1175	1760
HPO terms saved	10,050	13,566
Time to enter terms	15 min mean	5 min mean, 2 min mode
Viewing terms for a patient	Each encounter must be opened individually	All at once across encounters
Discrete?	No	Yes

terms in the ontology tree with ease. For example, as in the HPO paper from Robinson et al. [49], a physician can navigate the tree and where they would previously annotate "hip dislocation" they can now write "congenital bilateral hip dislocation." Alternatively, a user may wish to be less specific, when they do not have enough facts to identify a specific feature. PheNominal therefore makes overall diagnosis easier, and assists in accurate gene-phenotype association for downstream bioinformatics and sequencing pipelines.

As an example of how PheNominal assists in downstream analysis, we took a sample of unannotated patient notes from Yu et al. on a patient with hereditary spastic paraplegia [50] (Fig. 5). We can add the terms from the patient notes with PheNominal, under the assumption it would prevent input error with its autocompletion,

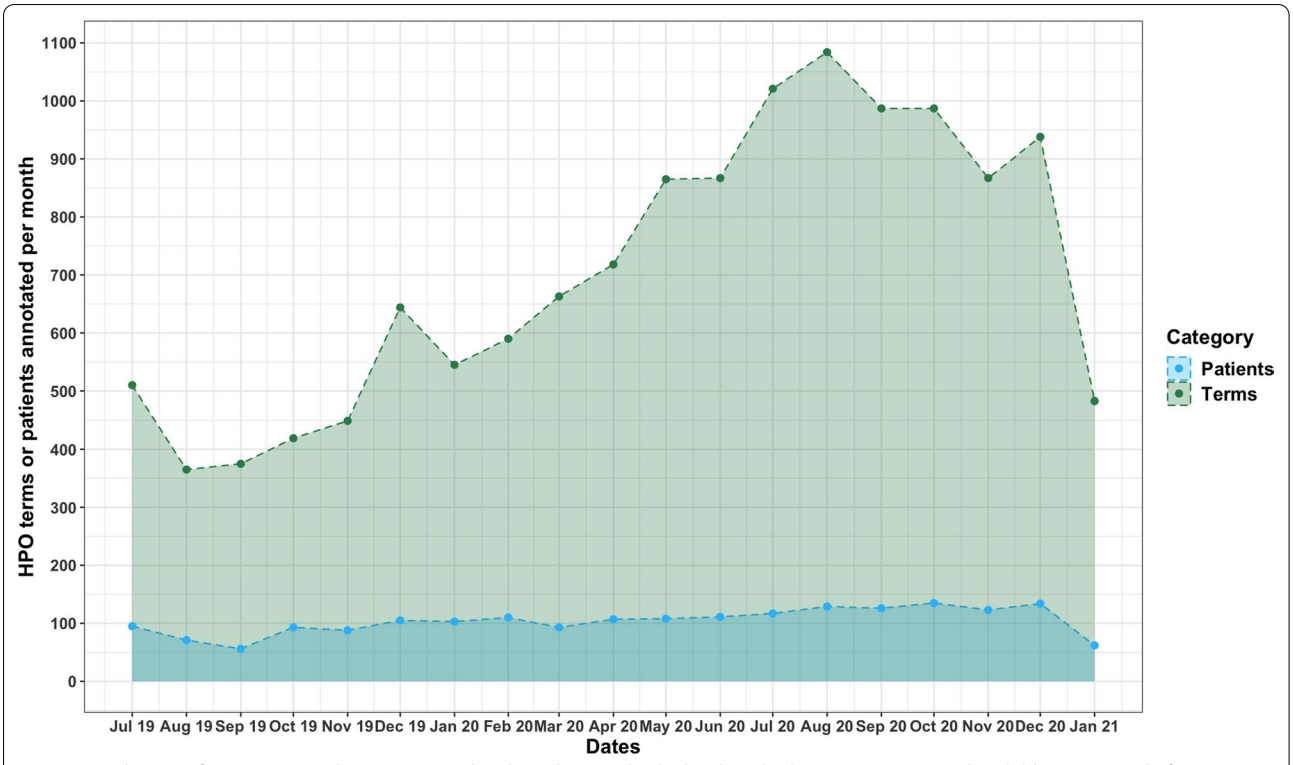

Fig. 4 Distribution of HPO terms, and patients served at the Roberts Individualized Medical Genetics Center at the Children's Hospital of Philadelphia after the initial implementation of PheNominal implementation from July 2019 to January 2021. The green line number of unique patients with HPO terms annotated by PheNominal per month. The blue line is the number of HPO terms annotated for all patients per month by PheNominal. This figure was generated in R using the abovementioned data gathered from the RIMGC directly

which is the set of HPO terms seen on the left side of Fig. 5. However, if we allowed physicians to type the notes or manually enter or copy-paste the HPO terms into patient notes and parse them using the basic Aho-Corasick algorithm from Doc2Hpo, we could run into some preventable user errors, such as pasting incomplete terms or misspelling terms. As mentioned previously, the RIMGC produced thousands of incorrect annotations. For the sake of argument, if these mistakes remove even 1 term from the 5 term list in Fig. 5, we go from scoring the causal gene (without any genetic variant knowledge) in the top 3 genes, to scoring it in the top 127 genes, using Phen2Gene in our downstream pipeline. PheNominal is critical for preventing simple errors like these that can seriously hinder or prevent the proper diagnosis of disease for troubled patients.

Discussion

Here, we report PheNominal, a point-of-care web application, embedded within Epic electronic health record (EHR) workflows, to permit capture of standardized phenotype data. PheNominal has improved the speed and ease with which physicians and genetic counselors can input discrete, reliable phenotypic data in the EHR. Specificity of terms can now be easily modulated, and the more specific an HPO term, the more information it contains. There are clear advantages of using PheNominal to consistently and reproducibly capture discrete HPO terms for downstream use in computational pipelines, as opposed to the makeshift post-hoc extraction of the past. Thanks to PheNominal, annotation mistakes and sparse phenotyping made by manual copy-and-paste entry can become an ancient memory for all healthcare systems and providers.

Future improvements to PheNominal will address limitations in compatibility, interoperability and comprehensiveness of the tool. Because PheNominal acts as a general-purpose ontology browser and annotation tool, it can be expanded to include other ontologies beyond HPO. We can also provide sorted autocomplete information that is ranked by relevance so that users are less likely to choose the wrong term, so we are not simply substituting typing errors for term selection errors. Term recommendations may be further improved by future work in NLP tools that can predict potential HPO terms based on a set of already chosen terms, though at the moment no tools can fulfill this purpose. Additionally, while PheNominal currently uses vendor-specific

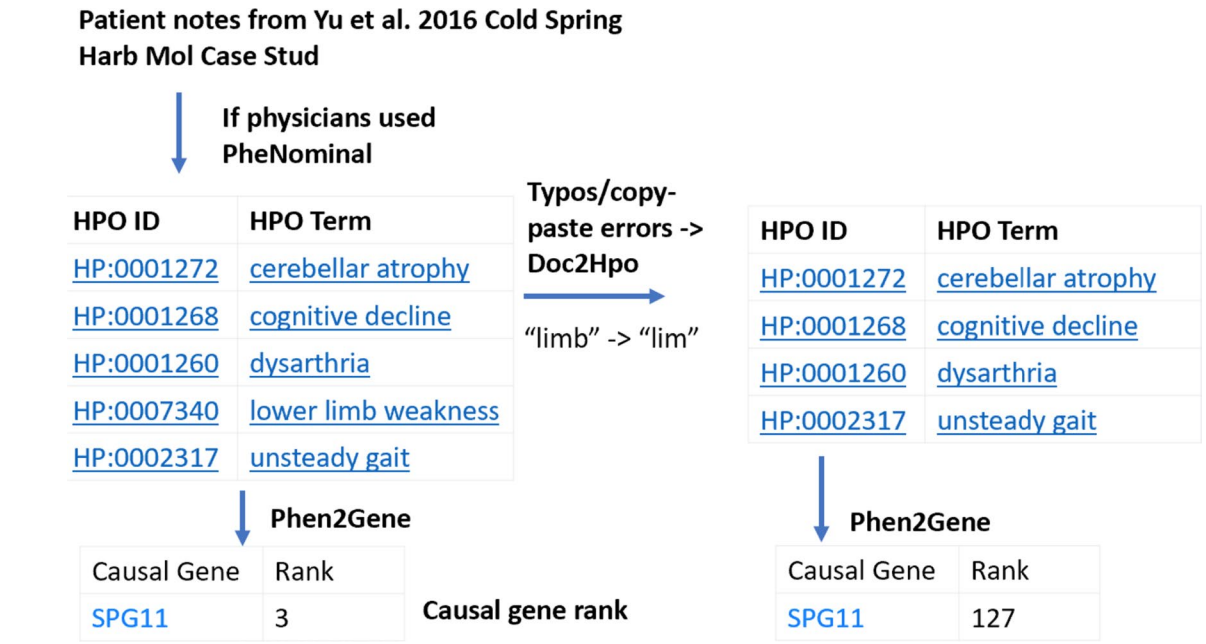

Fig. 5 Inputting terms with the support of PheNominal versus manual extraction downstream with NLP tools. In this example, we derived several HPO terms derived from patient notes with the help of PheNominal. We provide realistic examples of 3 reproducible copy-paste errors like "limb" to "lim" that produce missing terms when parsed by NLP algorithms such as Doc2Hpo. The derived HPO terms are then passed to Phen2Gene, which ranks potential disease genes based on phenotype information alone. In this example, the rank drops from 3 to 127, which may result in a variant in a candidate gene (*SPG11*) not to appear on the final diagnostic report

web services to read/write Epic SDEs, there is strong motivation to port the platform to use FHIR resources to improve cross-platform interoperability.

Because EHR vendors differ in their implementation of FHIR, and because PheNominal is dependent on the use of SDEs to store raw JSON data, there are some important considerations for future FHIR integration efforts. Some of the point-of-care affordances, such as the real-time availability of the entered terms in clinical notes, are harder to port to other EHRs or in a pure FHIR app, and the SmartPhrase concept may no longer be supported in all cases, which makes porting currently annotated Progress Notes difficult. But while the FHIR standard is still going through major changes, early conversations with the FHIR Genomics Working Group at CHOP have been positive and suggest a few potential options for conversion to FHIR-based resources.

Integrating new information into PheNominal is relatively easy. There is already an API in place for Phen2Gene [19] to take the gene annotations from PheNominal and return scores for each term within a second. With Phen2Gene ranking genes, we can sort by score, and prioritize potentially causal genes for physicians and genetic counselors at the point of care. Since we are using the Bioontology API, porting other ontologies contained in NCBO BioPortal is also easy: OMIM [51], SNOMED [52], MeSH [53], ICD-10 [54], ORDO [55], and DOID [56]. It is also entirely possible to combine other data collection tools at the point of care with the help of PheNominal and with the help of SDEs, integrating this information into downstream clinical workflows for Clinical Decision Support (CDS). We hope that clinicians and counselors will find this application a useful resource for improving the sensitivity and reproducibility of phenotypic annotation, and would use it to improve diagnosis speed and accuracy at the point of care.

Conclusion

PheNominal is a pilot effort to incorporate structured phenotype information using precise dictionaries into downstream diagnostic and research pipelines by reducing manual input during phenotype documentation and generation of patient notes. This reduces errors in data entry that lead to more accurate downstream results in delineating the subphenotype, and predicting candidate genes for disease.

We believe there are 5 main innovations of PheNominal. It is a tool for discrete, point-of-care capture of clinically precise phenotype terms with minimal effort. It utilizes secure and standard-compliant encoding and storage of the HPO terms directly into the EHR. The data is dual-formatted both to permit downstream reuse in JSON format, containing date and version information, as well as

to allow point-of-care insertion into a clinical note in RTF format. PheNominal is integrated into the EHR directly through native web services, permitting generalizability to any other EHR implementations and the FHIR standard. Lastly, it permits integration of other ontologies and evolving standards like Phenopackets.

Abbreviations

EHR: Electronic health record; FHIR: Fast healthcare interoperability resources; HPO: Human phenotype ontology; CDS: Clinical decision support; RTF: Rich text format; API: Application programming interface; SDE: SmartData Element; CHOP: Children's Hospital of Philadelphia.

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s12911-022-01927-1.

Additional file 1. PheNominal Demonstration and Tutorial.

Acknowledgements

We would like to thank the entire Emerging Technology and Transformation Team at CHOP's IS department for their help with developing the tool, as well as the efforts of every user at the RIMGC who contributed to the development of the tool.

About this Supplement

This article has been published as part of BMCÐedical Informatics and Decision Making Volume 22 Supplement 2, 2022:Ðelected articles from the International Conference on Intelligent Biology andÐedicine (ICIBM 2021): medical informatics and decision making. The fullÜontents of the supplement are available online at https://bmcmedinformdecismak.biomedcentral.com/articles/supplements/volume-Ò2-supplement-2.

Author contributions

JMH wrote the manuscript, generated the figures, and performed and organized the analysis. BRD conceived, designed, and organized the study and tool development. KW advised the study and computational experiments to perform comparative evaluation. IH provided critical feedback on the implementation of the tools and point-of-care usage of the tools. IK assisted in the tool implementation and evaluation on CHOP samples. All co-authors read and reviewed the manuscript. The Emerging Technology and Transformation Team members (AS, DMD, LM) developed the tool and AS generated the data. All authors read and approved the final manuscript.

Funding

Funding for the tool and its development came from the Information Services and Emerging Technology & Transformation team's account at CHOP. Open-access publication costs, as well as JMH and KW, were funded by NIH/NLM/NHGRI grant LM012895 and NIH/NIGMS grant GM132713 and the Penn/CHOP Intellectual and Developmental Disabilities Research Center grant - NIH/NICHD P50 HD105354. The funding bodies themselves had no role in the design or study/collection/analysis of data or writing of the manuscript.

Availability of data and materials

There are no new data associated with this article, as no new data were generated or analyzed in support of this research. Patients annotated with the tool cannot have their annotation data shared for HIPAA privacy reasons. The software tool was developed at The Children's Hospital of Philadelphia, and is available with appropriate institutional usage and license agreement.

Declarations

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Competing interests

The authors declare that they have no competing interests.

Author details

[1]Center for Cellular and Molecular Therapeutics, Children's Hospital of Philadelphia, Philadelphia, PA 19104, USA. [2]Emerging Technology and Transformation Team, Information Services, Children's Hospital of Philadelphia, Philadelphia, PA 19104, USA. [3]Division of Neurology, Children's Hospital of Philadelphia, Philadelphia, PA 19104, USA. [4]The Epilepsy NeuroGenetics Initiative (ENGIN), Children's Hospital of Philadelphia, Philadelphia, USA. [5]Department of Biomedical and Health Informatics, Children's Hospital of Philadelphia, Philadelphia, PA 19104, USA. [6]Department of Neurology, University of Pennsylvania, Perelman School of Medicine, Philadelphia, PA 19104, USA. [7]Roberts Individualized Medical Genetics Center, Children's Hospital of Philadelphia, Philadelphia, PA 19104, USA. [8]Department of Pathology and Laboratory Medicine, University of Pennsylvania Perelman School of Medicine, Philadelphia, PA 19104, USA. [9]Department of Pediatrics, University of Pennsylvania Perelman School of Medicine, Philadelphia, PA 19104, USA.

Received: 13 June 2022 Accepted: 6 July 2022
Published online: 28 July 2022

References

1. Pivovarov R, Perotte AJ, Grave E, Angiolillo J, Wiggins CH, Elhadad N. Learning probabilistic phenotypes from heterogeneous EHR data. J Biomed Inform. 2015;58:156–65. https://doi.org/10.1016/j.jbi.2015.10.001.
2. Hripcsak G, Albers DJ. Next-generation phenotyping of electronic health records. J Am Med Inform Assoc. 2013;20:117–21. https://doi.org/10.1136/amiajnl-2012-001145.
3. Weiskopf NG, Weng C. Methods and dimensions of electronic health record data quality assessment: enabling reuse for clinical research. J Am Med Inform Assoc. 2013;20:144–51. https://doi.org/10.1136/amiajnl-2011-000681.
4. Blobel B, Interoperable EHR. Interoperable EHR systems—challenges, standards and solutions. ejbi. 2018. https://doi.org/10.24105/ejbi.2018.14.2.3.
5. Reisman M. EHRs: the challenge of making electronic data usable and interoperable. P T. 2017;42:572–5.
6. Pryor TA, Hripcsak G. Sharing MLM's: an experiment between Columbia-Presbyterian and LDS Hospital. In: Proceedings of the annual symposium on computer applications in medical care; 1993. p. 399–403. https://www.ncbi.nlm.nih.gov/pubmed/8130503.
7. Hripcsak G, Kuperman GJ, Friedman C. Extracting findings from narrative reports: software transferability and sources of physician disagreement. Methods Inf Med. 1998;37:1–7.
8. Griffiths AJF, Gelbart WM, Miller JH, Lewontin RC. Genetics begins with Variation. In: Freeman WH editors. Modern genetic analysis; 1999. https://www.ncbi.nlm.nih.gov/books/NBK21344/. Accessed 1 Oct 2020.
9. Li Q, Zhao K, Bustamante CD, Ma X, Wong WH. Xrare: a machine learning method jointly modeling phenotypes and genetic evidence for rare disease diagnosis. Genet Med. 2019;21:2126–34. https://doi.org/10.1038/s41436-019-0439-8.
10. Jia J, Wang R, An Z, Guo Y, Ni X, Shi T. RDAD: a machine learning system to support phenotype-based rare disease diagnosis. Front Genet. 2018;9:587. https://doi.org/10.3389/fgene.2018.00587.
11. Díaz-Santiago E, Jabato FM, Rojano E, Seoane P, Pazos F, Perkins JR, et al. Phenotype-genotype comorbidity analysis of patients with rare disorders provides insight into their pathological and molecular bases. PLoS Genet. 2020;16:e1009054. https://doi.org/10.1371/journal.pgen.1009054.
12. Yang H, Robinson PN, Wang K. Phenolyzer: phenotype-based prioritization of candidate genes for human diseases. Nat Methods. 2015;12:841–3. https://doi.org/10.1038/nmeth.3484.
13 Robinson PN, Köhler S, Oellrich A, Sanger Mouse Genetics Project, Wang K, Mungall CJ, et al. Improved exome prioritization of disease

genes through cross-species phenotype comparison. Genome Res. 2014;24:340–8. https://doi.org/10.1101/gr.160325.113.
14. Helbig I, Lopez-Hernandez T, Shor O, Galer P, Ganesan S, Pendziwiat M, et al. A recurrent missense variant in AP2M1 impairs Clathrin-mediated endocytosis and causes developmental and epileptic encephalopathy. Am J Hum Genet. 2019;104:1060–72. https://doi.org/10.1016/j.ajhg.2019.04.001.
15. Galer PD, Ganesan S, Lewis-Smith D, McKeown SE, Pendziwiat M, Helbig KL, et al. Semantic similarity analysis reveals robust gene-disease relationships in developmental and epileptic encephalopathies. Am J Hum Genet. 2020;107:683–97. https://doi.org/10.1016/j.ajhg.2020.08.003.
16. Ganesan S, Galer PD, Helbig KL, McKeown SE, O'Brien M, Gonzalez AK, et al. A longitudinal footprint of genetic epilepsies using automated electronic medical record interpretation. Genet Med. 2020;22:2060–70. https://doi.org/10.1038/s41436-020-0923-1.
17. Cheng H, Capponi S, Wakeling E, Marchi E, Li Q, Zhao M, et al. Missense variants in TAF1 and developmental phenotypes: challenges of determining pathogenicity. Hum Mutat. 2019. https://doi.org/10.1002/humu.23936.
18. Son JH, Xie G, Yuan C, Ena L, Li Z, Goldstein A, et al. Deep phenotyping on electronic health records facilitates genetic diagnosis by clinical exomes. Am J Hum Genet. 2018;103:58–73. https://doi.org/10.1016/j.ajhg.2018.05.010.
19 Zhao M, Havrilla JM, Fang L, Chen Y, Peng J, Liu C, et al. Phen2Gene: rapid phenotype-driven gene prioritization for rare diseases. NAR Genom Bioinform. 2020;2:lqaa032. https://doi.org/10.1093/nargab/lqaa032.
20. Köhler S, Carmody L, Vasilevsky N, Jacobsen JOB, Danis D, Gourdine J-P, et al. Expansion of the Human Phenotype Ontology (HPO) knowledge base and resources. Nucleic Acids Res. 2019;47:D1018–27. https://doi.org/10.1093/nar/gky1105.
21. 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;14:51–9. https://doi.org/10.1038/gim.0b013e318232a005.
22. Louden DN. MedGen: NCBI's portal to information on medical conditions with a genetic component. Med Ref Serv Q. 2020;39:183–91. https://doi.org/10.1080/02763869.2020.1726152.
23. Landrum MJ, Lee JM, Riley GR, Jang W, Rubinstein WS, Church DM, et al. ClinVar: public archive of relationships among sequence variation and human phenotype. Nucleic Acids Res. 2014;42 Database issue:D980–5. https://doi.org/10.1093/nar/gkt1113.
24. Rubinstein WS, Maglott DR, Lee JM, Kattman BL, Malheiro AJ, Ovetsky M, et al. The NIH genetic testing registry: a new, centralized database of genetic tests to enable access to comprehensive information and improve transparency. Nucleic Acids Res. 2013;41 Database issue:D925–35. https://doi.org/10.1093/nar/gks1173.
25 Firth HV, Wright CF, DDD Study. The deciphering developmental disorders (DDD) study. Dev Med Child Neurol. 2011;53:702–3. https://doi.org/10.1111/j.1469-8749.2011.04032.x.
26. Bragin E, Chatzimichali EA, Wright CF, Hurles ME, Firth HV, Bevan AP, et al. DECIPHER: database for the interpretation of phenotype-linked plausibly pathogenic sequence and copy-number variation. Nucleic Acids Res. 2014;42 Database issue:D993–1000. https://doi.org/10.1093/nar/gkt937.
27. Turnbull C, Scott RH, Thomas E, Jones L, Murugaesu N, Pretty FB, et al. The 100 000 Genomes Project: bringing whole genome sequencing to the NHS. BMJ. 2018;361:k1687. https://doi.org/10.1136/bmj.k1687.
28. 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;36:915–21. https://doi.org/10.1002/humu.22858.
29 Gall T, Valkanas E, Bello C, Markello T, Adams C, Bone WP, et al. Defining disease, diagnosis, and translational medicine within a homeostatic perturbation paradigm: The National Institutes of Health Undiagnosed Diseases Program Experience. Front Med. 2017. https://doi.org/10.3389/fmed.2017.00062.
30 Zemojtel T, Köhler S, Mackenroth L, Jäger M, Hecht J, Krawitz P, et al. Effective diagnosis of genetic disease by computational phenotype analysis of the disease-associated genome. Sci Transl Med. 2014;6:252ra123. https://doi.org/10.1126/scitranslmed.3009262.
31. Wright CF, Fitzgerald TW, Jones WD, Clayton S, McRae JF, van Kogelenberg M, et al. Genetic diagnosis of developmental disorders in the DDD study: a scalable analysis of genome-wide research data. Lancet. 2015;385:1305–14. https://doi.org/10.1016/S0140-6736(14)61705-0.
32. Soden SE, Saunders CJ, Willig LK, Farrow EG, Smith LD, Petrikin JE, et al. Effectiveness of exome and genome sequencing guided by acuity of illness for diagnosis of neurodevelopmental disorders. Sci Transl Med. 2014;6:265ra168-265ra168. https://doi.org/10.1126/scitranslmed.3010076.
33 Rockowitz S, LeCompte N, Carmack M, Quitadamo A, Wang L, Park M, et al. Children's rare disease cohorts: an integrative research and clinical genomics initiative. npj Genom Med. 2020;5:1–12. https://doi.org/10.1038/s41525-020-0137-0.
34. Liu C, Peres Kury FS, Li Z, Ta C, Wang K, Weng C. Doc2Hpo: a web application for efficient and accurate HPO concept curation. Nucleic Acids Res. 2019;47:W566–70. https://doi.org/10.1093/nar/gkz386.
35. Taboada M, Rodriguez H, Gudivada RC, Martinez D. A new synonym-substitution method to enrich the human phenotype ontology. BMC Bioinform. 2017;18:446. https://doi.org/10.1186/s12859-017-1858-7.
36. Ehrmeyer SS. Plan for quality to improve patient safety at the point of care. Ann Saudi Med. 2011;31:342. https://doi.org/10.4103/0256-4947.83203.
37. Frey LJ, Lenert L, Lopez-Campos G. EHR big data deep phenotyping: contribution of the IMIA Genomic Medicine Working Group. Yearb Med Inform. 2014;9:206. https://doi.org/10.15265/IY-2014-0006.
38. Kohane IS. Deeper, longer phenotyping to accelerate the discovery of the genetic architectures of diseases. Genome Biol. 2014;15:115. https://doi.org/10.1186/gb4175.
39. Hughes RG, editor. Patient safety and quality: an evidence-based handbook for nurses. Rockville: Agency for Healthcare Research and Quality (US); 2011.
40. Köhler S, Vasilevsky NA, Engelstad M, Foster E, McMurry J, Aymé S, et al. The human phenotype ontology in 2017. Nucleic Acids Res. 2017;45:D865–76. https://doi.org/10.1093/nar/gkw1039.
41. Köhler S, Gargano M, Matentzoglu N, Carmody LC, Lewis-Smith D, Vasilevsky NA, et al. The human phenotype ontology in 2021. Nucleic Acids Res. 2021;49:D1207–17. https://doi.org/10.1093/nar/gkaa1043.
42. Birgmeier J, Haeussler M, Deisseroth CA, Steinberg EH, Jagadeesh KA, Ratner AJ, et al. AMELIE speeds Mendelian diagnosis by matching patient phenotype and genotype to primary literature. Sci Transl Med. 2020. https://doi.org/10.1126/scitranslmed.aau9113.
43. Javed A, Agrawal S, Ng PC. Phen-Gen: combining phenotype and genotype to analyze rare disorders. Nat Methods. 2014;11:935–7. https://doi.org/10.1038/nmeth.3046.
44. Sifrim A, Popovic D, Tranchevent L-C, Ardeshirdavani A, Sakai R, Konings P, et al. eXtasy: variant prioritization by genomic data fusion. Nat Methods. 2013;10:1083–4. https://doi.org/10.1038/nmeth.2656.
45. Hidalgo ES. Adapting the scrum framework for agile project management in science: case study of a distributed research initiative. Heliyon. 2019;5:e01447. https://doi.org/10.1016/j.heliyon.2019.e01447.
46 Maglott D, Ostell J, Pruitt KD, Tatusova T. Entrez Gene: gene-centered information at NCBI. Nucleic Acids Res. 2011;39 Database issue:D52–7. https://doi.org/10.1093/nar/gkq1237.
47 Braschi B, Denny P, Gray K, Jones T, Seal R, Tweedie S, et al. Genenames.org: the HGNC and VGNC resources in 2019. Nucleic Acids Res. 2019;47:D786–92. https://doi.org/10.1093/nar/gky930.
48. Jonquet C, Lependu P, Falconer S, Coulet A, Noy NF, Musen MA, et al. NCBO resource index: ontology-based search and mining of biomedical resources. Web Semant. 2011;9:316–24. https://doi.org/10.1016/j.websem.2011.06.005.
49. 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;83:610–5. https://doi.org/10.1016/j.ajhg.2008.09.017.
50. Yu AC-S, Chan AY-Y, Au WC, Shen Y, Chan TF, Chan H-YE. Whole-genome sequencing of two probands with hereditary spastic paraplegia reveals novel splice-donor region variant and known pathogenic variant in SPG11. Cold Spring Harb Mol Case Stud. 2016;2:a001248. https://doi.org/10.1101/mcs.a001248.
51. McKusick VA. Mendelian inheritance in man and its online version. OMIM Am J Hum Genet. 2007;80:588–604. https://doi.org/10.1086/514346.
52. Elkin PL, Brown SH, Husser CS, Bauer BA, Wahner-Roedler D, Rosenbloom ST, et al. Evaluation of the content coverage of SNOMED CT: ability of

SNOMED clinical terms to represent clinical problem lists. Mayo Clin Proc. 2006;81:741–8. https://doi.org/10.4065/81.6.741.
53. Lipscomb CE. Medical subject headings (MeSH). Bull Med Libr Assoc. 2000;88:265.
54. World Health Organization. International statistical classification of diseases and related health problems: tabular list. World Health Organization; 2004. https://play.google.com/store/books/details?id=Tw5eAtsatiUC.
55. Weinreich SS, Mangon R, Sikkens JJ, Teeuw ME, Cornel MC. Orphanet: a European database for rare diseases. Ned Tijdschr Geneeskd. 2008;152:518–9.
56 Schriml LM, Arze C, Nadendla S, Chang Y-WW, Mazaitis M, Felix V, et al. Disease ontology: a backbone for disease semantic integration. Nucleic Acids Res. 2012;40 Database issue:D940–6. https://doi.org/10.1093/nar/gkr972.

Publisher's Note