

RESEARCH

Open Access



Validation of a computational phenotype for finding patients eligible for genetic testing for pathogenic *PTEN* variants across three centers

Cartik Kothari^{1†}, Siddharth Srivastava^{2†}, Youssef Kousa^{3,4}, Rima Izem⁵, Marcin Gierdalski⁶, Dongkyu Kim⁶, Amy Good⁷, Kira A. Dies², Gregory Geisel², Hiroki Morizono^{8,9}, Vittorio Gallo¹⁰, Scott L. Pomeroy¹¹, Gwenn A. Garden^{12,13}, Lisa Guay-Woodford¹⁴, Mustafa Sahin² and Paul Avillach^{1*} 

Abstract

Background: Computational phenotypes are most often combinations of patient billing codes that are highly predictive of disease using electronic health records (EHR). In the case of rare diseases that can only be diagnosed by genetic testing, computational phenotypes identify patient cohorts for genetic testing and possible diagnosis. This article details the validation of a computational phenotype for *PTEN* hamartoma tumor syndrome (PHTS) against the EHR of patients at three collaborating clinical research centers: Boston Children's Hospital, Children's National Hospital, and the University of Washington.

Methods: A combination of billing codes from the International Classification of Diseases versions 9 and 10 (ICD-9 and ICD-10) for diagnostic criteria postulated by a research team at Cleveland Clinic was used to identify patient cohorts for genetic testing from the clinical data warehouses at the three research centers. Subsequently, the EHR—including billing codes, clinical notes, and genetic reports—of these patients were reviewed by clinical experts to identify patients with PHTS.

Results: The *PTEN* genetic testing yield of the computational phenotype, the number of patients who needed to be genetically tested for incidence of pathogenic *PTEN* gene variants, ranged from 82 to 94% at the three centers.

Conclusions: Computational phenotypes have the potential to enable the timely and accurate diagnosis of rare genetic diseases such as PHTS by identifying patient cohorts for genetic sequencing and testing.

Keywords: Rare disease, Genetic disease, Computational phenotype, Electronic health records, Autism

Background

Computational phenotypes for rare diseases

Computational phenotypes [1, 2] are combinations of clinical billing and diagnostic codes that are highly

indicative of disease and are identified either manually [3–6] or by machine learning algorithms from patients' electronic health records (EHR) [7–13]. The Phenotype Knowledge Base (PheKB, <http://phekb.org>) lists computational phenotypes for over fifty diseases with more submissions under review. The identification of computational phenotypes or “phenotyping” is reliant on the availability of large patient cohorts, common knowledge

*Correspondence: paul_avillach@hms.harvard.edu

†Cartik Kothari and Siddharth Srivastava are co-first authors.

¹ Department of Biomedical Informatics, Harvard Medical School, Boston, MA 02115, USA

Full list of author information is available at the end of the article



about disease symptoms, and standardized codes for clinical diagnoses, procedures, and lab tests.

In the case of rare diseases [14] where patient populations are small, knowledge of the breadth of patient symptoms can be limited [15–17], and diagnosis may depend on genetic testing, computational phenotypes serve two different purposes. They (1) enable the identification of patients who may have a suspected genetic disorder and who could be referred for appropriate confirmatory genetic testing and (2) reveal previously undiscovered patterns of clinical comorbidities that can enhance the clinical characterization of the disease [9].

***PTEN* hamartoma tumor syndrome**

PTEN hamartoma tumor syndrome (PHTS) [18] is a rare genetic disorder, which encompasses four major clinically distinct syndromes: (a) Cowden syndrome (CS; OMIM: 615107, 615108, 615109) [19, 20], (b) Bannayan-Riley-Ruvalcaba syndrome (OMIM: 158350) [21], (c) Proteus syndrome (OMIM: 176920) [22], and (d) Proteus-like syndrome. All the disorders are associated with germline pathogenic variants of the phosphatase and tensin homolog (*PTEN*) tumor suppressor gene (NCBI Gene ID: 5728; HGNC ID: 9588), located on the long arm of chromosome 10. The clinical manifestations of PHTS are diverse and constitute a wide spectrum of neurological and developmental characteristics (e.g., macrocephaly, intellectual disability, autism spectrum disorder, attention deficit hyperactivity disorder, and anxiety), gastrointestinal manifestations (e.g., gastrointestinal polyps), vascular and nonvascular skin findings (e.g., arteriovenous malformations, hemangiomas, trichilemmomas, acral keratoses, lipomas, fibromas), and oncological concerns (e.g., increased risk for various cancers including thyroid cancer, breast cancer, colorectal cancer, renal cancer, endometrial cancer, and melanoma) [18]. Some of these symptoms in isolation (such as thyroid cancer) may be found in the general population. Because of this, patients may be undiagnosed and thus do not benefit from available cancer surveillance strategies. The timely recognition of the phenotypic patterns typical to PHTS is therefore critical to patient outcomes [18].

The definitive diagnosis of PHTS is based on detection of a pathogenic germline variant in the *PTEN* gene. The

presence of specific clinical features may pinpoint a need for molecular testing. In 2011, Cleveland Clinic established criteria for pathogenic *PTEN* variant screening in children (https://www.lerner.ccf.org/gmi/ccscore/documents/pediatric_criteria.html; hereon referred to as Cleveland Clinic criteria for *PTEN* testing or Cleveland Clinic criteria) [23]. The objective of this study is to determine the effectiveness of a computational phenotype of the Cleveland Clinic criteria in finding patients who need to be genetically tested for pathogenic *PTEN* variants, using data from EHR.

Outline

In this paper, we describe a cross-institutional initiative among three participating clinical research centers: (a) Boston Children's Hospital (BCH), (b) Children's National Hospital (CNH), and (c) the University of Washington (UW), to evaluate the predictive power of a computational phenotype for PHTS in identifying patients requiring genetic testing for diagnosis of *PTEN* syndrome. This initiative was coordinated by the Intellectual and Developmental Disabilities Research Centers at each of the institutions (IDDRC, <https://www.iddrc.org/>).

Methods

We used a workflow adopted for the evaluation of the predictive power of a computational phenotype for PHTS (Fig. S1).

Data

The data used in this study are complete patient electronic health records (EHR)—comprising clinical notes, genetic reports, and billing codes—sourced from the clinical data warehouses at the three centers: BCH, CNH, and UW. The patient cohorts were identified by querying the clinical data warehouses for patients with the criteria in Table 1.

The criteria above were proposed by a team of researchers at Cleveland Clinic after evaluation of a cohort of pediatric individuals with *PTEN* mutations [23] and will be referred to henceforth in this paper as the Cleveland Clinic pediatric clinical criteria or simply as the Cleveland Clinic criteria. A clinical expert identified the billing codes from the International Classification of Diseases

Table 1 Cleveland Clinic criteria for identifying pediatric patients who would benefit from *PTEN* sequencing

1. Macrocephaly (≥ 2 Standard Deviations from Mean)
AND
2. At least one of the following four additional criteria
A. Autism or Developmental Delay
B. Dermatologic features, including lipomas, trichilemmomas, oral papillomas, or penile freckling
C. Vascular features, such as arteriovenous malformations or hemangiomas
D. Gastrointestinal Polyyps

versions 9 and 10 (ICD-9 and ICD-10) [24, 25] that correspond to the conditions in the Cleveland Clinic criteria. The list of the identified billing codes can be found in Table S1. The sizes of the patient cohorts identified by the Cleveland Clinic criteria at the three centers are shown in Table 2.

The Institutional Review Board (IRB) at Boston Children's Hospital served as the single IRB with reliance agreements and approved this study (P00029725). The clinical data warehouses at the three participating centers were queried for patients whose clinical visits were assigned a combination of ICD-9 and ICD-10 codes that satisfied the Cleveland Clinic criteria. The complete EHR of these patients—comprising clinical notes, genetic reports, and billing codes—were extracted. At each site, the charts of a subset of these patients were reviewed by a team of clinical experts from that site in order to determine (A) if that patient indeed met Cleveland Clinic criteria, (B) if that patient had any genetic testing, (C) if the genetic testing included *PTEN* sequencing and/or deletion duplication analysis, and (D) if there was a likely pathogenic or pathogenic variant detected in *PTEN*.

Determination of whether the patient satisfied the Cleveland Clinic criteria

The presence of macrocephaly was assumed to be true for all patients due to inconsistent documentation about head circumference or inability to ascertain this clinical feature. To determine if each patient satisfied Cleveland Clinic (CC) criteria (i.e., if the patient had *at least* one of the four additional clinical features mentioned in the criteria), the reviewing team iteratively evaluated each patient record to be reviewed using a protocol (detailed in [Supplementary Methods](#) under Protocol for determination of whether patient satisfied Cleveland Clinic criteria)

Determination of whether the patient had genetic testing

To determine if the patient had genetic testing, the reviewing team followed a protocol (detailed in [Supplementary Methods](#) under "Protocol for determination of whether patient had genetic testing").

Table 2 Number of patients identified as having met Cleveland Clinic criteria using informatics approach across the three sites

	Boston Children's Hospital From January 2001 to October 2019	University of Washington January 2001 to August 2019	Children's National Hospital From August 2012 to October 2019
Total patients with at least one ICD-9 or ICD-10 code available for review	1.78 M	2.7 M	2.11 M
Number of patients identified by informatics approach as having met Cleveland Clinic criteria	1,215	104	481
Number of patients identified by informatics approach as having met Cleveland Clinic criteria whose charts were manually reviewed	396	94	481
Number of patients who had any genetic testing (from among those whose charts were manually reviewed)	204	29	227
Number of patients who had any genetic testing which included <i>PTEN</i> sequencing and/or deletion duplication analysis (from among those whose charts were manually reviewed)	90	17	43
Number of patients who satisfied Cleveland Clinic criteria after human review (from among those whose charts were manually reviewed)	371	77	438
Number of patients with pathogenic or likely pathogenic variant in <i>PTEN</i> (from among those whose charts were manually reviewed)	14	0	13
Yield of the informatics approach in identifying patients who meet Cleveland Clinic criteria (from among those whose charts were manually reviewed)	371/396 (94%)	77/94 (82%)	438/481 (91%)
Number of patients with PHTS divided by number of those identified as having met Cleveland Clinic criteria using informatics approach whose charts were manually reviewed	3.54% (14/396)	0 (0/94)	2.70% (13/481)
Number of patients with PHTS divided by number of those identified as having met Cleveland Clinic criteria using informatics approach and who also had genetic testing which included detection of <i>PTEN</i> variants (from among those whose charts were manually reviewed)	15.6% (14/90)	0% (0/17)	30.2% (13/43)

Determination of whether genetic testing included *PTEN* analysis

If the patient had genetic testing, the reviewing team reviewed the list of genetic tests to identify inclusion of *PTEN* analysis. The following tests were among those automatically deemed to include *PTEN* analysis: *PTEN* single gene sequencing, *PTEN* deletion/duplication analysis, and whole exome sequencing. For gene panels, the team manually reviewed the report if available to determine if *PTEN* was included in the panel. If the report was not available, then the testing laboratory's website was queried to see if *PTEN* was part of the panel. See Fig. S2.

Determination of whether the patient had a pathogenic or likely pathogenic variant in *PTEN*

If the patient had genetic testing that included *PTEN* sequencing, the team reviewed the original report, or references to the test results, to see if there was a reported likely pathogenic or pathogenic variant in *PTEN*. If so, the patient was deemed to *have PHTS*. Otherwise, the patient was deemed to *not have PHTS*.

Results

The yield—the number of patients who needed to be genetically tested for a pathogenic *PTEN* variant—of the Cleveland Clinic criteria ranged from 82 to 94% at the three centers (Table 2).

Review of yield of informatics approach

Boston Children's Hospital (BCH)

With this informatics approach, there were 1215 patients at Boston Children's Hospital identified as having met Cleveland Clinic criteria. Human review of clinical documentation of 396 randomly selected patients was performed. Of these 396 patients, 371 patients did indeed satisfy Cleveland Clinic criteria (see Table 1). For the BCH site, the yield of this informatics approach in correctly identifying patients who met Cleveland Clinic criteria was 93.69%.

Children's National Hospital (CNH)

With this informatics approach, there were 481 patients at Children's National Hospital identified as having met Cleveland Clinic criteria. Human review of clinical documentation of all of these patients identified 438 patients as having truly met Cleveland Clinic criteria. For the CNH site, the yield of this informatics approach in correctly identifying patients who met Cleveland Clinic criteria was 91.06%.

University of Washington (UW)

At the University of Washington, 94 patients were randomly selected for human review, out of the 104 patients

who satisfied the Cleveland Clinic criteria using the informatics approach. After human review, 77 out of the 94 patients indeed satisfied the Cleveland Clinic criteria, resulting in a yield of 81.91%.

Review of genetic testing

We also evaluated the number of patients who had molecular confirmation of the PHTS diagnosis. Among those patients who met Cleveland Clinic criteria identified by this informatics approach, and whose charts were reviewed, the percentage of patients with a molecular diagnosis of PHTS was 0.0% at UW, 2.7% at CNH, and 3.5% at BCH. Among those patients who met Cleveland Clinic criteria identified by this informatics approach, whose charts were reviewed, and who also had any genetic testing done which would have captured *PTEN* variants, this percentage is higher: 30.2% at CNH and 15.6% at BCH (Table 2).

Discussion

Conditions associated with rare genetic diseases are largely underrepresented [26, 27] in commonly used clinical terminologies such as the ICD-10 and ICD-9. The problem persists in the latest version of the International Classification of Diseases (ICD-11) terminology [28], where conditions associated with genetic diseases are either categorized in counterintuitive ways, too broadly generalized, or not defined at all [29]. In this work, we have demonstrated the feasibility of using a computational phenotype across multiple institutions to identify patients who satisfy Cleveland Clinic criteria and who may therefore benefit from *PTEN* genetic analysis. The positive predictive value of this approach at each of the three sites exceeded 80%, suggesting that an informatics approach may be able to bypass the shortcoming of the ICD9/10 code system in explicitly including "PTEN hamartoma tumor syndrome."

We also evaluated the percentage of patients who were correctly identified as having PHTS, out of the total number of patients identified as satisfying Cleveland Clinic criteria through this informatics approach. While this number was low across the three sites—between 0 and 3.5%—several factors account for this. First, the number of patients with PHTS identified may reflect the very low prevalence of PHTS, which according to one estimate is 1:200,000 [30]. Second, these percentages do not take into account those who did not undergo any genetic testing in the first place. Third, not every patient who underwent genetic testing had genetic testing that included *PTEN* sequencing.

These percentages become higher (15.6%, 30.2%) when the denominator is further limited by those who have undergone genetic testing which would have captured

PTEN variants. It is worthwhile to compare this higher range of percentages (i.e., *PTEN* molecular diagnosis among those identified by informatics approach who had genetic testing that included detection of *PTEN* variants) to clinical scenarios reported in prior studies of diagnostic yield of *PTEN* testing in different cohorts. For example, in the original data serving as the basis for the Cleveland Clinic pediatric *PTEN* criteria, there were 92 pediatric patients who met relaxed International Cowden Consortium operational criteria for CS [23], of whom 28 had a *PTEN* mutation (30.4%). In a retrospective study of the percentage of patients with a confirmed *PTEN* mutation among different pediatric cohorts, 2/14 (14.2%) had PHTS among those with ASD and macrocephaly, 3/13 (23.1%) had PHTS among those with ASD and developmental delay/ID and macrocephaly, and 6/32 (18.8%) had PHTS among those with developmental delay/ID and macrocephaly [31]. Hence, the informatics approach used in our study not only shows promise in identifying those who may meet Cleveland Clinic *PTEN* criteria but also underscores that there were many patients who may have benefited from genetic testing but who did not actually undergo genetic testing. This is evident by the large percentage of patients in our study identified by informatics approach as having met Cleveland Clinic *PTEN* criteria, who either did not have genetic testing or had genetic testing which did not include analysis of *PTEN* variants.

The approach taken here across three academic research centers can be used at several other institutions around the country in the future to identify patients that would benefit from *PTEN* sequencing. Furthermore, similar computational phenotypes can be developed and tested for other rare genetic disorders. For example, if a clinician is evaluating a patient for whom only electronic health records are available, the use of a computational phenotype could help delineate a phenotype caused by a particular gene defect.

Limitations

Limitations in this informatics approach for detecting patients who met Cleveland Clinic criteria for *PTEN* testing are evident in the instances of false positives, that is, those who met Cleveland Clinic criteria by the informatics approach but who on review of the medical records did not actually meet Cleveland Clinic criteria. A large contributing factor is that the billing codes may not accurately or completely encompass the clinical phenotype. In addition, there may be inaccuracies in the billing codes. For instance, in some cases, providers coded patients as having developmental delay, when the clinical documentation specifically mentioned “normal development.” There can be a mismatch in actual clinical information vs. intention behind billed ICD codes.

For example, there was an instance in which a patient postoperatively lost speech but regained this ability later on. The provider coded this as expressive language disorder, perhaps because another more suitable billing code was not identifiable.

Coding systems such as ICD-10 and ICD-9 were developed primarily for administrative purposes [32]. Given the lack of precise clinical codes for genetic diseases and their symptoms, errors in coding can be difficult to avoid [33]. Studies have revealed widespread inconsistencies in the precision of billing codes in capturing clinical symptoms [34, 35]. In other words, though it is feasible to use billing codes to ascertain Cleveland Clinic criteria, there is a need for improved precision of clinical codes in capturing clinical phenotype diversity to address this limitation. Deep phenotyping [36, 37], using finer-grained representations of disease phenotypes as defined in terminologies such as the Human Phenotype Ontology (HPO) [38] and SNOMED CT [39], is essential for precise characterization and phenome-based diagnosis of rare diseases such as PHTS.

There were several additional limitations. First, we did not analyze whether patients identified as having met Cleveland Clinic criteria, and whose charts were reviewed and confirmed to meet Cleveland Clinic criteria, reported another clinical reason to suspect a diagnosis other than PHTS. Second, we did not ascertain whether macrocephaly was truly present, due to inconsistent availability and documentation of head circumference. This may help account for the low fraction of individuals who fulfill Cleveland Clinic criteria who have pathogenic *PTEN* variants. For example, at the BCH site, we identified an example of one patient with PHTS with macrocephaly and related dermatological findings who would have fulfilled Cleveland Clinic criteria, but macrocephaly was not billed as a diagnosis. Third, we did not limit EMR data to that *prior* to the diagnosis (given that a patient diagnosis would influence what clinical features are referenced in the notes), since it was not straightforward to ascertain age of diagnosis (though report date is one possibility, patient knowledge and provider knowledge of this diagnosis may lag). Finally, we did not have the data to evaluate race/ethnicity/social vulnerability index. On review of data from the BCH site, nearly 60% of the patients identified as having met Cleveland Clinic criteria using the informatics approach were white, suggesting that minorities were underrepresented, which limits generalizability. This point underscores continued need for attention to inclusion and diversity in ongoing research efforts, especially to the question of why minorities are underrepresented in research databases and clinical encounters.

Conclusions

Computational phenotypes have the potential to greatly reduce the difficulties in diagnosing rare genetic disorders by identifying patient cohorts for genetic testing and also to enhance the clinical characterization of these diseases. In this paper, we have discussed the evaluation and effectiveness of a computational phenotype in identifying patients who need to be genetically screened for pathogenic *PTEN* variants from the EHR of patients. The observed yield of this computational phenotype results from the following: (A) the lack of emphasis on fine-grained representation of clinical symptoms in billing codes used at healthcare centers, (B) the slow pace of adoption of diagnostic methods based upon genetic testing into clinical practice, and (C) the limited understanding of the phenotypic diversity of genetic diseases.

However, the availability of genomic and phenotypic data from significantly larger patient populations and improvements in the representational capabilities of clinical terminologies in the long-term will greatly facilitate the drive towards precise clinical characterization of PHTS and its symptoms.

Abbreviations

PTEN: Phosphatase and tensin homolog; PHTS: *PTEN* hamartoma syndrome; IDDC: Intellectual and Developmental Disabilities Research Centers; BCH: Boston Children's Hospital; CNH: Children's National Hospital; EHR: Electronic health records; OMIM: Online Mendelian Inheritance in Man; HGNC: HUGO Gene Nomenclature Committee; NCBI: National Center for Biotechnology Information; ICD-9: International Classification of Diseases version 9; ICD-10: International Classification of Diseases version 10; UW: University of Washington; CC: Cleveland Clinic; IRB: Institutional Review Board.

Supplementary Information

The online version contains supplementary material available at <https://doi.org/10.1186/s11689-022-09434-0>.

Additional file 1: Figure S1. Workflow for the identification and validation of a computational phenotype for identifying patients who meet Cleveland Clinic criteria.

Additional file 2: Figure S2. Workflow for the review of genetic testing to determine if patients had a pathogenic variant in the *PTEN* gene that would confirm a molecular diagnosis of PHTS.

Additional file 3: Supplementary Methods. Protocol for determination of whether patient satisfied Cleveland Clinic criteria.

Additional file 4: Table S1. Billing Codes from ICD-9 and ICD-10 for Cleveland Clinic criteria for PHTS.

Acknowledgements

Not applicable.

Authors' contributions

1. CK ran the computational phenotype queries on the clinical data warehouse at Boston Children's Hospital, coordinated the testing of the computational phenotype at Children's National Hospital and at the University of Washington and the collection of the patient counts for the tables, generated all the figures and the tables, and drafted and edited the manuscript. 2. SS provided clinical input, identified the ICD-9 and ICD-10 codes for the

Cleveland Clinic criteria, standardized chart review procedures for all the centers, performed chart reviews, and edited the manuscript. 3. YK contributed to study design, provided clinical input, performed chart reviews, coordinated the work at Children's National Hospital, contributed to data analysis, and edited the manuscript. 4. RI created a data collection sheet in REDCap for those manually reviewing the charts at Children's National Hospital and evaluating them against Cleveland Clinic criteria for *PTEN* testing. She collated and reviewed summary data from manual chart reviews to derive flow charts and positive predictive values. She answered clarifying questions about the data and provided statistical expertise. 5. MG performed queries on the Children's National Hospital enterprise data warehouse and cross-referenced persons of interest with the Children's National Hospital EHR. 6. DK provided the access methodology and the mapping logic for EHR data from the Children's National Hospital enterprise data warehouse. 7. AG reviewed, abstracted, and compiled patient data from the University of Washington Medical Center enterprise data warehouse. 8. KAD served as the regulatory manager for the study, managed data collection and analysis at BCH, and edited the manuscript. 9. GG assisted with chart reviews and served as study coordinator at Boston Children's Hospital, the central IRB. 10. HM performed queries on the Children's National Hospital enterprise data warehouse and edited the manuscript. 11. VG coordinated the effort at Children's National Hospital as Director of the DC-IDDC. 12. SP coordinated the effort at Boston Children's Hospital as Director of the Harvard IDDC. 13. GAG coordinated the effort at the University of Washington as Director of the Clinical Translational Core of the UW IDDC. 14. LGW coordinated the effort at Children's National Hospital as Director of the DC-IDDC Clinical Translational Core. 15. MS conceived and designed the study, supervised the data analysis, and edited the manuscript. 16. PA conceived and designed the study, supervised the data analysis, and drafted and edited the manuscript. All authors approved the final version of the manuscript.

Funding

Boston Children's Hospital

This publication was supported by Boston Children's Hospital IDDC Clinical Translational Core, funded by NIH U54 HD090255.

Children's National Hospital

This publication was supported by the District of Columbia Intellectual and Developmental Disabilities Research Center (DC-IDDC; U54HD090257) and Clinical and Translational Science Award Number UL1TR001876 from the NIH National Center for Advancing Translational Sciences. Its contents are solely the responsibility of the authors and do not necessarily represent the official views of the National Center for Advancing Translational Sciences or the National Institutes of Health. SS receives support from the National Institute of Health/National Institute of Neurological Disorders and Stroke (NIH/NINDS) (1K23NS119666).

University of Washington Medicine

This publication was supported by the University of Washington IDDC Clinical Translational Core funded by U54 HD083091 and by the National Center for Advancing Translational Sciences of the National Institutes of Health under Award Number UL1 TR002319. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

Availability of data and materials

The datasets used in this study are stored in the clinical data warehouses at the three centers: Boston Children's Hospital, Children's National Hospital, and the University of Washington. Because of patient privacy concerns, this data is not available to the general public. Requests for this data will need to be formally submitted to the Information Technology teams at these centers, and IRB clearances from the appropriate organizations will be needed as well. The code developed for this paper is available on [GitHub](#). The ICD-9 and ICD-10 codes corresponding to the Cleveland Clinic criteria, which were used for searching the clinical data warehouses at Boston Children's Hospital, Children's National Hospital, and the University of Washington are listed in Table S1.

Declarations

Ethics approval and consent to participate

Boston Children's Hospital IRB-P00029725.

Consent for publication

Not applicable.

Competing interests

1. CK has no competing interests to declare.
2. SS has received consulting fees from GLG, Guidepoint (which connected to a client, Fortress Biotech), and Novartis.
3. YK has no competing interests to declare.
4. RI has no competing interests to declare.
5. MG has no competing interests to declare.
6. DK has no competing interests to declare.
7. AG has no competing interests to declare.
8. KAD has no competing interests to declare.
9. GG has no competing interests to declare.
10. HM has no competing interests to declare.
11. VG has no competing interests to declare.
12. SP has no competing interests to declare.
13. GAG has no competing interests to declare.
14. LGW has no competing interests to declare.
15. MS reports grant support from Novartis, Roche, Biogen, Astellas, Aeovian, Bridgebio, Aucta, and Quadrant Biosciences. He has served on Scientific Advisory Boards for PTEN Research, Novartis, Roche, Celgene, Regenxbio, Alkermes, and Takeda.
16. PA has no competing interests to declare

Author details

¹Department of Biomedical Informatics, Harvard Medical School, Boston, MA 02115, USA. ²Department of Neurology, Rosamund Stone Zander Translational Neuroscience Center, Boston Children's Hospital, Harvard Medical School, Boston, MA 02115, USA. ³Division of Neurology, Children's National Hospital, Washington, DC 20010, USA. ⁴Department of Genomics and Precision Medicine, The George Washington University School of Medicine and Health Sciences, Washington, DC 20052, USA. ⁵Division of Biostatistics and Study Methodology, Children's National Research Institute, Silver Spring, MD 20910, USA. ⁶Division of Biostatistics and Study Methodology, Children's National Hospital, Washington, DC 20010, USA. ⁷Institute for Translational Health Sciences, University of Washington, Seattle, WA 98195, USA. ⁸Center for Genetic Medicine Research, Children's National Hospital, Washington, DC 20010, USA. ⁹Department of Genomics and Precision Medicine, The George Washington University School of Medicine and Health Sciences, Washington, DC 20052, USA. ¹⁰Center for Neuroscience Research, Children's National Research Institute, Children's National Hospital, Washington, DC 20010, USA. ¹¹Department of Neurology, Boston Children's Hospital, Harvard Medical School, Boston, MA 02115, USA. ¹²Department of Neurology and Center on Human Development and Disability, University of Washington, Seattle, WA 98195, USA. ¹³Department of Neurology, University of North Carolina, Chapel Hill, NC 27599, USA. ¹⁴Center for Translational Research, Children's National Hospital, Washington, DC 20010, USA.

Received: 10 April 2021 Accepted: 4 March 2022

Published online: 23 March 2022

References

1. Kirby JC, Speltz P, Rasmussen LV, et al. PheKB: a catalog and workflow for creating electronic phenotype algorithms for transportability. *J Am Med Inform Assoc.* 2016;23:1046–52.
2. Che Z, Kale D, Li W, Taha Bahadori M, Liu Y. Deep computational phenotyping. Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. <https://doi.org/10.1145/2783258.2783365>.
3. Pathak J, Wang J, Kashyap S, Basford M, Li R, Masys DR, et al. Mapping clinical phenotype data elements to standardized metadata repositories and controlled terminologies: the eMERGE Network experience. *J Am Med Inform Assoc.* 2011;18:376–86.
4. Conway M, Berg RL, Carrell D, et al. Analyzing the heterogeneity and complexity of electronic health record oriented phenotyping algorithms. *AMIA Annu Symp Proc.* 2011;2011:274–83.
5. Newton KM, Peissig PL, Kho AN, et al. Validation of electronic medical record-based phenotyping algorithms: results and lessons learned from the eMERGE network. *J Am Med Inform Assoc.* 2013;20:e147–54.
6. McCarty CA, Chisholm RL, Chute CG, et al. The eMERGE Network: a consortium of biorepositories linked to electronic medical records data for conducting genomic studies. *BMC Med Genomics.* 2011;4:13.
7. Liao KP, Cai T, Savova GK, et al. Development of phenotype algorithms using electronic medical records and incorporating natural language processing. *BMJ.* 2015;350:h1885.
8. Pathak J, Kho AN, Denny JC. Electronic health records-driven phenotyping: challenges, recent advances, and perspectives. *J Am Med Inform Assoc.* 2013;20:e206–11.
9. Kohane IS. Using electronic health records to drive discovery in disease genomics. *Nat Rev Genet.* 2011;12:417–28.
10. Hodapp C. Unsupervised learning for computational phenotyping. *arXiv [stat.ML];* 2016.
11. Marlin BM, Kale DC, Khemani RG, Wetzel RC. Unsupervised pattern discovery in electronic health care data using probabilistic clustering models. Proceedings of the 2nd ACM SIGHIT International Health Informatics Symposium. <https://dl.acm.org/doi/10.1145/2110363.2110408>. Accessed 30 Jan 2020.
12. Lasko TA, Denny JC, Levy MA. Computational phenotype discovery using unsupervised feature learning over noisy, sparse, and irregular clinical data. *PLoS One.* 2013;8:e66341.
13. Yu S, Liao KP, Shaw SY, Gainer VS, Churchill SE, Szolovits P, et al. Toward high-throughput phenotyping: unbiased automated feature extraction and selection from knowledge sources. *J Am Med Inform Assoc.* 2015;22:993–1000.
14. Jackson M, Marks L, May GHW, Wilson JB. The genetic basis of disease. *Essays Biochem.* 2018;62:643–723.
15. Zurynski Y, Frith K, Leonard H, Elliott E. Rare childhood diseases: how should we respond? *Arch Dis Child.* 2008;93:1071–4.
16. Kliegman RM, Bordini BJ, Basel D, Nocton JJ. How doctors think: common diagnostic errors in clinical judgment—lessons from an undiagnosed and rare disease program. *Pediatr Clin North Am.* 2017;64:1–15.
17. Lopes MT, Koch VH, Sarrubbi-Junior V, Gallo PR, Carneiro-Sampaio M. Difficulties in the diagnosis and treatment of rare diseases according to the perceptions of patients, relatives and health care professionals. *Clinics.* 2018;73:e68.
18. Hobert JA, Eng C. PTEN hamartoma tumor syndrome: an overview. *Genet Med.* 2009;11:687–94.
19. Orloff MS, He X, Peterson C, Chen F, Chen J-L, Mester JL, et al. Germline PIK3CA and AKT1 mutations in Cowden and Cowden-like syndromes. *Am J Hum Genet.* 2013;92:76–80.
20. Bennett KL, Mester J, Eng C. Germline epigenetic regulation of KILLIN in Cowden and Cowden-like syndrome. *JAMA.* 2010;304:2724–31.
21. Arch EM, Goodman BK, Van Wesepe RA, Liaw D, Clarke K, Parsons R, et al. Deletion of PTEN in a patient with Bannayan-Riley-Ruvalcaba syndrome suggests allelism with Cowden disease. *Am J Med Genet.* 1997;71:489–93.
22. Ou M, Sun Z, Zhu P, Sun G, Dai Y. Proteus syndrome: a case report and review of the literature. *Mol Clin Oncol.* 2017;6:381–3.
23. Tan M-H, Mester J, Peterson C, et al. A clinical scoring system for selection of patients for PTEN mutation testing is proposed on the basis of a prospective study of 3042 probands. *Am J Hum Genet.* 2011;88:42–56.
24. ICD - ICD-9-CM - International Classification of Diseases, Ninth Revision, Clinical Modification. 2019. <https://www.cdc.gov/nchs/icd/icd9cm.htm>. Accessed 31 Jan 2020.
25. ICD - ICD-10-CM - International Classification of Diseases, Tenth Revision, Clinical Modification. 2020. <https://www.cdc.gov/nchs/icd/icd10cm.htm>. Accessed 31 Jan 2020.
26. Kodra Y, Fantini B, Taruscio D. Classification and codification of rare diseases. *J Clin Epidemiol.* 2012;65:1026–7.
27. van Karnebeek CDM, Beumer D, Pawliuk C, Goetz H, Mostafavi S, Andrews G, et al. A novel classification system for research reporting in rare and progressive genetic conditions. *Dev Med Child Neurol.* 2019;61:1208–13.
28. ICD-11. <https://icd.who.int/en>. Accessed 6 Feb 2020.
29. Aymé S, Bellet B, Rath A. Rare diseases in ICD11: making rare diseases visible in health information systems through appropriate coding. *Orphanet J Rare Dis.* 2015;10:35.

30. Nelen MR, Kremer H, Konings IB, et al. Novel PTEN mutations in patients with Cowden disease: absence of clear genotype-phenotype correlations. *Eur J Hum Genet.* 1999;7:267–73.
31. Varga E, Pastore M, Prior T, et al. The prevalence of PTEN mutations in a clinical pediatric cohort with autism spectrum disorders, developmental delay, and macrocephaly. *Genet Med.* 2009;11:111–7.
32. Ladha KS, Eikermann M. Codifying healthcare—big data and the issue of misclassification. *BMC Anesthesiol.* 2015;15:179.
33. Tran J, Cennimo D, Chen S, Altschuler EL. Teaching billing and coding to medical students: a pilot study. *Med Educ Online.* 2013;18:21455.
34. Horsky J, Drucker EA, Ramelson HZ. Accuracy and completeness of clinical coding using ICD-10 for ambulatory visits. *AMIA Annu Symp Proc.* 2017;2017:912–20.
35. Fawcett N, Young B, Peto L, et al. “Caveat emptor”: the cautionary tale of endocarditis and the potential pitfalls of clinical coding data—an electronic health records study. *BMC Med.* 2019;17:169.
36. Delude CM. Deep phenotyping: the details of disease. *Nature.* 2015;527:S14–5.
37. Robinson PN. Deep phenotyping for precision medicine. *Hum Mutat.* 2012;33:777–80.
38. 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.
39. SNOMED Home page. In: SNOMED. <http://www.snomed.org/>. Accessed 6 Feb 2020.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Ready to submit your research? Choose BMC and benefit from:

- fast, convenient online submission
- thorough peer review by experienced researchers in your field
- rapid publication on acceptance
- support for research data, including large and complex data types
- gold Open Access which fosters wider collaboration and increased citations
- maximum visibility for your research: over 100M website views per year

At BMC, research is always in progress.

Learn more biomedcentral.com/submissions

