Published in final edited form as:

J Pediatr. 2017 September ; 188: 224–231.e5. doi:10.1016/j.jpeds.2017.05.037.

A Computable Phenotype Improves Cohort Ascertainment in a Pediatric Pulmonary Hypertension Registry

Alon Geva, MD, MPH^{1,2,9}, Jessica L. Gronsbell, BA³, Tianxi Cai, ScD³, Tianrun Cai, MD⁴, Shawn N. Murphy, MD, PhD^{5,6,10}, Jessica C. Lyons, MS¹⁰, Michelle M. Heinz, BS¹, Marc D. Natter, MD^{1,11}, Nandan Patibandla, MS⁷, Jonathan Bickel, MD, MS^{1,7,11}, Mary P. Mullen, MD, PhD^{8,11}, Kenneth D. Mandl, MD, MPH^{1,10,11}, and for the PPHNet and NHLBI Pediatric Pulmonary Vascular Disease Outcomes Bioinformatics Clinical Coordinating Center Investigators*

¹Computational Health Informatics Program, Boston Children's Hospital, Boston, MA

²Division of Critical Care Medicine, Department of Anesthesiology, Perioperative, and Pain Medicine, Boston Children's Hospital, Boston, MA

³Department of Biostatistics, Harvard T. H. Chan School of Public Health, Boston, MA

⁴Division of Rheumatology, Immunology and Allergy, Brigham and Women's Hospital, Boston, MA

⁵Department of Research Information Services and Computing, Partners Healthcare, Boston, MA

⁶Department of Neurology, Massachusetts General Hospital, Boston, MA

⁷Information Services Department, Boston Children's Hospital, Boston, MA

⁸Department of Cardiology, Boston Children's Hospital, Boston, MA

⁹Department of Anaesthesia, Harvard Medical School, Boston, MA

¹⁰Department of Biomedical Informatics, Harvard Medical School, Boston, MA

¹¹Department of Pediatrics, Harvard Medical School, Boston, MA

Abstract

Objectives—To compare registry and EHR data mining approaches for cohort ascertainment in patients with pediatric pulmonary hypertension (PH) in an effort to overcome some of the limitations of registry enrollment alone in identifying patients with disease phenotypes.

Study design—This study was a single-center retrospective analysis of EHR and registry data at Boston Children's Hospital. The local Informatics for Integrating Biology and the Bedside (i2b2)

Address correspondence to: Kenneth D. Mandl, MD, MPH, Computational Health Informatics Program, Boston Children's Hospital, 300 Longwood Avenue, 1 Autumn 535, Mail Stop BCH3187, Boston, MA 02115, Tel. 617-355-4145, kenneth_mandl@harvard.edu.

*List of additional member of the PPHNet and NHLBI Pediatric Pulmonary Vascular Disease Outcomes Bioinformatics Clinical Coordinating Center is available at www.jpeds.com (Appendix).

Publisher's Disclaimer: This is a PDF file of an unedited manuscript that has been accepted for publication. As a service to our customers we are providing this early version of the manuscript. The manuscript will undergo copyediting, typesetting, and review of the resulting proof before it is published in its final citable form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

The authors declare no conflicts of interest.

data warehouse was queried for billing codes, prescriptions, and narrative data related to pediatric PH. Computable phenotype algorithms were developed by fitting penalized logistic regression models to a physician-annotated training set. Algorithms were applied to a candidate patient cohort and performance was evaluated using a separate set of 136 records and 179 registry patients. We compared clinical and demographic characteristics of patients identified by computable phenotype and the registry.

Results—The computable phenotype had an area under the ROC curve of 90% (95% CI 85% – 95%), positive predictive value of 85% (95% CI 77% – 93%), and identified 413 patients (an additional 231%) with pediatric PH not enrolled in the registry. Patients identified by the computable phenotype were clinically distinct from registry patients, with greater prevalence of diagnoses related to perinatal distress and left heart disease.

Conclusions—Mining of EHRs using computable phenotypes identified a large cohort of patients not recruited using a classic registry. Fusion of EHR and registry data can improve cohort ascertainment for the study of rare diseases.

Trial Registration—[ClinicalTrials.gov](https://clinicaltrials.gov/ct2/show/study/NCT02249923): NCT02249923

Keywords

pulmonary hypertension; pediatrics; registry; computer-based model; bioinformatics

As with many rare diseases, the challenge of obtaining adequate sample sizes to study pediatric pulmonary hypertension (PH) has in the past been addressed by enrolling patients into registries.^{1, 2} However, registries are resource-intensive and can only capture patients who come to recruiters' attention. More recently, widespread adoption of electronic health records (EHRs) has allowed clinical data collected at the point of care to be mined in order to identify patients with particular disease phenotypes.^{3, 4} Statistically derived “computable phenotypes,” comprised of a composite of varying EHR data elements, accurately identify patients of interest in clinical data repositories.^{5–12}

Our study had two primary objectives. First, we sought to compare two rare disease cohort ascertainment methods, a classic registry approach and mining of EHR data using a computable phenotype. Second, we examined the relative contribution of different EHR data types—standard codified (COD) data, domain-specific coding, and narrative features extracted by natural language processing (NLP)—to the optimal computable phenotype algorithm.

METHODS

The goal of EHR-based cohort ascertainment is to identify patients with a high probability of having the phenotype of interest—in this case, PH—and thus who can be used as subjects for future studies without further clinician or researcher review. An overview of our EHR-based approach for cohort ascertainment is outlined in Figure 1 (available at www.jpeds.com). Beginning with the entire EHR database, we: (1) created a candidate cohort of all possible patients with PH using broad criteria and extracted relevant COD and NLP features; (2) developed “computable phenotype” classification algorithms using a

training set annotated with gold standard PH status; (3) applied the algorithms to all patients in the candidate cohort and validated the algorithm's predicted PH status using an independent test set of manually reviewed records; and (4) compared the cohort defined by the best performing computable phenotype algorithm with a cohort enrolled in the traditional registry. The institutional review board at Boston Children's Hospital approved the study protocol, which included informed consent for participation in the registry and waiver of informed consent for review and analysis of EHR data.

There were two main data sources for this study. The first was a traditional registry started by the Pediatric Pulmonary Hypertension Network (PPHNet) in 2014 ([ClinicalTrials.gov: NCT02249923](https://clinicaltrials.gov/ct2/show/study/NCT02249923)).¹³ This is a multi-center pediatric registry enrolling new and established patients diagnosed with PH between birth and 18 years of age. Eligible patients were identified at Boston Children's Hospital beginning in October 2014 based on referral to pulmonary hypertension clinic or when the pulmonary hypertension service was consulted during hospital admission. For the purposes of this study, only patients recruited to the PPHNet registry at Boston Children's Hospital were included. The second data source was the local Informatics for Integrating Biology and the Bedside (i2b2)¹⁴ data warehouse at Boston Children's Hospital, which is part of the Scalable Collaborative for a Learning Health System (SCILHS)¹⁵ network. i2b2 contains clinical data such as medication prescriptions, billing data including diagnostic codes, demographic data, and clinical notes.

A candidate cohort enriched in the phenotype of interest—that is, PH—was created in order to increase the positive predictive value for the subsequent machine learning algorithm given the rarity of PH in the general EHR population. Identification of the candidate cohort was designed to maximize sensitivity for capturing appropriate patients. We included patients with potential pediatric PH based on diagnostic codes and medications in i2b2. Patients with at least one International Classification of Diseases, Ninth Revision (ICD-9) code related to pulmonary vascular disease or persistent fetal circulation (416.0, 416.8, 416.9, 747.83) were added. Also included were patients who were prescribed or had recorded in their medication history at least one of the following medications used to treat pediatric PH: ambrisentan, bosentan, epoprostenol, iloprost, sildenafil, tadalafil, and treprostinil. Analysis was limited to patients with relevant ICD-9 codes or medications documented after January 1, 2013 in order to minimize capture of patients no longer followed at our institution for PH. The EHR candidate cohort consisted of 3,494 patients, of whom 3,205 had notes. One patient's PH status could not be determined on chart review and was removed from further analysis.

For each patient in the candidate cohort, we queried i2b2 for diagnoses and procedures known to be associated with pediatric PH. We also included diagnoses in which right ventricular (RV) hypertension led to inaccurate coding of PH. Fyler codes, which define a taxonomy of congenital heart disease and related concepts,¹⁶ were available from the Department of Cardiology database at the study institution. Fyler codes for PH and RV hypertension were incorporated into the dataset. We refer to COD data, excluding Fyler codes, as “standard codified data,” to distinguish it from domain-specific coding. Indicators of healthcare utilization were created by calculating the total number of ICD-9 codes for each patient. Table I (available at www.jpeds.com) shows features extracted. Features present in fewer than 5% of patients were dropped from further analysis.

In order to use narrative data in the electronic health record as features for development of a computable phenotype of pulmonary hypertension (PH), we used natural language processing (NLP) to identify concepts relevant to PH in patients' clinical notes. Review articles of PH^{17, 18} and notes of sample patients with and without PH were used to create a dictionary of clinical concepts relevant to PH. We processed these sources using the Narrative Information Linear Extraction (NILE) package,^{19, 20} which includes a named entity recognition package that extracts terms and maps them to clinical concepts using the Unified Medical Language System (UMLS).²¹ The final dictionary consisted of 137 concepts, including all concepts appearing in both review articles and notes and additional concepts selected by a clinician (AG) that appeared in either the articles or notes, but not both. Clinical notes, including inpatient progress notes, discharge summaries, and outpatient clinic notes, for the patients identified in the candidate cohort were then processed using NILE to search for the 137 PH dictionary concepts. We dropped 79 concepts appearing in fewer than 5% of notes that mentioned PH. Of 211,806 notes for patients in the candidate cohort, 181,524 eligible notes with length greater than 500 characters were used for analysis. Each note was annotated with the number of times each dictionary concept appeared in that note, and the total number of mentions for each concept in all clinical notes of each patient were used for candidate NLP features. The list of concepts used is shown in Table I.

We randomly selected 242 patients from the candidate cohort to create a training set. 250 patients were initially reviewed; 7 were subsequently found not to have had a PH-related diagnosis or medication recorded after 1/1/2013, as specified in our inclusion criteria, and were excluded from further analysis. An additional patient's PH status could not be determined from chart review, as noted above, and was also removed from further analyses. A single physician (AG) manually reviewed the medical records for the remaining 242 patients using a structured case report form to determine the gold standard PH status for these patients. A second physician (MPM) reviewed 50 of these charts to establish inter-rater reliability, which was 92%. Discordant classification was resolved by consensus.

Clinicians used the following criteria to determine PH status during manual review of the medical record. Cardiac catheterization, when available, was considered the gold standard for a diagnosis of PH when thresholds of mean distal pulmonary arterial pressure ≥ 25 mm Hg and indexed pulmonary vascular resistance ≥ 3 Wood units/m² were met. When catheterization data were not available, we used echocardiographic evidence of elevated RV pressure (estimates of RV pressure based on tricuspid regurgitation jet, gradient across a ventricular septal defect or patent ductus arteriosus, or ventricular septal configuration) without evidence of RV outflow tract obstruction (RVOTO) to classify a patient as having PH. When RVOTO was present, patients without catheterization evidence of PH were presumed not to have PH. In cases of unrestrictive left-to-right shunts, if evidence of elevated RV pressure did not persist at least 6 months after repair of the shunt, patients were presumed to have elevated RV pressure secondary to the unrestrictive lesion and not PH. Patients with only single echocardiograms showing elevated RV pressure within the first 3 days of life, without other documentation of follow-up for PH or with documented subsequent resolution of PH, were considered not to have PH in order to exclude patients with transitional circulation but include patients with persistent pulmonary hypertension of the newborn.

The total number of COD and NLP candidate features was large relative to the size of the training set. Therefore, to reduce overfitting, we first used the Automated Feature Extraction for Phenotyping (AFEP) procedure to reduce the number of candidate features.²⁰ With AFEP, features were included in model estimation if their absolute correlation with the NLP variable for mention of “primary pulmonary hypertension” and with ICD-9 codes for other chronic pulmonary heart diseases (416.8) or persistent fetal circulation (747.83) exceeded 0.2 (Table II; available at www.jpeds.com). All models also included the ICD-9 codes used to obtain the candidate cohort and the total number of ICD-9 codes for the patient to adjust for healthcare utilization.

To predict the probability of having PH, we used the *glm*path package²² to fit adaptive least absolute shrinkage and selection operator (LASSO) penalized logistic regression models²³ to the training data with the response being the gold standard PH status and predictors being (i) COD features only or (ii) COD and NLP features. The penalized regression reduces overfitting and simultaneously performs variable selection. The tuning variable was selected based on a modified Bayes’ Information Criterion.^{24, 25} We also estimated models (i) and (ii) excluding Fyler codes, as they are not available at all institutions, for a total of four algorithms. Count data, which tended to be highly skewed, were log-transformed and all features were standardized to have unit variance prior to model fitting. Because NLP data were not available for all patients, the models including NLP features were trained with the 224 patients in the training set with narrative information. The predicted probabilities of PH for patients who did not have narrative information were obtained with the corresponding COD models.

Receiver operating characteristic (ROC) analysis was conducted with the training set to determine an appropriate threshold for classifying PH status. We corrected for overfitting bias with the 0.632 bootstrap procedure.^{26, 27} To tune the algorithm for high positive predictive value, the threshold probability was selected to achieve 95% specificity. Each algorithm was applied to the patients in the candidate cohort (including to the 242 patients used for model training) and subjects with predicted probabilities exceeding the estimated threshold were classified as having PH.

The performance of the four computable phenotype algorithms was further validated using an additional 136 randomly selected patients (150 patients were initially sampled, of whom 12 were registry patients and 2 were sampled twice) from the candidate cohort whose gold standard PH status was determined by a single physician (AG). Included also for validation were the 179 patients with PH who had been enrolled in the traditional registry at the study institution at the time of data extraction. Inverse probability weighting was used to correct for the non-random inclusion of these validation samples. Standard error estimates were obtained with the standard bootstrap with 500 replicates.

To determine the number of eligible patients identified by the computable phenotype algorithm who were not identified in the registry, we selected patients from those predicted by the bestperforming computable phenotype algorithm to have PH. In order to maximize validity of the comparison for prospective patient enrollment, selected EHR patients were matched to registry eligibility criteria of PH diagnosis prior to age 19 years and being seen

at the study institution after October 28, 2014, the initial date of recruitment for the PPHNet registry. These criteria thus avoided an invalid comparison between the PPHNet registry patients and EHR patients with adult-onset PH seen at our institution (typically adults with congenital heart disease) as well as with counting historical PH patients who were no longer followed at our hospital and thus could not have been recruited to the PPHNet registry.

We compared the patients identified by computable phenotype and by the registry across demographic and clinical variables. Diagnoses were determined based on ICD-9 billing codes for patients, and only diagnostic codes present in at least 5% of the cohort were included. The computable phenotype and registry cohorts were compared using Student *t* test for continuous variables and Fisher exact test for categorical variables. A *P* value < 0.05 was considered statistically significant. Analyses were performed using R version 3.3.1.²⁸

RESULTS

As of October 9, 2015, there were 179 patients with PH enrolled in the PPHNet registry at our institution. An additional 55 patients had been screened for participation in the registry but were not enrolled, due to either not meeting inclusion criteria or informed consent not being given. Prevalence of PH among the 242 patients reviewed to create the gold standard training dataset was 38%. After frequency control, the initial set of candidate features included a total of 24 COD and 57 NLP features. After automated feature selection, a total of 23 features (9 COD, one of which was a domain-specific Fyler code, and 14 NLP) were included in the candidate set for model estimation (Table II). The most influential features—that is, those with the most positive or negative coefficient values—in the computable phenotype algorithms are shown in Table III (available at www.jpeds.com). COD variables that were inversely associated with PH (tetralogy of Fallot, pulmonary valve or subpulmonic stenosis, pulmonary atresia, or transposition of the great arteries) were not influential if NLP variables were included in the computable phenotype algorithm.

As compared with 179 patients enrolled in the registry at our institution at the time of data analysis, the best computable phenotype model derived using EHR data identified 413 patients who may have been eligible for the registry but not enrolled (Figure 2 and Table IV; Figure 2 available at www.jpeds.com). An additional 151 patients with pediatric PH were identified by the algorithm but were not seen at the study institution after the start of recruitment for the PPHNet registry. The computable phenotype algorithm also identified 300 patients whose first PH-related diagnosis or medication was documented after age 19 years. 162 (91%) of 179 registry patients were correctly classified as having PH by the computable phenotype algorithm. Addition of Fyler codes and NLP features to the models increased sensitivity from 48% for COD data without Fyler codes to 66%. Models including either Fyler codes or NLP had a similar area under the ROC curve (AUC) of 89%, and AUC with all data types included was 90%.

Patients identified by the computable phenotype algorithm but not captured in the registry were younger than registry patients (7.8 ± 10.9 y v. 9.2 ± 8.1 y, $P < .001$) and included a higher proportion of deceased patients (29 (7%) v. 5 (3%), $P = 0.05$) (Table V). A higher proportion of registry patients had congenital diaphragmatic hernia (CDH), reflecting

institutional practice of referring all CDH patients to PH clinic. Patients in the registry appeared to be more medically complex, with a higher number of hospital visits and number of diagnoses per patient. Finally, we noted differences in PH coding practices: a higher proportion of computable phenotype patients not in the registry had “persistent fetal circulation” billed as compared with registry patients (122 (30%) v. 27 (15%), $P < .001$), whereas “primary pulmonary hypertension,” “other chronic pulmonary heart disease,” and “chronic pulmonary heart disease, unspecified,” were more commonly billed for patients in the registry ($P < .01$ for each; Figure 3, available at www.jpeds.com). Inclusion of the 151 patients identified by the computable phenotype but not seen at the study institution after commencement of registry recruitment did not significantly alter these results, with the exception of an even higher proportion of deceased patients (65 of 564 (12%), $P < .001$).

We compared high frequency diagnoses between the cohorts identified by the computable phenotype and by the registry. Among computable phenotype patients as compared with registry patients, two patterns of more frequent diseases emerged: congenital and acquired heart disease, including mitral valve disease and heart failure; and neonatal distress, including meconium aspiration and respiratory distress syndrome (Figure 4). Among patients enrolled in the registry, gastrointestinal disease and malformations, chronic respiratory disease—including CDH, obstructive sleep apnea, asthma, and machine dependence—and prematurity, developmental disorders, and seizures were 1.7 – 2.4 times more frequent than among patients identified by computable phenotype (Figure 5).

DISCUSSION

An EHR-based computable phenotype identified a pediatric PH cohort more than three times as large as a traditional disease registry. Notably, the two populations are phenotypically different, demonstrating that certain patient populations are not captured effectively by the registry. Thus, we demonstrate the utility of cohort ascertainment through EHR mining for powering future studies of pediatric PH, and directly compare the cohorts using computable phenotypes versus a traditional registry.

Traditional patient registries have been useful for the study of rare diseases,²⁹ but PH registries have had limited success enrolling children. For example, the Registry to Evaluate Early and Long-term PH Disease Management (REVEAL) enrolled approximately 3,000 patients but only included about 200 children.¹ The Tracking Outcomes of Practice in Pediatric Pulmonary Hypertension (TOPP) registry enrolled 456 pediatric patients, of whom 362 had confirmed pulmonary hypertension.² In contrast, at a single institution, application of a computable phenotype to EHR data identified 575 patients with PH.

Registry-based cohorts may be limited if not all patients of interest are seen in the enrolling clinics or by the enrolling services, as occurred in our patient population. The PPHNet registry is recruiting patients seen by PH specialists, but such a referral was not made for many of the patients identified only by the computable phenotype algorithm. Many of these patients may not come to the attention of PH specialists because their PH resolves with age, as evidenced by the increased relative frequency of neonatal diagnoses associated with PH in EHR versus registry patients, and by “persistent fetal circulation” being more likely to be

used as the diagnostic code for patients in the EHR cohort. These patients are expected to be important for understanding the natural history and pathophysiology of persistent pulmonary hypertension of the newborn. Another group of patients that was relatively over-represented in the computable phenotype cohort were patients with diagnoses of left heart disease. We suspect that many of these patients, in whom PH was not the primary but rather secondary pathophysiology, were being managed by their primary cardiologists and thus did not come to PH specialist attention. This has the potential to lead to ascertainment bias in registry-based studies when recruitment is dependent on patients being referred to disease specialists. Lastly, some patients who were identified through screening to be candidates for inclusion in the PPHNet registry were unable to be approached by recruiters, or their families declined to give consent. These patients can nonetheless be identified by a computable phenotype and followed in an EHR-based cohort based on clinical data routinely collected at the point of care. These complementary benefits of an EHR-based approach to cohort ascertainment allow more complete patient capture, particularly important for rare diseases and for particular patient sub-populations that may be more difficult to recruit to traditional registries.

Similar to previous efforts in this domain,³ developing a computable phenotype for PH poses the challenge of a complex disorder with unique determinants of disease state that are often confounded by similar traits not indicative of the disease of interest. In particular, we found that RV hypertension often, but not always, led to inaccurate coding of PH in billing codes. Most, but not all, patients with RVOTO who had an ICD-9 code related to PH had RV hypertension but did not actually have PH. Thus, rules-based approaches were either overly broad or narrow. More sophisticated, statistically derived computable phenotypes allowed us to obtain algorithms with adequate sensitivity and specificity. In such algorithms, diagnoses such as those associated with RVOTO were inversely associated with PH in the computable phenotype models that did not include NLP variables. Models including NLP-derived variables, by incorporating data contained in expert clinicians' notes, did not rely on these billing codes to achieve adequate performance, demonstrating the utility of NLP in overcoming some of the limitations of claims data. Domain-specific coding systems such as Fyler codes also helped overcome such limitations. Similar performance might be achieved by carefully curated problem lists or other specific annotations by domain experts.

Like other studies of EHR data, this study has several inherent limitations. Although the best performing algorithm achieved a PPV of 85%, misclassification of PH status may lead to slight bias in results of analyses of this cohort. This work was performed at a single institution, and institution-specific coding and documentation practices may affect generalizability of this computable phenotype to other institutions. Lastly, our current computable phenotype algorithm relies on ICD-9 codes, which have recently been replaced with ICD-10 in the United States. Given that the four ICD-9 codes used for cohort definition in the present study can be mapped to six ICD-10 codes and that coding practices will likely differ across institutions, further work will be necessary to adapt the current algorithm in order to allow ongoing patient recruitment based on EHR data across sites.

In this study, we identified patients based on computable phenotype retrospectively, but the EHR-based approach used here provides the backbone for a prospective approach to

ongoing patient recruitment. Based on the queried i2b2 database, which is continually updated with new patient data, the computable phenotype can continue to identify new patients for recruitment into future studies. Furthermore, the predicted status of having or not having PH can be incorporated as a “fact” into i2b2, creating a permanent notation for identifying patients for research purposes outside the scope of the current project. These facts can be validated and integrated with other data, including registry data and textual information in notes processed by NLP, using knowledge management pipelines such as tranSMART.³⁰ Thus, by integrating traditional registry information with patient identification and phenotyping based on EHR data, we create a sustainable infrastructure to significantly grow the number of patients and amount of data available for pediatric PH research.

EHR- and registry-based methods of patient recruitment are complementary, with a registry infrastructure allowing collection of detailed phenotypic information while an EHR-based computable phenotype allows breadth of patient recruitment, including patient subtypes that are not included in the registry. The computable phenotype infrastructure allows for scalable, ongoing patient identification not only at the study institution but also, if the algorithm is validated, at other centers caring for pediatric PH. We anticipate with these methods recruiting the largest cohort of pediatric PH patients to date for research on this rare, highly morbid disorder.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

Supported by the National Institutes of Health (NHLBI U01HL121518, NHLBI L40HL133929, NIGMS R01GM104303, PCORI CDRN130604608, and NICHD T32HD040128).

Abbreviations

AUC	area under the receiver operating characteristic curve
CDH	congenital diaphragmatic hernia
COD	codified
EHR	electronic health record
ICD-9	International Classification of Diseases, Ninth Revision
NLP	natural language processing
PH	pulmonary hypertension
PPV	positive predictive value
ROC	receiver operating characteristic
RV	right ventricular

RVOTO right ventricular outflow tract obstruction

Appendix

The PPHNet and NHLBI Pediatric Pulmonary Vascular Disease Outcomes Bioinformatics Clinical Coordinating Center Investigators include:

Steven Abman, MD (University of Colorado, Denver, Denver, CO), Ian Adatia, MBChB, FRCP(C) (Stollery Children's Hospital, Mazankowski Alberta Heart Institute, University of Alberta, Edmonton, Alberta, Canada), Eric D. Austin, MD, MSCI (Monroe Carell Jr. Children's Hospital at Vanderbilt, Nashville, TN), Jeffrey Feinstein, MD (Stanford University, Stanford, CA), Jeff Fineman, MD (University of California, San Francisco, San Francisco, CA), Brian Hanna, MD (Children's Hospital of Philadelphia, Philadelphia, PA), Rachel Hopper, MD (Children's Hospital of Philadelphia, Philadelphia, PA), Dunbar Ivy, MD (University of Colorado, Denver, Denver, CO), Roberta Keller, MD (University of California, San Francisco, San Francisco, CA), Usha Krishnan, MD (Columbia University, New York-Presbyterian, Morgan Stanley Children's Hospital, New York, New York), Thomas Kulik, MD (Boston Children's Hospital, Boston, MA), Mary Mullen, MD, PhD (Boston Children's Hospital, Boston, MA), Usha Raj, MD (University of Illinois, Chicago, Chicago, IL), Erika Berman Rosenzweig, MD (Columbia University, New York-Presbyterian, Morgan Stanley Children's Hospital, New York, NY)

References

1. McGoon MD, Miller DP. Reveal: A contemporary US pulmonary arterial hypertension registry. *Eur Respir Rev.* 2012; 21:8–18. [PubMed: 22379169]
2. Berger RM, Beghetti M, Humpl T, Raskob GE, Ivy DD, Jing ZC, et al. Clinical features of paediatric pulmonary hypertension: A registry study. *Lancet.* 2012; 379:537–46. [PubMed: 22240409]
3. Murphy S, Churchill S, Bry L, Chueh H, Weiss S, Lazarus R, et al. Instrumenting the health care enterprise for discovery research in the genomic era. *Genome Res.* 2009; 19:1675–81. [PubMed: 19602638]
4. Kohane IS. Using electronic health records to drive discovery in disease genomics. *Nat Rev Genet.* 2011; 12:417–28. [PubMed: 21587298]
5. Wei WQ, Denny JC. Extracting research-quality phenotypes from electronic health records to support precision medicine. *Genome Med.* 2015; 7:41. [PubMed: 25937834]
6. Wei WQ, Teixeira PL, Mo H, Cronin RM, Warner JL, Denny JC. Combining billing codes, clinical notes, and medications from electronic health records provides superior phenotyping performance. *J Am Med Inform Assoc.* 2016; 23:e20–7. [PubMed: 26338219]
7. Mo H, Thompson WK, Rasmussen LV, Pacheco JA, Jiang G, Kiefer R, et al. Desiderata for computable representations of electronic health records-driven phenotype algorithms. *J Am Med Inform Assoc.* 2015; 22:1220–30. [PubMed: 26342218]
8. Liao KP, Ananthakrishnan AN, Kumar V, Xia Z, Cagan A, Gainer VS, et al. Methods to develop an electronic medical record phenotype algorithm to compare the risk of coronary artery disease across 3 chronic disease cohorts. *PLoS One.* 2015; 10:e0136651. [PubMed: 26301417]
9. Gainer VS, Cagan A, Castro VM, Duey S, Ghosh B, Goodson AP, et al. The biobank portal for Partners personalized medicine: A query tool for working with consented biobank samples, genotypes, and phenotypes using i2b2. *J Pers Med.* 2016; 6:e11. [PubMed: 26927184]
10. Kirby JC, Speltz P, Rasmussen LV, Basford M, Gottesman O, Peissig PL, et al. PheKB: A catalog and workflow for creating electronic phenotype algorithms for transportability. *J Am Med Inform Assoc.* 2016; 23:1046–52. [PubMed: 27026615]

11. 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. [PubMed: 26464024]
12. Lingren T, Chen P, Bochenek J, Doshi-Velez F, Manning-Courtney P, Bickel J, et al. Electronic health record based algorithm to identify patients with autism spectrum disorder. *PLoS One.* 2016; 11:e0159621. [PubMed: 27472449]
13. The Pediatric Pulmonary Hypertension Network. [Accessed June 15, 2016] The pediatric pulmonary hypertension network. <http://www.pphnet.org/>. Published 2016
14. Klann JG, Abend A, Raghavan VA, Mandl KD, Murphy SN. Data interchange using i2b2. *J Am Med Inform Assoc.* 2016; 23:909–15. [PubMed: 26911824]
15. Mandl KD, Kohane IS, McFadden D, Weber GM, Natter M, Mandel J, et al. Scalable collaborative infrastructure for a learning healthcare system (SCILHS): Architecture. *J Am Med Inform Assoc.* 2014; 21:615–20. [PubMed: 24821734]
16. Colan, SD. Early database initiatives: The Fyler codes. In: Barach, PR, Jacobs, J, Lipshultz, SE, Laussen, PC., editors. *Pediatric and congenital cardiac care.* London: Springer-Verlag; 2015. p. 163–9.
17. Nicolarsen J, Ivy D. Progress in the diagnosis and management of pulmonary hypertension in children. *Curr Opin Pediatr.* 2014; 26:527–35. [PubMed: 25136947]
18. Pandya KA, Puligandla PS. Pulmonary hypertension management in neonates. *Semin Pediatr Surg.* 2015; 24:12–6. [PubMed: 25639804]
19. Yu S, Cai T. A short introduction to NILE. *CoRR.* 2013 abs/1311.6063.
20. 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. *Journal of the American Medical Informatics Association.* 2015; 22:993–1000. [PubMed: 25929596]
21. Bodenreider O. The Unified Medical Language System (UMLS): Integrating biomedical terminology. *Nucleic Acids Res.* 2004; 32:D267–70. [PubMed: 14681409]
22. Park MY, Hastie T. Glmpath: L1 regularization path for generalized linear models and Cox proportional hazards model [computer program]. 2013
23. Zou H. The adaptive lasso and its oracle properties. *Journal of the American Statistical Association.* 2006; 101:1418–29.
24. Minnier J, Tian L, Cai T. A perturbation method for inference on regularized regression estimates. *Journal of the American Statistical Association.* 2011; 106:1371–82. [PubMed: 22844171]
25. Agniel D, Liao KP, Cai T. Estimation and testing for multiple regulation of multivariate mixed outcomes. *Biometrics.* 2016; 72:1194–205. [PubMed: 26910481]
26. Efron B. Estimating the error rate of a prediction rule: Improvement on cross-validation. *Journal of the American Statistical Association.* 1983; 78:316–31.
27. Efron B, Tibshirani R. Improvements on cross-validation: The .632+ bootstrap method. *Journal of the American Statistical Association.* 1997; 92:548–60.
28. R Core Team. R: A language and environment for statistical computing [computer program]. Vienna, Austria: R Foundation for Statistical Computing; 2016.
29. Gliklich, R., Dreyer, N., editors. *Registries for evaluating patient outcomes: A user's guide.* 2. Rockville, MD: Agency for Healthcare Research and Quality; 2010.
30. Scheufele E, Aronzon D, Coopersmith R, McDuffie MT, Kapoor M, Uhrich CA, et al. tranSMART: An open source knowledge management and high content data analytics platform. *AMIA Jt Summits Transl Sci Proc.* 2014; 2014:96–101. [PubMed: 25717408]

Figure 1. Overview of our EHR-based approach for cohort ascertainment

Figure 2. Venn diagram showing computable phenotype- and registry-based cohorts
The cohorts used for subsequent comparisons are labeled in small caps.

Figure 3. Comparison of ICD-9 coding practices between the computable phenotype and registry cohorts

Bars represent percent of patients who had each ICD-9 code recorded at least once. $P < 0.0001$ (****), $P < 0.001$ (***), $P < 0.01$ (**) for pairwise comparisons between cohorts.

Figure 4. ICD-9 codes with greater relative frequency among patients identified only in the EHR as compared with patients enrolled in the registry

Magenta lines indicate ICD-9 codes related to heart disease and green lines indicate ICD-9 codes related to neonatal distress. Trivial diagnoses are shown in light grey to emphasize diagnoses of interest. Only conditions with $P < 1.5$ are shown.

Figure 5. ICD-9 codes with greater relative frequency among registry patients as compared with patients identified only in the EHR

Magenta lines indicate ICD-9 codes related to gastrointestinal disease and malformations, blue lines indicate ICD-9 codes related to chronic respiratory disease, and green lines indicate ICD-9 codes related to prematurity, developmental disorders, and seizures. Trivial diagnoses are shown in light grey to emphasize diagnoses of interest. Only conditions with $P < 0.05$ are shown. Some ICD-9 code descriptions appear twice due to unique ICD-9 codes mapping to similar descriptions.

Table 4**Model performance**

Eligible patients captured represents the number of patients identified by the EHR computable phenotype algorithm that could have qualified for the registry. Registry patients missed shows the number of patients enrolled in the registry (N = 179) that the computable phenotype did not identify.

	SC only	SC/Fyler	SC/NLP	SC/NLP/Fyler
PPV, % (95% CI)	82 (72 – 91)	83 (75 – 92)	84 (75 – 93)	85 (77 – 93)
Sensitivity, % (95% CI)	48 (42 – 54)	59 (53 – 65)	64 (57 – 71)	66 (60 – 73)
AUC, % (95% CI)	85 (78 – 92)	89 (83 – 95)	89 (83 – 95)	90 (85 – 95)
Total eligible patients captured, No.	470	518	575	575
New eligible patients captured, No.	323	364	414	413
Registry patients missed, No.	32	25	18	17

AUC = area under the receiver operating characteristic curve; NLP = natural language processing; PPV = positive predictive value; SC = standard codified

Table 5
Comparison between computable phenotype and registry patients

New patients refers to patients identified by the computable phenotype algorithm but not enrolled in the registry.

	New EHR patients	Registry patients	<i>P</i> value
	(N = 413)	(N = 179)	
Age, mean (sd), y	7.8 (10.9)	9.2 (8.1)	< 0.001
Deceased, No. (%)	29 (7)	5 (3)	0.05
CDH, No. (%)	31 (8)	59 (33)	< 0.001
Prematurity, No. (%)	66 (16)	35 (20)	0.29
Trisomy 21, No. (%)	43 (10)	12 (7)	0.17
PH medication, No. (%)	117 (28)	84 (47)	< 0.001
Right heart catheterization, No. (%)	212 (51)	102 (57)	0.21
No. of unique encounters, mean (sd)	132 (161)	191 (164)	< 0.001
No. of unique ICD-9 codes, mean (sd)	66 (56)	80 (51)	< 0.001

CDH = congenital diaphragmatic hernia; EHR = electronic health record; PH = pulmonary hypertension