Developing and evaluating pediatric phecodes for high-throughput phenotyping using electronic

health records

By

Monika Grabowska

Thesis

Submitted to the Faculty of the

Graduate School of Vanderbilt University

in partial fulfillment of the requirements

for the degree of

MASTER OF SCIENCE

in

Biomedical Informatics

December 16, 2023

Nashville, Tennessee

Approved:

Wei-Qi Wei, M.D., Ph.D.

Prince J. Kannankeril, M.D., M.S.C.I.

Robert Carroll, Ph.D.

To my Dad, thank you for always being my greatest supporter.

ACKNOWLEDGEMENTS

I would first like to thank my research mentor, Wei-Qi Wei, for his invaluable guidance, encouragement, and wisdom. I also wish to thank my committee members, Prince Kannankeril and Robert Carroll, for their mentorship throughout this project. I thank Sara Van Driest, who helped develop this work from the start, QiPing Feng for generously sharing her time and expertise, and Anna Patrick and Jamie Robinson for their clinical insights. Thank you all for continuously supporting and inspiring me. I am incredibly grateful for everything you have taught me.

Finally, I thank my parents Marek Grabowski and Renata Polanowska-Grabowska, my sister Anita, my grandparents Jacek and Aleksandra Grabowski, and my cats Ellen and Missy for their endless love and support. Dad, although you could not experience this adventure with me, I know you are always watching over me.

TABLE OF CONTENTS

LIST OF FIGURES

CHAPTER 1

INTRODUCTION AND BACKGROUND

In 2009, the Health Information Technology for Economic and Clinical Health (HITECH) Act was signed into law in the United States.[1] Intended to incentivize the use of electronic health records (EHRs), this legislation played a significant role in driving widespread EHR adoption across the nation,[2,3] with profound effects for both healthcare and biomedical research.[4] The transition from paper-based charts to EHRs increased the availability of patient health data for research, creating unprecedented opportunities for scientific discovery. EHRs are a rich source of longitudinal clinical information, including patient diagnoses, laboratory values, and medications, which can be harnessed to examine larger patterns of human health and disease in real-time. However, extracting clinically meaningful information from EHRs has proven to be a challenging task for researchers given the high volume and velocity of data accumulation, as well as the complex and often fragmented nature of the data itself, prompting the development of novel informatics tools and approaches to optimize EHR data for research use.

Electronic health record-based phenotyping

Phenotyping, the process of identifying patients with common traits, symptoms, conditions, or other characteristics of interest, plays a fundamental role in the research conducted using EHRs by defining the patient cohorts that are analyzed.[5] Accurate phenotyping is critical; however,

aggregating and translating EHR data into precise and clinically meaningful phenotypes is often difficult because of the highly heterogeneous nature of EHR data.

EHRs contain multimodal data from a variety of sources recorded at different points in time, resulting in considerable variability in data representation and quality. Data in the EHR can be broadly divided into two main formats: structured data organized into discrete fields (e.g., demographic information, billing codes, vital signs, laboratory measurements, and prescriptions) and unstructured data with no pre-defined data model (e.g., free-text clinical notes, discharge summaries, and radiology and pathology reports, as well as medical images). Different healthcare providers may record the same clinical concept differently within a patient's EHR based on personal preference or institutional policies and practices, enabling a single disease phenotype to be represented in multiple ways. Furthermore, data in the EHR may be missing, incomplete, outdated, or even conflicting, making it difficult to reliably identify patients with a specific clinical phenotype.

To address these issues, multiple phenotyping strategies have been proposed leveraging a variety of data types from the EHR, including rule-based algorithms that apply clearly defined inclusion and exclusion criteria, typically in the form of structured data,[6–9] natural language processing (NLP) pipelines that extract phenotypic concepts from unstructured data,[10,11] machine learning approaches that identify phenotypes from patterns detected in EHR data,[12,13] and hybrid approaches.[14–16] Traditional methods for EHR-based phenotyping methods have primarily relied on the characterization of a single phenotype (for example, type 2 diabetes), with extensive manual chart review required to evaluate phenotype performance.[17] More recently, EHR phenotyping efforts have also involved the development of scalable, high-throughput

2

phenotyping approaches capable of capturing thousands of phenotypes simultaneously without the need for manual review.[5]

Beyond reducing the time and effort for phenotype generation and review, high-throughput phenotyping presents unique opportunities for new scientific discovery by allowing researchers to quickly obtain a comprehensive clinical overview of many patients at once, facilitating large-scale analyses of EHR data capable of identifying phenotypic associations that may be missed in studies with smaller sample sizes. In particular, the development of high-throughput phenotyping tools paved the way for phenome-wide association studies (PheWAS), which investigate the association between a variable of interest and a wide variety of clinical phenotypes.[18–20] PheWAS approaches have most commonly been used to identify phenotypes associated with a single genetic variant in the form of a small nucleotide polymorphism (SNP), but combinations of SNPs and even continuous measures (e.g., laboratory values such as white blood cell count [21]) have also been used as the target variables in PheWAS.

Phecodes for high-throughput electronic health records phenotyping

Phecodes (phenotype codes) were originally developed to demonstrate the feasibility of PheWAS for identification of SNP-disease associations.[18,22] Phecodes leverage International Classification of Diseases (ICD) codes, a common form of structured diagnosis data present in virtually all EHR systems in the U.S. due to their use in medical billing and reimbursement. Before 2015, the U.S. used the ICD-9-CM coding system to report patient diagnoses in the EHR, which included >13,000 numeric diagnosis codes.[23] Given the size of the ICD-9-CM code set, the initial phecodes created custom groupings of ICD-9-CM codes in order to reduce the phenotype space, allowing for clinically meaningful PheWAS outcomes.[18] The phecode groupings were

3

found to align better with diseases mentioned in clinical practice and produced superior PheWAS results compared with ICD-9-CM codes and the Clinical Classifications Software (CCS) for ICD-9-CM.[24] In October 2015, ICD-9-CM was replaced by ICD-10-CM in the U.S. (>68,000 alphanumeric codes), and the phecode system was extended to include a mapping for ICD-10-CM codes and internationally used ICD-10 codes to phecodes, providing comprehensive coverage of existing ICD codes.[25]

Phecodes offer several unique advantages for PheWAS. Importantly, the phecodes system defines exclusion phecodes for each phecode, which are used to generate appropriate case and control groups for all phecodes. The exclusion phecodes restrict the ICD codes that patients in the control group can have (i.e., controls for a phecode cannot exhibit any of the phenotypes represented by that phecode's exclusion criteria), effectively ensuring that controls do not have diseases related to the phecode of interest, which could reduce the strength of the signals detected in PheWAS. Similarly, the phecodes system applies sex restrictions for sex-specific phecodes, which pertain only to males (e.g., prostate cancer) or females (e.g., endometriosis). For these sex-specific phecodes, only patients with the relevant sex are counted as cases and controls, again ensuring that the case and control groups are appropriately formed for PheWAS analysis. Finally, phecodes provide a means of standardizing phenotype definitions across different studies and datasets, allowing for reliable and reproducible phenotyping.

Phecodes applied to pediatric cohorts

Phecodes have been used widely to perform PheWAS in a wide range of patient populations;[26–29] however, they are best suited for use in adult populations. The phenotypes represented by the phecodes include a variety of chronic conditions that predominantly affect older patients (e.g.,

dementia, colorectal cancer, and osteoarthritis), and do not capture certain pediatric phenotypes with appropriate granularity (e.g., chromosomal anomalies). However, the existing phecodes could be optimized for use in pediatric cohorts by removing phecodes representing uncommon pediatric phenotypes and creating new, more detailed phecodes for common pediatric conditions. Creating such specialized pediatric phecodes presents a valuable opportunity to conduct meaningful large-scale PheWAS analyses in pediatric populations and develop a better understanding of pediatric disease.

Motivation and research aims

The study presented in this thesis was motivated by a critical need for high-throughput EHR phenotyping tools specialized for use in pediatric populations. Pediatric disease remains understudied in comparison to adult disease, and existing high-throughput EHR phenotyping tools do not account for disease differences between pediatric and adult patients, necessitating the development of new pediatric-specific phenotyping tools. This thesis is presented in three chapters. The first chapter introduces the background and motivation for this research. The second chapter describes the development and validation of specialized pediatric phecodes (Peds-Phecodes) to support large-scale analyses of pediatric EHR data. The final chapter summarizes and discusses the limitations and future directions of this work.

CHAPTER 2

DEVELOPING AND EVALUATING PEDIATRIC PHECODES (PEDS-PHECODES) FOR
HIGH-THROUGHPUT PHENOTYPING USING ELECTRONIC HEALTH RECORDS

Background and significance

The pediatric spectrum of disease is distinct from its adult counterpart.[1] Children experience a variety of illnesses not commonly observed in adults. This includes congenital anomalies, genetic disorders with early mortality, and certain infectious diseases.[2,3] Conversely, adults experience many diseases unlikely to affect pediatric patients, such as Alzheimer's disease, breast and prostate cancer, osteoarthritis, and many other conditions associated with aging.[4] Although pediatric data have rapidly accumulated in electronic health records (EHRs), pediatric patients have not been prioritized in developing high-throughput phenotyping tools such as phecodes, preventing researchers from performing focused large-scale analyses of pediatric data, including pediatric-specific phenome-wide association studies (PheWAS) and genome-wide association studies (GWAS), contributing to missed opportunities for scientific discovery.

The development of phecodes represents a key effort in EHR phenotyping. Phecodes aggregate relevant International Classification of Diseases (ICD-9-CM, ICD-10-CM, and ICD-10) codes into distinct phenotypes to better represent clinically meaningful diseases and traits (e.g., grouping ICD-9-CM codes 162.* representing lung cancer and ICD-9-CM codes V10.1* representing a history of lung cancer).[5,6] Phecodes are represented using numeric codes arranged in a three-level hierarchy, allowing phenotypes to be captured at various levels of granularity. Root phecodes, located at the top of the phecode hierarchy, provide the broadest phenotype definitions

and are represented using whole numbers. These root phecodes can then branch into progressively more detailed sub-phecodes, indicated by decimal digits. In the current phecodes (version 1.2), up to two levels of additional phenotypic granularity (i.e., two decimal places) are available for each root phecode. Phecodes with a single digit following the decimal point are referred to as level 1 sub-phecodes; level 2 sub-phecodes have two digits after the decimal point. For example, root phecode 008 "Intestinal infection" branches into level 1 sub-phecode 008.5 "Bacterial enteritis", which then branches into level 2 sub-phecode 008.51 "Intestinal E. coli".

We previously demonstrated that phecodes produced superior results in PheWAS compared to other coding systems, including ICD and the Clinical Classifications Software.[5] Since their introduction, phecodes have been globally used in multiple studies to both replicate known genotype-phenotype associations and discover new ones.[7–9] Phecodes have also been used beyond genetics to study long-term disease effects.[10] The latest version of phecodes can be found at: https://wei-lab.app.vumc.org/phecode-data/phecodes

While the existing phecodes are valuable in biomedical research, they need to be optimized for use in pediatric phenomic analyses. Phecodes were developed using population-based diagnoses predominantly from adult patients, which do not accurately reflect pediatric conditions. Modifying the existing phecodes to capture diseases primarily affecting pediatric patients with increased granularity and exclude age-related diseases uncommon in the pediatric population presents an opportunity to increase statistical power for identifying signals more efficiently and accurately.

In this study, we use current phecodes as a starting point to create Peds-Phecodes, specialized pediatric phecodes that more appropriately reflect the unique spectrum of pediatric disease, and evaluate their performance in large-scale PheWAS analyses using real-world data.

7

Materials and methods

Data source

We used de-identified EHRs from Vanderbilt University Medical Center's (VUMC) Synthetic Derivative, a repository of rich, longitudinal clinical information encompassing data from >3.5 million patients, including data from >1 million pediatric individuals.[11] For our validation analyses, we also used data from VUMC's DNA biobank, BioVU,[12] which links the genetic data of >100,000 individuals to their de-identified EHR data (>50,000 individuals with available pediatric EHR data).

Pediatric phecode (Peds-Phecodes) development

We mapped ICD diagnoses captured while patients were <18 years (pediatric) and ≥18 years (adult) in VUMC's EHR to current phecodes (N = 1,866) and compared the prevalence of each phecode in the pediatric and adult populations. We used the chi-square test to assess for significant differences in prevalence between the two groups if both proportions were over 5%. Otherwise, we used Fisher's exact test to assess the difference. We used a Bonferroni-corrected significance threshold to adjust for multiple testing. Additionally, we flagged phecodes with low pediatric patient representation (N<50 pediatric patients) as phecodes for potential removal, as 50 cases has been suggested to be the minimum sample size required to detect an association in PheWAS analyses of binary traits.[13]

We adopted a hybrid data- and knowledge-driven approach to develop the pediatric phecodes (Peds-Phecodes) (Figure 1). Because of the three-tiered hierarchical structure of the phecodes (root phecodes, level 1 sub-phecodes, and level 2 sub-phecodes), an important component of this work involved ensuring that all modified phecodes continued to belong to the correct phecode "tree".

From bottom-up, we pruned the phecodes without sufficient pediatric patient representation to identify potential phecodes for removal. We iteratively remapped each sub-phecode with <50 pediatric patients (from a study population of >1 million pediatric patients) to the phecode directly above it in the phecode hierarchy, starting at the most granular level of the hierarchy (level 2 sub-phecodes) and working up to the root phecode. In this way, we collapsed uncommon sub-phecodes while preserving the underlying hierarchical structure of the phecodes. For example, sub-phecodes 153.2 "Colon cancer" and 153.3 "Malignant neoplasm of rectum, rectosigmoid junction, and anus" were consolidated into their root phecode 153 "Colorectal cancer". Two pediatricians (SLV and PJK) reviewed the remaining low count phecodes to identify those that represented diseases not applicable to the pediatric population (e.g., phecode 453 "Chronic venous hypertension" and phecode 796 "Elevated prostate specific antigen"), which were then removed. Low pediatric count phecodes representing rare diseases (e.g., phecode 209 "Neuroendocrine tumors") were preserved.

From top-down, we focused on the phecodes with significantly higher prevalence in our pediatric cohort. For each phecode with higher pediatric prevalence, the two pediatricians reviewed the distribution of pediatric and adult patient diagnoses mapped to the phecode and provided recommendations for new phecodes reflecting diseases of particular importance to the pediatric population. To provide a clearer picture of the intra-phecode ICD distributions, ICD-10-CM diagnoses were converted to ICD-9-CM using the General Equivalence Mappings (GEMS) provided by the Centers for Medicare & Medicaid Services.[14] Both the ICD-9-CM codes and matching ICD-10-CM codes were remapped to the newly created phecodes. We also mapped several ICD-10-CM codes without ICD-9-CM analogs (e.g., ICD-10-CM codes related to COVID-19) to new phecodes.
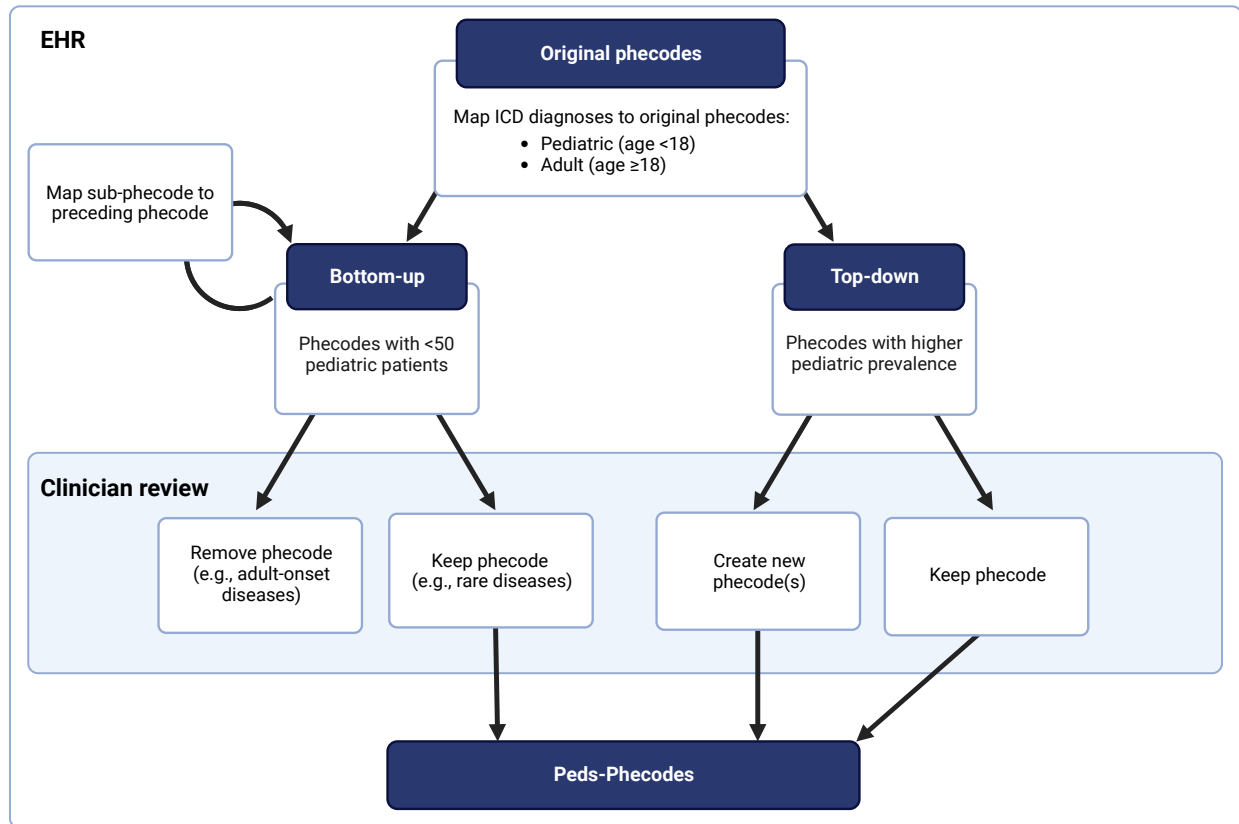
Figure 1. Workflow for Peds-Phecodes development. We used a hybrid data- and knowledge-driven approach combining patient diagnosis counts from VUMC with clinician-led manual review to prune out phenotypes with little pediatric relevance (left) and create new phecodes for important pediatric phenotypes (right).

PheWAS analyses

To validate the Peds-Phecodes, we conducted PheWAS analyses using Peds-Phecodes and the current phecodes independently and compared their ability to replicate known genotype-phenotype associations from previous pediatric studies. We queried the NHGRI-EBI GWAS Catalog[15] to find genetic variants associated with pediatric phenotypes for our PheWAS replication studies. We identified 687 SNPs in the GWAS Catalog that could be investigated using VUMC's DNA biobank. These SNPs were associated with eight different pediatric phenotypes: congenital heart disease, pyloric stenosis, Hirschsprung's disease, hypospadias, café-au-lait spots (observed in neurofibromatosis type I), pediatric eosinophilic esophagitis, childhood-onset asthma, and juvenile

idiopathic arthritis. We performed PheWAS analyses for all 687 SNPs using binary logistic regression, adjusting for sex and race. We only used diagnoses made during the pediatric age window (i.e., ICD-9-CM and ICD-10-CM codes recorded at <18 years of age) in creating the phenotypes for PheWAS. We required a minimum of two ICD codes (both recorded at <18 years of age) for a patient to be counted as a case.

As in previous studies, replication was defined as the detection of signals related to the phenotype of interest with concordant effect directionality and p<0.05.[6] We evaluated replication using both an exact and approximate phenotype match. In the exact phenotype match, we focused only on the PheWAS signal for the singular phecode best representing the phenotype of the genetic association from the GWAS Catalog (e.g., Peds-Phecode 747.1113 "Transposition of great vessels" and phecode 747.13 "Congenital anomalies of great vessels" for replication of the association between rs150246290 and transposition of the great arteries[16]). In the approximate match, we broadened our replication phenotype, examining the signals for all sub-phecodes stemming from the root phecode best matching the phenotype of the known association. For example, the approximated replication phenotype for the association between rs150246290 and transposition of the great arteries was represented by root phecode 747 "Cardiac and circulatory congenital anomalies" and all its sub-phecodes (N = 57 sub-phecodes in Peds-Phecodes, N = 5 in the current phecodes).

Beyond replication, we also evaluated the ability of the Peds-Phecodes to detect novel genetic associations compared to the existing phecodes. We compared the overlapping and unique signals generated in each PheWAS and examined the significant signals detected using Peds-Phecodes but not phecodes (p<Bonferroni-corrected p).

We developed the PedsPheWAS R package to facilitate PheWAS analysis using the Peds-Phecodes (https://github.com/The-Wei-Lab/PedsPheWAS).

Results

Differences in disease prevalence in the pediatric and adult populations

We observed marked differences in the distribution of disease between the pediatric and adult populations. Phecode 465 "Acute upper respiratory infections of multiple or unspecified sites" represented the most common overall pediatric disease phenotype, with a prevalence of 26.4% in the pediatric population compared to a prevalence of 16.9% in the adult population (p<0.001), followed by phecode 656 "Other perinatal conditions of fetus or newborn" (pediatric prevalence = 23.7%, adult prevalence = 1.7%, p<0.001). In contrast, the most common overall disease phenotype among the adult population was phecode 401 "Hypertension" (adult prevalence = 31.8%, pediatric prevalence = 1.8%, p<0.001).

Consistent with our hypothesis that pediatric and adult patients are differentially affected by disease, we found that nearly all phecodes (1,783/1,866, or 95.6%) differed significantly in prevalence in the pediatric and adult populations. We identified 317 phecodes with higher prevalence among pediatric patients at VUMC compared to adult patients and 1,466 phecodes with lower prevalence in pediatric patients compared to adults (Bonferroni-corrected p = 0.05/1,866 = 2.7×10^{-5}). The phecodes with significantly higher pediatric prevalence included congenital anomalies (N = 51), disorders affecting the sense organs (N = 30), diseases affecting the digestive system (N = 26), and injuries and poisonings (N = 26). Among the phecodes with significantly higher prevalence in the adult population were diseases of the circulatory system (N = 164), genitourinary conditions (N = 157), and endocrine/metabolic conditions (N = 136). Figure 2 shows the top five level 2 sub-phecodes with significantly higher prevalence in the pediatric population,

ordered by descending prevalence, along with the corresponding sub-phecodes with significantly higher prevalence in the adult population.
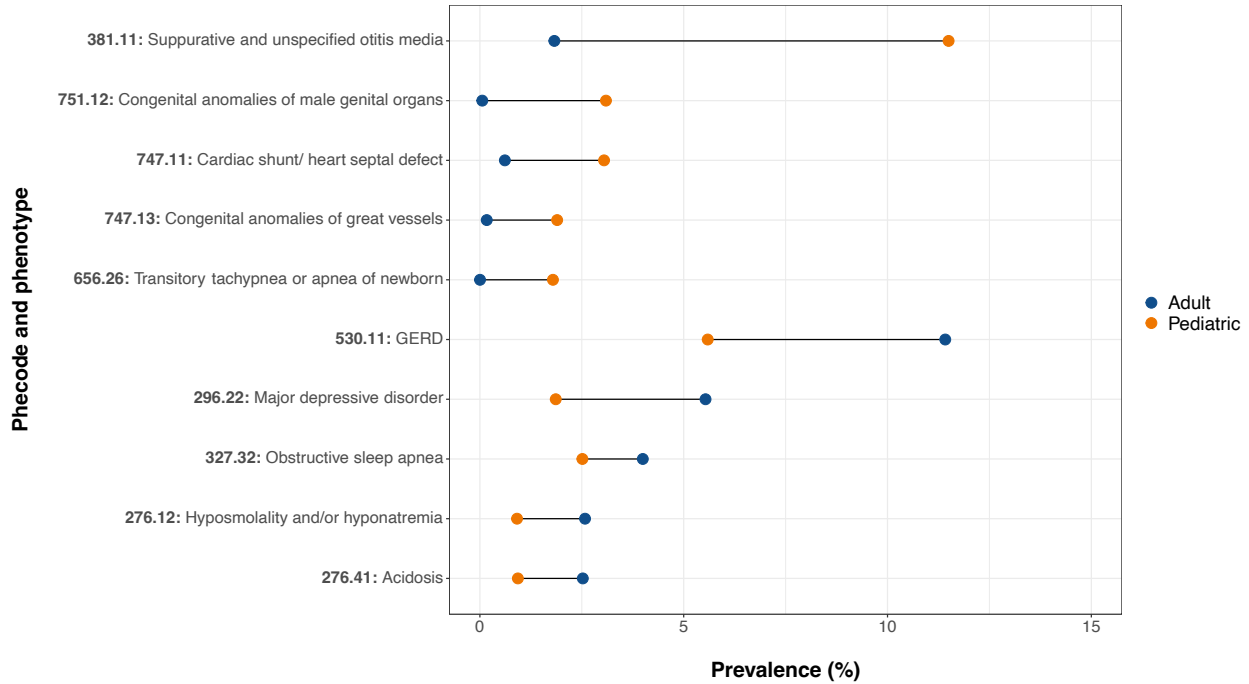

Figure 2. Comparison of phecode prevalence between the pediatric and adult populations at VUMC. Patient diagnoses at <18 years of age were mapped to the most recent phecodes version. The top five phecodes with significantly higher pediatric prevalence are displayed in the top half the dumbbell plot; the top five phecodes with significantly higher prevalence among adults are shown in the bottom half. Because of the multiple levels of phecode granularity available, direct comparisons can only be made between phecodes existing at the same level of the phecode hierarchy. Here, we show results for the level 2 (i.e., most granular) sub-phecodes in the current phecode hierarchy.

Pediatric phecodes (Peds-Phecodes)

The final Peds-Phecodes set consists of 2,051 phecodes, compared with 1,866 phecodes in the most recent phecodes iteration. The 2,051 Peds-Phecodes are subdivided into 586 root phecodes with five possible levels of sub-phecodes (N = 1,016 level 1 sub-phecodes, N = 322 level 2, N = 106 level 3, N = 18 level 4, and N = 3 level 5), in contrast to the 1,866 phecodes, which contain 585 root phecodes with two possible levels of sub-phecodes (N = 1,012 level 1, N = 269 level 2). These 2,051 phecodes provide a mapping for 15,533 ICD-9-CM codes and 82,949 ICD-10-CM

codes. As the pediatric phecodes were developed using the existing phecodes as a starting point, there is still much overlap between the two, with 1,693 shared phecodes (Figure 3). In developing the Peds-Phecodes, 173 phecodes were removed from the starting phecode set and 358 new phecodes were created. The majority of the new phecodes created (351/358) were sub-phecodes of existing root phecodes, resulting in an extension of the phecode hierarchy to a maximum of six levels rather than three.
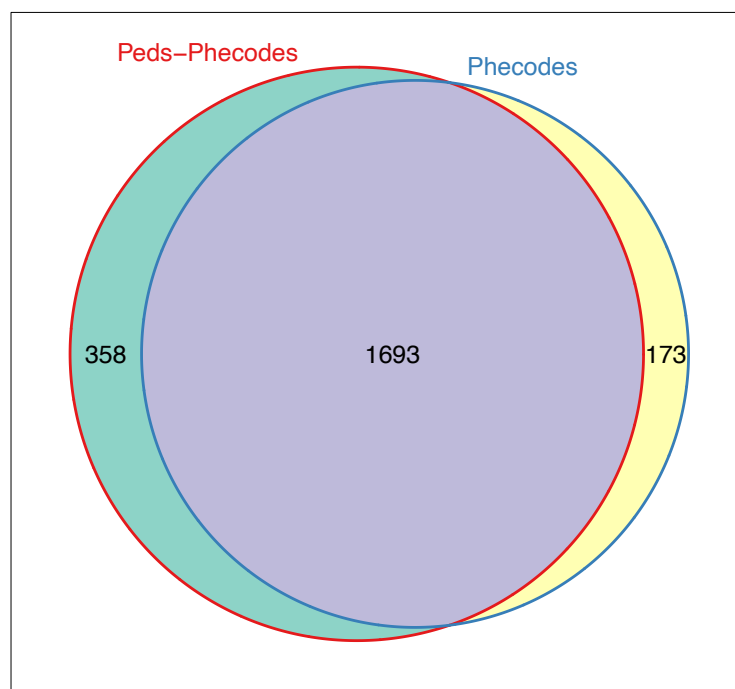
Figure 3. Comparison of Peds-Phecodes and phecodes. Weighted Venn diagram showing the overlap between Peds-Phecodes and phecodes.

Furthermore, the Peds-Phecodes were recategorized into 19 disease categories: circulatory system, dermatologic, endocrine/metabolic, digestive, genitourinary, hematologic, musculoskeletal, neurological, renal/kidney, respiratory, sense organs, congenital anomalies and genetic disorders, infections, immunologic and inflammatory disorders, neoplasms, mental disorders, pregnancy complications, symptoms, and injuries and poisonings. The renal/kidney and

immunologic and inflammatory disorders categories were newly created for the Peds-Phecodes. We sought to group the Peds-Phecodes into disease categories based on the affected anatomical location – for instance, Peds-Phecode 747 "Cardiac and circulatory congenital anomalies" was placed into the circulatory system disease category, although phecode 747 belongs to the congenital anomalies category in phecodes version 1.2.

Replication of known genotype-phenotype associations

Considering only the signal for the singular phecode best matching the phenotype of interest (i.e., exact phenotype match), use of the Peds-Phecodes replicated 88/687 (12.8%) of the SNP-phenotype associations extracted from the GWAS Catalog with $p < 0.05$, compared to 80/687 (11.6%) replicated using the phecodes. Given this relatively low replication rate for both phecode sets in the context of limited pediatric GWAS availability and the smaller pediatric sample sizes for PheWAS, we expanded our definition of the replication phenotype. Rather than using a single phecode as the replication phenotype, we defined the replication phenotype using the root phecode best matching the phenotype of interest as well as all its sub-phecodes (i.e., approximate phenotype match).

Using the approximate phenotype match, we successfully replicated 248/687 (36.1%) associations at $p < 0.05$ using the Peds-Phecodes, a significant increase from the 192/687 (27.9%) replicated using the phecodes ($p < 0.001$ using McNemar's exact test). Additionally, use of the Peds-Phecodes generally yielded smaller p-values for known genotype-phenotype associations compared to the phecodes, despite similar statistical power based on the number of identified cases (Figure 4).

15
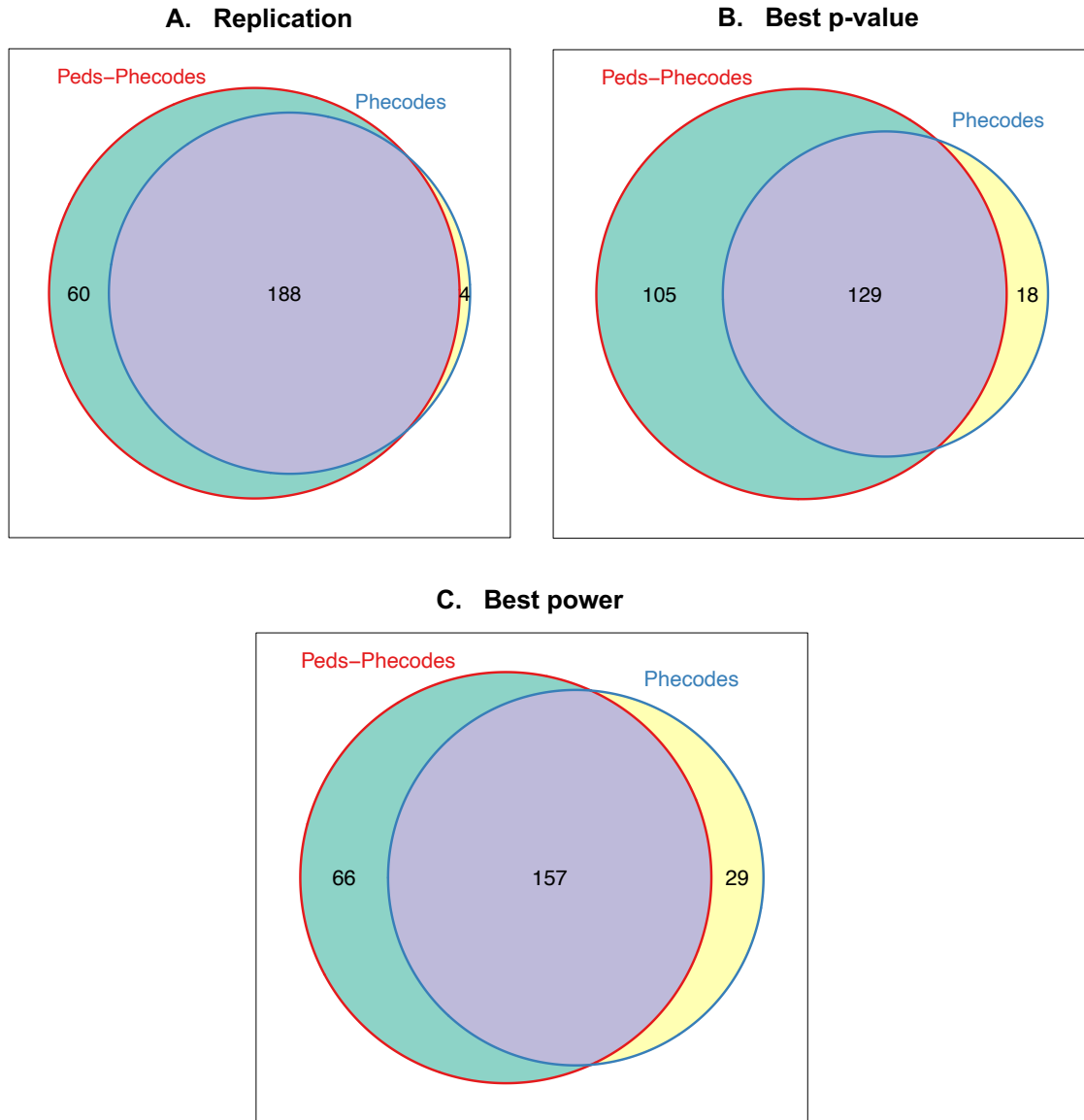
Figure 4. Performance of Peds-Phecodes and phecodes in PheWAS replication studies.
Weighted Venn diagrams comparing Peds-Phecodes and phecodes in terms of their ability to replicate known genotype-phenotype associations (panel A), the replication p-values (panel B), and statistical power (panel C).

Using Peds-Phecodes, we were able to replicate several genotype-phenotype associations that we were unable to replicate using the phecodes. In Figure 5, we show the results of our PheWAS for SNP rs2505998, a genetic variant in the *RET* gene, which has previously been shown

to be associated with Hirschsprung's disease, a congenital defect of the intestine that most commonly affects the colon and rectum.[17] The PheWAS performed using Peds-Phecodes captured the expected association between phecode 750.214 "Hirschsprung's disease and other congenital functional disorders of colon" (OR = 2.23, p = 6.02 x 10^{-4}); however, the PheWAS using the phecodes was not able to exactly replicate the association (OR = 1.13, p = 0.22);. Our PheWAS analyses also reflected the higher granularity of phenotypes available through the Peds-Phecodes. In the PheWAS for rs138741144, an *ASIC2* variant previously associated with cardiac septal defects,[16] using Peds-Phecodes revealed significant associations with phecode 747.113 "Common ventricle", phecode 747.128 "Atresia and stenosis of aorta", phecode 747.13 "Congenital anomalies of great vessels", and phecode 747.133 "Patent ductus arteriosus". In contrast, the PheWAS using the phecodes produced a much less granular signal, with only one significant association related to congenital heart disease: phecode 747.13 "Congenital anomalies of great vessels" (Figure 6).

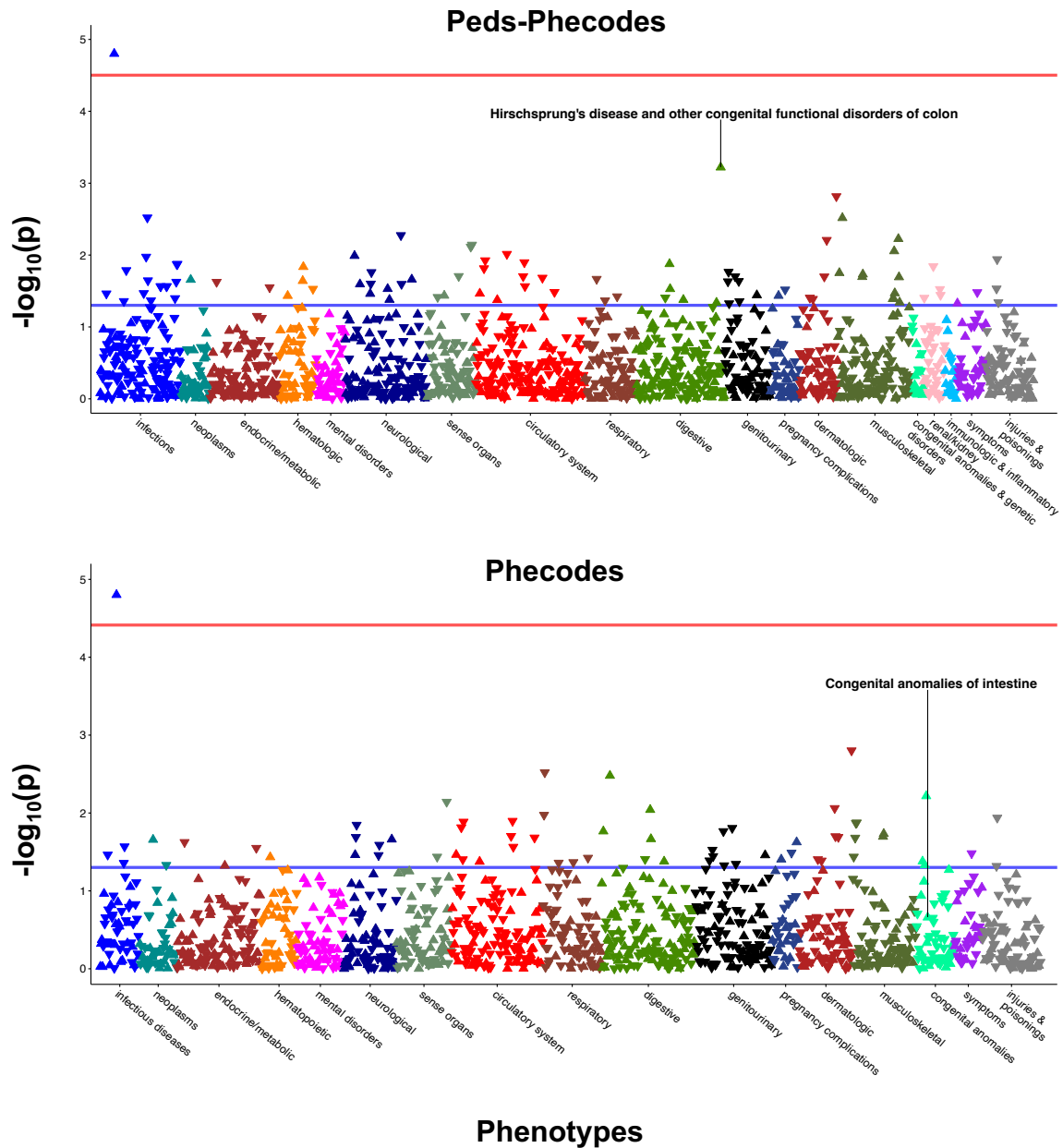**Figure 5. Comparative PheWAS analyses for rs2505998, which has previously been associated with Hirschsprung's disease.** PheWAS was performed using Peds-Phecodes (top) and phecodes (bottom). The horizontal blue line indicates $p = 0.05$; the red line indicates the Bonferroni-corrected p-value threshold. In both analyses, the most significant PheWAS association for rs2505998 was observed for phecode 070.2 "Viral hepatitis B".
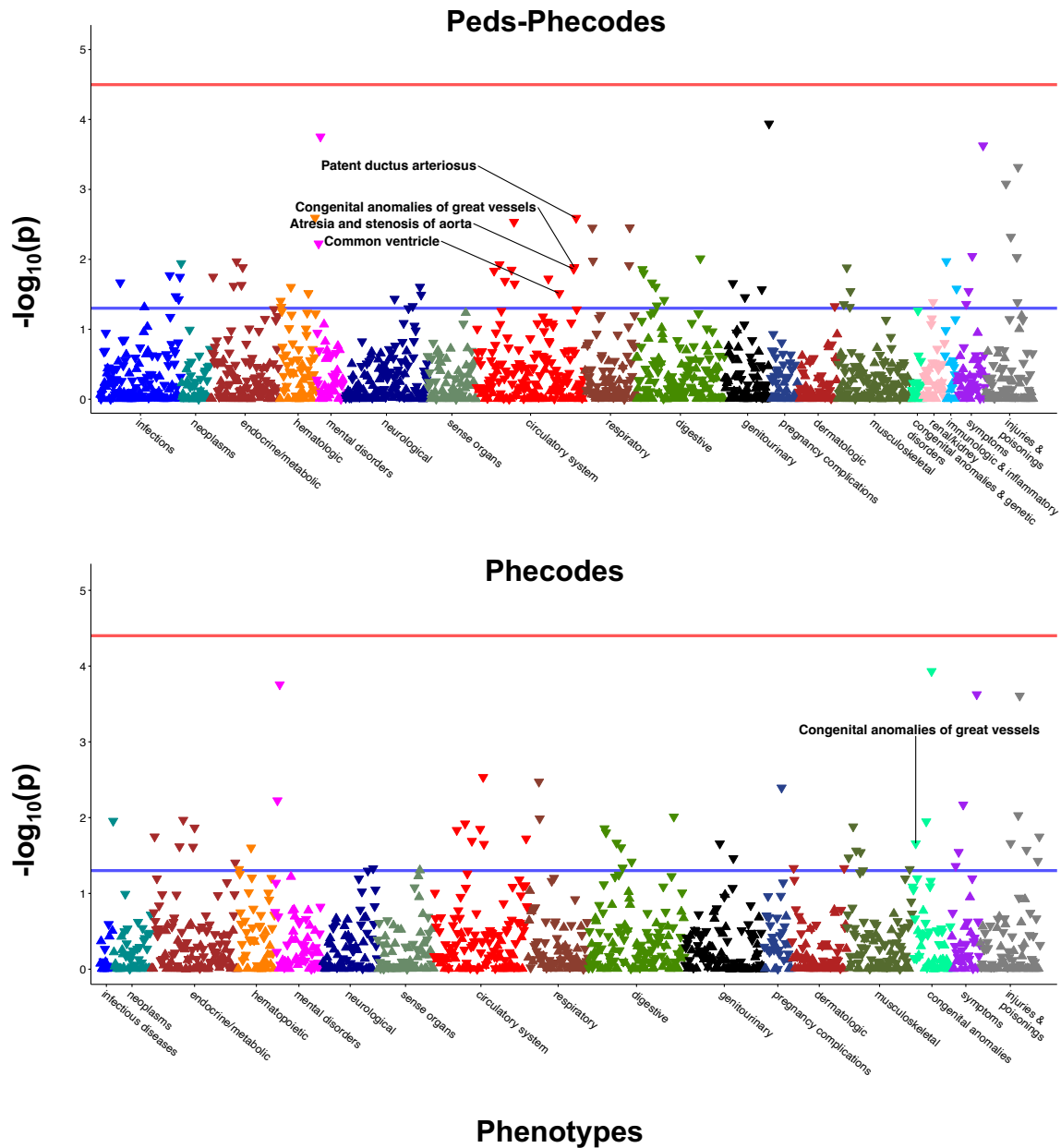
Figure 6. Comparative PheWAS analyses for rs138741144, which has previously been associated with cardiac septal defects. PheWAS was performed using Peds-Phecodes (top) and phecodes (bottom). The horizontal blue line indicates p = 0.05. No signals cross the Bonferroni-corrected p-value threshold (red line).

We also examined the overlap between the top signals produced in PheWAS using Peds-Phecodes compared to phecodes. For 99.7% (685/687) of SNP-phenotype pairs, the PheWAS

analysis using Peds-Phecodes resulted in a larger number of signals with p<0.05 compared to the PheWAS using the phecodes. We found a significant difference between the number of significant signals in the Peds-Phecodes PheWAS compared to the phecode PheWAS using a paired t-test (p<0.001). The number of overlapping signals with p<0.05 ranged from 2.3% (16/687) to 29.5% (203/687). For each genotype-phenotype association, the PheWAS using Peds-Phecodes revealed several signals (minimum N = 3, maximum N = 75) that were not detected using the phecodes, representing potentially novel SNP-phenotype associations.

To further explore the novel associations identified using the Peds-Phecodes, we examined the unique signals detected in the Peds-Phecodes PheWAS analyses applying a Bonferroni-corrected significance threshold rather than a replication p-value threshold. We found that the Peds-Phecodes identified novel associations for 147/687 (21.4%) of the SNPs tested in PheWAS, with a total of 348 novel associations detected (maximum N = 16 for the *TLR1* variant rs5743618). Interestingly, these new associations were not limited to new phecodes created when developing the Peds-Phecodes from the existing phecodes – we identified new genetic associations for 55 distinct phenotypes, 27 of which were phecodes shared by Peds-Phecodes and the phecodes. For example, we identified 97 new genetic associations for phecode 465.3 "Acute upper respiratory infection" (a new phecode) and 43 new associations for phecode 079 "Viral infection" (a shared phecode).

Discussion

The Peds-Phecodes represent a first attempt to develop a high-throughput EHR phenotyping tool specifically designed for pediatric research, with the aim of facilitating large-scale pediatric-specific phenome-wide and genome-wide studies by streamlining and standardizing the phenotyping process in a manner that reflects the unique pediatric spectrum of disease.

EHRs provide an archive of longitudinal clinical data with vast potential for research use. However, defining phenotypes efficiently and accurately for research within the EHR remains a challenging and nuanced process. Phecodes emerged from an effort to characterize the phenome by creating standardized groupings of ICD codes to represent clinically meaningful phenotypes, making it possible to perform PheWAS analyses quickly and effectively using EHR data. However, phecodes were developed using population-based diagnoses predominantly affecting adult patients, and thus are best suited for cohorts primarily composed of adult patients.

When we mapped patient diagnoses in VUMC's EHR to the current phecodes version, we observed distinct differences in the diagnoses recorded while patients were <18 years of age (i.e., pediatric patients) and those recorded after patients were ≥18 years of age. As shown in Figure 2, phecodes related to infections (e.g., otitis media) and congenital anomalies had higher prevalence in the pediatric population, whereas chronic diseases (e.g., gastroesophageal reflux disease, obstructive sleep apnea, and major depressive disorder) were more common among adults. These differences reflect the unique distributions of disease among pediatric and adult populations and highlight the need for modified phecodes for pediatric-specific PheWAS. Other studies have reported similar differences in patterns of disease among children and adults,[18,19] providing further motivation for specialized pediatric phecodes.

We developed the pediatric phecodes using a hybrid data- and knowledge-driven approach – we leveraged diagnosis code counts to help us approximate the importance (or lack thereof) of each existing phecode in the pediatric population. Diagnosis codes provided a consistent and replicable way of performing the initial assessment and significantly reduced manual review efforts. Clinical domain knowledge then fine-tuned the detected signals and ascertained their practical importance. With this insight, we modified the existing phecodes to capture diseases

primarily affecting pediatric patients with increased granularity and exclude diseases not relevant to pediatric patients, capturing clinically meaningful pediatric phenotypes. The results of our PheWAS analyses reflect the ability of Peds-Phecodes to better capture pediatric conditions. Using Peds-Phecodes, we replicated more of the genotype-phenotype associations extracted from the GWAS Catalog compared to the phecodes. We also found that using the Peds-Phecodes tended to produce smaller p-values in our PheWAS replication studies.

Peds-Phecodes substantially increased the granularity with which certain phenotypes could be characterized. In developing Peds-Phecodes, we did not remove the underlying hierarchical structure of phecodes; rather, we further extended the hierarchy vertically to accommodate a maximum of five levels of added granularity rather than two. We also extended the phecode hierarchy horizontally, incorporating alphanumeric codes to allow a phecode to directly branch into more than nine possible sub-phecodes (e.g., phecode 758.1A "Gonadal dysgenesis"). The results of our comparative PheWAS analyses highlight the advantages of the added granularity built into the Peds-Phecodes. When attempting to find an exact phenotype match for the replication studies, phecodes tended to represent phenotypes that were much broader than the target phenotype. For instance, phecodes 751.12 "Congenital anomalies of male genital organs" and 750.21 "Congenital anomalies of intestine" were the best matches for hypospadias and Hirschsprung's disease, respectively. However, the Peds-Phecodes directly represented these phenotypes (phecodes 751.124 "Hypospadias" and 750.214 "Hirschsprung's disease and other congenital functional disorders of colon"). Using an approximate phenotype match, we were able to replicate more known genotype-phenotype associations using the Peds-Phecodes largely because of their increased granularity.

Differences in the significant signals obtained in PheWAS using Peds-Phecodes and phecodes also revealed the potential of the Peds-Phecodes to detect novel associations. Although we found at least sixteen overlapping phecodes with p<0.05 in each PheWAS comparison, we also found at least three signals with p<0.05 unique to each Peds-Phecodes PheWAS for each SNP. Furthermore, we identified a total of 348 novel SNP-phenotype associations with Bonferroni significance using the Peds-Phecodes. These signals warrant further investigation, and we hope to further explore the novel associations identified using the Peds-Phecodes in large-scale pediatric PheWAS analyses in future studies.

This study has several limitations. While we created over 350 new phecodes in developing the Peds-Phecodes, we faced difficulties in finding SNPs associated with pediatric phenotypes for PheWAS validation, particularly in the realm of infectious diseases. This may in part have been due to our reliance on the GWAS Catalog to identify genetic associations; use of other repositories of genotype-phenotype data (e.g., Online Mendelian Inheritance in Man), particularly those focusing specifically on rare diseases, may have yielded other genetic variants for testing, and should be explored in subsequent studies. Still, pediatric patients are comparatively understudied in GWAS and the pediatric-specific studies available tend to focus on a relatively small number of conditions, limiting the breadth of the replication PheWAS analyses that we were able to perform. We extracted 687 SNPs associated with eight pediatric phenotypes from the GWAS Catalog. Four of these eight pediatric phenotypes were congenital anomalies, two were pediatric autoimmune diseases, one was a common inflammatory disease (asthma), and one was a rare disease (neurofibromatosis type I), providing some level of variety but nothing close to comprehensive coverage of the pediatric phenome. Our replication analyses were also limited by smaller sample sizes resulting in lower statistical power, as there are less pediatric data contained

within EHRs. However, pediatric EHR data will accumulate over time, paving the way for more high-powered PheWAS analyses in pediatric cohorts in the future. Finally, we acknowledge that the ICD code to phecode mapping will likely always have room for improvement. Knowledge assembly is an iterative process – as researchers apply the pediatric phecodes in their work, we hope to incorporate their feedback and continue to update the mapping. We will also continue to extend the pediatric phecode map to include newly added ICD-10-CM codes.

The Peds-Phecodes can be browsed and downloaded at:

https://wei-lab.app.vumc.org/phecode-data/pediatric_phecodes

Conclusion

There is an urgent need for high-throughput EHR phenotyping tools specialized for use in pediatric populations. Building on the existing phecodes, we developed new pediatric phecodes (Peds-Phecodes) using a hybrid approach integrating diagnosis data and clinical domain knowledge. Using the Peds-Phecodes to conduct PheWAS analyses, we were able to replicate significantly more known genotype-phenotype associations for pediatric conditions compared to using the phecodes, as well as identify potentially novel genotype-phenotype associations. We have made a PedsPheWAS R package to perform PheWAS using the Peds-Phecodes.

CHAPTER 3

SUMMARY

This study introduces Peds-Phecodes as a novel high-throughput EHR phenotyping tool for use in pediatric cohorts. Like the existing phecodes, the Peds-Phecodes group ICD-9-CM and ICD-10-CM codes into clinically meaningful diseases and traits; however, in contrast to phecodes, Peds-Phecodes were specifically designed to capture the unique pediatric spectrum of disease. We first demonstrated the need for specialized pediatric phecodes by confirming the presence of significant differences in disease distribution among pediatric and adult populations. We then developed the Peds-Phecodes from the existing phecodes using a hybrid data- and knowledge-driven approach. Briefly, patient diagnosis counts from VUMC were used to flag phecodes that could potentially be removed (i.e., phecodes with low pediatric counts) as well as phecodes that could be extended (i.e., phecodes with high pediatric prevalence). These initial signals were then reviewed by two clinicians, who decided how the flagged phecodes should be modified to best reflect the pediatric spectrum of disease. The final Peds-Phecodes set contained 2,051 distinct phecodes, providing a mapping for 15,533 ICD-9-CM codes and 82,949 ICD-10-CM codes to phecodes. During the development process, 173 phecodes were removed from the starting phecode set and 358 new phecodes were created.

Next, we evaluated the performance of the Peds-Phecodes in comparison to the existing phecodes using PheWAS analyses. We showed that the Peds-Phecodes replicated more known pediatric genotype-phenotype associations compared to phecodes (248 versus 192 out of 687 SNP-

phenotype associations extracted from the GWAS Catalog, p<0.001). We also demonstrated the potential of Peds-Phecodes to identify novel associations. Using the Peds-Phecodes to perform PheWAS analyses, we identified 348 SNP-phenotype associations (p<Bonferroni-corrected p) that were not detected in the PheWAS analyses using the phecodes. Together, these findings demonstrate the ability of the Peds-Phecodes to capture higher-quality pediatric phenotypes and deliver superior PheWAS outcomes compared to phecodes, suggesting that the Peds-Phecodes are an effective tool for performing large-scale phenotypic analyses in pediatric populations.

Limitations

Our study has some limitations. First, although phecodes are a well-validated tool that have been widely used to perform PheWAS analyses, it is important to acknowledge that the Peds-Phecodes were created from the existing phecodes using data-driven insights about disease prevalence from a single academic medical center (VUMC), which may limit the generalizability of the Peds-Phecodes to other pediatric populations. Future studies will need to evaluate the performance of Peds-Phecodes in pediatric-specific PheWAS using EHR data from other institutions to assess whether the Peds-Phecodes reasonably cover the pediatric clinical phenome or require additional modification. Second, both phecodes and Peds-Phecodes use only one form of structured EHR data, ICD codes, in phenotype definition, and thus may not reliably capture certain disease phenotypes, particularly for rare diseases which often do not have designated ICD codes.[47] Deeper phenotyping efforts leveraging other types of EHR data will likely be needed to appropriately phenotype patients with pediatric rare diseases.

Future directions of research

We plan to use the Peds-Phecodes to perform large-scale PheWAS analyses combining pediatric EHR data from multiple sources. We are particularly interested in further exploring the ability of Peds-Phecodes to identify novel pediatric genotype-phenotype associations. Another avenue of future research may involve extending the phecode mapping. Currently, both phecodes and Peds-Phecodes map only ICD codes to phecodes; however, other types of EHR data may also feasibly be mapped to phecodes, such as lab values, prescriptions, and concepts extracted from clinical notes. Incorporating other data types into the phecodes structure may improve phenotyping performance, particularly for rare diseases.

Conclusion

As pediatric data continues to accumulate in EHRs, there is an urgent need for high-throughput EHR phenotyping tools intended for use in pediatric cohorts. This work proposes Peds-Phecodes as a new high-throughput EHR phenotyping tool specifically designed to capture the distinctive pediatric spectrum of disease, building on the existing concept of phecodes. Our evaluation showed that Peds-Phecodes delivered superior results in pediatric-specific PheWAS analyses compared to phecodes. We expect that the Peds-Phecodes will facilitate future large-scale phenomic and genomic analyses in pediatric populations, contributing to new opportunities for scientific discovery in pediatrics.

REFERENCES

1   Burde H. THE HITECH ACT: An Overview. *AMA Journal of Ethics* 2011;**13**:172–5. doi:10.1001/virtualmentor.2011.13.3.hlaw1-1103

2   Adler-Milstein J, DesRoches CM, Kralovec P, *et al.* Electronic Health Record Adoption In US Hospitals: Progress Continues, But Challenges Persist. *Health Aff (Millwood)* 2015;**34**:2174–80. doi:10.1377/hlthaff.2015.0992

3   GOLD M, McLAUGHLIN C. Assessing HITECH Implementation and Lessons: 5 Years Later. *Milbank Q* 2016;**94**:654–87. doi:10.1111/1468-0009.12214

4   Atasoy H, Greenwood BN, McCullough JS. The Digitization of Patient Care: A Review of the Effects of Electronic Health Records on Health Care Quality and Utilization. *Annu Rev Public Health* 2019;**40**:487–500. doi:10.1146/annurev-publhealth-040218-044206

5   Banda JM, Seneviratne M, Hernandez-Boussard T, *et al.* Advances in Electronic Phenotyping: From Rule-Based Definitions to Machine Learning Models. *Annu Rev Biomed Data Sci* 2018;**1**:53–68. doi:10.1146/annurev-biodatasci-080917-013315

6   Kho AN, Hayes MG, Rasmussen-Torvik L, *et al.* Use of diverse electronic medical record systems to identify genetic risk for type 2 diabetes within a genome-wide association study. *J Am Med Inform Assoc* 2012;**19**:212–8. doi:10.1136/amiajnl-2011-000439

7   Zheng HW, Ranganath VK, Perry LC, *et al.* Evaluation of an automated phenotyping algorithm for rheumatoid arthritis. *J Biomed Inform* 2022;**135**:104214. doi:10.1016/j.jbi.2022.104214

8   Slaby I, Hain HS, Abrams D, *et al.* An electronic health record (EHR) phenotype algorithm to identify patients with attention deficit hyperactivity disorders (ADHD) and psychiatric comorbidities. *J Neurodev Disord* 2022;**14**:37. doi:10.1186/s11689-022-09447-9

9   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. doi:10.1093/jamia/ocv202

10 Savova GK, Masanz JJ, Ogren PV, *et al.* Mayo clinical Text Analysis and Knowledge Extraction System (cTAKES): architecture, component evaluation and applications. *J Am Med Inform Assoc* 2010;**17**:507–13. doi:10.1136/jamia.2009.001560

11 Strauss JA, Chao CR, Kwan ML, *et al.* Identifying primary and recurrent cancers using a SAS-based natural language processing algorithm. *J Am Med Inform Assoc* 2013;**20**:349–55. doi:10.1136/amiajnl-2012-000928

12 Seneviratne MG, Banda JM, Brooks JD, *et al.* Identifying Cases of Metastatic Prostate Cancer Using Machine Learning on Electronic Health Records. *AMIA Annu Symp Proc* 2018;**2018**:1498–504.

13 Ogunyemi OI, Gandhi M, Lee M, *et al.* Detecting diabetic retinopathy through machine learning on electronic health record data from an urban, safety net healthcare system. *JAMIA Open* 2021;**4**:ooab066. doi:10.1093/jamiaopen/ooab066

14 Carroll RJ, Eyler AE, Denny JC. Naïve Electronic Health Record Phenotype Identification for Rheumatoid Arthritis. *AMIA Annu Symp Proc* 2011;**2011**:189–96.

15 Hong N, Wen A, Stone DJ, *et al.* Developing a FHIR-based EHR phenotyping framework: A case study for identification of patients with obesity and multiple comorbidities from discharge summaries. *J Biomed Inform* 2019;**99**:103310. doi:10.1016/j.jbi.2019.103310

16 Pacheco JA, Rasmussen LV, Wiley K, *et al.* Evaluation of the portability of computable phenotypes with natural language processing in the eMERGE network. *Sci Rep* 2023;**13**:1971. doi:10.1038/s41598-023-27481-y

17 Bastarache L. Using Phecodes for Research with the Electronic Health Record: From PheWAS to PheRS. *Annu Rev Biomed Data Sci* 2021;**4**:1–19. doi:10.1146/annurev-biodatasci-122320-112352

18 Denny JC, Ritchie MD, Basford MA, *et al.* PheWAS: demonstrating the feasibility of a phenome-wide scan to discover gene-disease associations. *Bioinformatics* 2010;**26**:1205–10. doi:10.1093/bioinformatics/btq126

19 Roden DM. Phenome-wide association studies: a new method for functional genomics in humans. *J Physiol* 2017;**595**:4109–15. doi:10.1113/JP273122

20 Bastarache L, Denny JC, Roden DM. Phenome-Wide Association Studies. *JAMA* 2022;**327**:75–6. doi:10.1001/jama.2021.20356

21 Carroll RJ, Bastarache L, Denny JC. R PheWAS: data analysis and plotting tools for phenome-wide association studies in the R environment. *Bioinformatics* 2014;**30**:2375–6. doi:10.1093/bioinformatics/btu197

22 Denny JC, Bastarache L, Ritchie MD, *et al.* Systematic comparison of phenome-wide association study of electronic medical record data and genome-wide association study data. *Nat Biotechnol* 2013;**31**:1102–10. doi:10.1038/nbt.2749

23 Cartwright DJ. ICD-9-CM to ICD-10-CM Codes: What? Why? How? *Adv Wound Care (New Rochelle)* 2013;**2**:588–92. doi:10.1089/wound.2013.0478

24 Wei W-Q, Bastarache LA, Carroll RJ, *et al.* Evaluating phecodes, clinical classification software, and ICD-9-CM codes for phenome-wide association studies in the electronic health record. *PLoS One* 2017;**12**:e0175508. doi:10.1371/journal.pone.0175508

25 Wu P, Gifford A, Meng X, *et al.* Mapping ICD-10 and ICD-10-CM Codes to Phecodes: Workflow Development and Initial Evaluation. *JMIR Med Inform* 2019;**7**:e14325. doi:10.2196/14325

26 Namjou B, Marsolo K, Carroll R, *et al.* Phenome-wide association study (PheWAS) in EMR-linked pediatric cohorts. *Frontiers in Genetics* 2014;**5**.https://www.frontiersin.org/articles/10.3389/fgene.2014.00401 (accessed 14 Aug 2023).

27 Gutiérrez-Sacristán A, Sáez C, De Niz C, *et al.* Multi-PheWAS intersection approach to identify sex differences across comorbidities in 59 140 pediatric patients with autism spectrum disorder. *Journal of the American Medical Informatics Association* 2022;**29**:230–8. doi:10.1093/jamia/ocab144

28 Fu M, Chang TS. Phenome-Wide Association Study of Polygenic Risk Score for Alzheimer's Disease in Electronic Health Records. *Front Aging Neurosci* 2022;**14**:800375. doi:10.3389/fnagi.2022.800375

29 Hyppönen E, Mulugeta A, Zhou A, *et al.* A data-driven approach for studying the role of body mass in multiple diseases: a phenome-wide registry-based case-control study in the UK Biobank. *The Lancet Digital Health* 2019;**1**:e116–26. doi:10.1016/S2589-7500(19)30028-7

30 Mahnke CB. The growth and development of a specialty: the history of pediatrics. *Clin Pediatr (Phila)* 2000;**39**:705–14. doi:10.1177/000992280003901204

31 Wojcik MH, Schwartz TS, Thiele KE, *et al.* Infant Mortality: the Contribution of Genetic Disorders. *J Perinatol* 2019;**39**:1611–9. doi:10.1038/s41372-019-0451-5

32 Madrid L, Seale AC, Kohli-Lynch M, *et al.* Infant Group B Streptococcal Disease Incidence and Serotypes Worldwide: Systematic Review and Meta-analyses. *Clin Infect Dis* 2017;**65**:S160–72. doi:10.1093/cid/cix656

33 Li Z, Zhang Z, Ren Y, *et al.* Aging and age-related diseases: from mechanisms to therapeutic strategies. *Biogerontology* 2021;**22**:165–87. doi:10.1007/s10522-021-09910-5

34 Safarova MS, Satterfield BA, Fan X, *et al.* A phenome-wide association study to discover pleiotropic effects of PCSK9, APOB, and LDLR. *npj Genomic Med* 2019;**4**:1–9. doi:10.1038/s41525-019-0078-7

35 Li S, Jia Z, Zhang Z, *et al.* Association Study of Genetic Variants in Calcium Signaling-Related Genes With Cardiovascular Diseases. *Front Cell Dev Biol* 2021;**9**:642141. doi:10.3389/fcell.2021.642141

36 Zhu X, Li Y, Yu T, *et al.* A hypothesis-driven study to comprehensively investigate the association between genetic polymorphisms in EPHX2 gene and cardiovascular diseases: Findings from the UK Biobank. *Gene* 2022;**822**:146340. doi:10.1016/j.gene.2022.146340

37 Kerchberger VE, Peterson JF, Wei W-Q. Scanning the medical phenome to identify new diagnoses after recovery from COVID-19 in a US cohort. *Journal of the American Medical Informatics Association* 2023;**30**:233–44. doi:10.1093/jamia/ocac159

38 Danciu I, Cowan JD, Basford M, *et al.* Secondary Use of Clinical Data: the Vanderbilt Approach. *J Biomed Inform* 2014;**52**:28–35. doi:10.1016/j.jbi.2014.02.003

39 Roden DM, Pulley JM, Basford MA, *et al.* Development of a large-scale de-identified DNA biobank to enable personalized medicine. *Clin Pharmacol Ther* 2008;**84**:362–9. doi:10.1038/clpt.2008.89

40 Verma A, Bradford Y, Dudek S, *et al.* A simulation study investigating power estimates in phenome-wide association studies. *BMC Bioinformatics* 2018;**19**:120. doi:10.1186/s12859-018-2135-0

41 2018 ICD-10 CM and GEMs | CMS. https://www.cms.gov/Medicare/Coding/ICD10/2018-ICD-10-CM-and-GEMs (accessed 11 May 2023).

42 Sollis E, Mosaku A, Abid A, *et al.* The NHGRI-EBI GWAS Catalog: knowledgebase and deposition resource. *Nucleic Acids Res* 2023;**51**:D977–85. doi:10.1093/nar/gkac1010

43 Lahm H, Jia M, Dreßen M, *et al.* Congenital heart disease risk loci identified by genome-wide association study in European patients. *J Clin Invest* 2021;**131**:e141837, 141837. doi:10.1172/JCI141837

44 Tang CS, Gui H, Kapoor A, *et al.* Trans-ethnic meta-analysis of genome-wide association studies for Hirschsprung disease. *Hum Mol Genet* 2016;**25**:5265–75. doi:10.1093/hmg/ddw333

45 Kuan V, Denaxas S, Gonzalez-Izquierdo A, *et al.* A chronological map of 308 physical and mental health conditions from 4 million individuals in the English National Health Service. *The Lancet Digital Health* 2019;**1**:e63–77. doi:10.1016/S2589-7500(19)30012-3

46 Fraser LK, Lidstone V, Miller M, *et al.* Patterns of diagnoses among children and young adults with life-limiting conditions: A secondary analysis of a national dataset. *Palliat Med* 2014;**28**:513–20. doi:10.1177/0269216314528743

47 Fung KW, Richesson R, Bodenreider O. Coverage of Rare Disease Names in Standard Terminologies and Implications for Patients, Providers, and Research. *AMIA Annu Symp Proc* 2014;**2014**:564–72.