Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

HHS Public Access

Author manuscript

Annu Rev Biomed Data Sci. Author manuscript; available in PMC 2022 July 22.

Published in final edited form as:

Annu Rev Biomed Data Sci. 2021 July 20; 4: 1–19. doi:10.1146/annurev-biodatasci-122320-112352.

Using Phecodes for Research with the Electronic Health Record: From PheWAS to PheRS

Lisa Bastarache

Department of Biomedical Informatics, Vanderbilt University Medical Center, Nashville, Tennessee 37232, USA

Abstract

Electronic health records (EHRs) are a rich source of data for researchers, but extracting meaningful information out of this highly complex data source is challenging. Phecodes represent one strategy for defining phenotypes for research using EHR data. They are a high-throughput phenotyping tool based on ICD (International Classification of Diseases) codes that can be used to rapidly define the case/control status of thousands of clinically meaningful diseases and conditions. Phecodes were originally developed to conduct phenome-wide association studies to scan for phenotypic associations with common genetic variants. Since then, phecodes have been used to support a wide range of EHR-based phenotyping methods, including the phenotype risk score. This review aims to comprehensively describe the development, validation, and applications of phecodes and suggest some future directions for phecodes and high-throughput phenotyping.

Keywords

phecodes; electronic health record; phenotyping; phenome-wide association study (PheWAS); genomics; Mendelian genetics; phenotype risk score; phenotype risk score (PheRS)

INTRODUCTION

The growth of electronic health records (EHRs) provides today's researchers with a wealth of opportunities. The primary purpose of an EHR is to facilitate patient care and hospital billing. In serving these purposes, the EHR stores information about patient symptoms, findings, and diagnoses over time. This information can be used to address scientific and medical questions in ways that have never before been possible. Creative methods to leverage EHR data have proliferated in the last decade, and it is likely that researchers have only begun to scratch the surface.

The sheer size and complexity of many EHRs can be overwhelming. Clinical notes, lab results, pathology reports, radiology reports, claims data, and medications—all these data elements are present in the EHR, though often not in a form that is easy to use. Information

lisa.bastarache@vumc.org .

DISCLOSURE STATEMENT

L.B. receives royalty payments from Nashville Biosciences, a Vanderbilt University Medical Center–owned entity.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

about patients is incomplete and, at times, contradictory or incorrect. To harness the potential of EHRs, researchers must find a way to extract quality phenotypes from a dense, complex, fragmented, and noisy entity.

EHR-based phenotyping methods can be roughly divided into two categories. First, there are validated algorithms created for a specific phenotype. These algorithms tie together multiple sources of data from the EHR—including lab results, medications, concepts extracted from notes using natural language processing, and billing codes—to identify cases and controls with maximum performance. Second, there are high-throughput phenotyping methods that attempt to define a wide array of phenotypes via automated processes. Their goal is to capture some of the enormous breadth of the EHR to support large-scale studies of the medical phenome.

Phecodes belong to the latter category. They are a high-throughput means of capturing thousands of phenotypes from the EHR. What phecodes lack in specificity and sensitivity, they gain in breadth, portability, and ease of use. This review describes how phecodes were created and validated, gives examples of some of their applications, and explains how phecodes may be used by researchers to gain a foothold in the complex realm of EHR phenotyping.

WHAT IS A PHECODE?

Phecodes are manually curated groups of International Classification of Diseases (ICD) codes intended to capture clinically meaningful concepts for research. They were created to rapidly characterize a wide swath of diagnoses, symptoms, and findings—a phenome-wide snapshot—needed to conduct phenome-wide association studies (PheWAS). Today, phecodes are used in a variety of EHR phenotyping methods.[1]

International Classification of Diseases Codes

The ICD is a medical classification list of codes for diseases, symptoms, findings, and injuries that is maintained by the World Health Organization (WHO). ICD codes are used by more than 100 countries to track global morbidity and mortality data (2). In some countries, ICD codes are integrated into the EHR to record medical findings of patients in a standardized format, which are then used to estimate healthcare costs by tracking the complexity and frequency of care. Today, over 70% of the world's health expenditures are allocated with ICD (3).

The ICD has evolved significantly over its long history, and the WHO periodically releases new versions (4, 5). In the context of EHR-based research, only the two most recent versions of the ICD, the ninth and tenth revisions (ICD-9 and ICD-10, respectively), are relevant. The United States uses an extended version of ICD called the "clinical modification" (CM)

[1]Phecodes were originally called "PheWAS codes" before the term "phecode" was introduced in 2015 (1), and they are still sometimes called by that name. In this review, medical findings, symptoms, and diagnoses are often referred to collectively as the "medical phenome" or "phenotypes," which is a general term used to describe any observable trait in an organism and is typically used in the context of genetics. Calling medical diagnoses "phenotypes" reflects the initial purpose of phecodes to facilitate genetics research.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

developed by the National Center for Health Statistics (6). While ten nations use ICD codes for reimbursement (including Canada, the United Kingdom, and Sweden), only the United States uses the CM extension (7).

In US hospitals, ICD-9-CM codes were used from the 1980s until 2015, when ICD-10-CM was adopted (8). While both code sets capture many of the same medical conditions, the coding structure is fundamentally different, and ICD-10-CM has over five times the number of diagnosis codes as ICD-9-CM (9). To facilitate the transition from ICD-9 to ICD-10, the Centers for Medicare and Medicaid Studies (CMS) released a publicly available map between the two versions (10).

Researchers using ICD codes should keep two key facts in mind: (*a*) First, ICD codes from the United States are not directly compatible with other countries that use ICD for reimbursement. (*b*) Second, EHR systems in the United States that have data from before and after 2015 contain both ICD-9-CM and ICD-10-CM codes.

Grouping International Classification of Diseases Codes into Useful Phenotypes

As ICD codes accumulated in hospital administrative databases, investigators began using them for healthcare research (11, 12). To facilitate secondary use, researchers have often defined groups of ICD codes to capture their outcome of interest (13, 14). Grouping ICD codes is necessary because ICD codes encode granularity that may not be useful in certain applications. For example, there are 20 ICD-9-CM codes and 113 ICD-10-CM for type 1 diabetes (T1D), including codes for T1D with various complications like ketoacidosis, renal disease, neurological manifestations, etc. Any one of these codes may be used to indicate a T1D diagnosis. ICD groupings are often created for one or a small set of phenotypes, and typically the groupings are created using domain expertise and consensus (15, 16). Cases of a given disease or condition are defined based on the presence of one of the codes in the grouping for that disease or condition. (Cases may be required to have multiple instances of the same phecode, as described below in the section titled The Rule of Two.)

Phecodes are an effort to scale this grouping process to include all available ICD diagnosis codes. Defining which level of granularity is useful is subjective and depends on the application. For phecodes, the decision of which concepts to define was made on the basis of clinical and researcher judgement, with a focus on capturing common adult diagnoses that would be useful in genetic association studies. The groupings were created manually and informed by the Clinical Classification Software (CCS) grouping schema, as well as the prevalence of codes in the EHRs of multiple medical centers.

Phecodes are maintained by the Center for Precision Medicine at Vanderbilt University Medical Center and are available at https://www.phewascatalog.org/phecodes. Revisions are published periodically and users can submit feedback through the website. The latest version of phecodes (version 1.2) includes 1,867 phecodes for different diagnoses, symptoms, and findings.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

Adding International Classification of Diseases, Tenth Revision, Clinical Modification

Phecodes were developed using ICD-9-CM codes. The adoption of ICD-10-CM codes in 2015 necessitated the definition of a new map linking ICD-10-CM codes into the existing phecode structure. Unlike the manual effort required to define phecodes, the ICD-10 to phecode map was developed primarily through an automated process using general equivalence mappings provided by the CCS, as well as mappings available in the United Medical Language System (UMLS), and validated with genetic replication studies (17). Phecodes version 1.2 condenses roughly 15,500 ICD-9-CM codes and 90,000 ICD-10-CM codes into 1,867 phecodes.

Phecode Hierarchy

Phecodes are hierarchical, which allows them to capture phenotypes at multiple levels of granularity. The hierarchy is an attempt to accommodate different research questions and cohort sizes. There is a phecode for the specific "Intestinal infection due to *C. difficile*" (008.52), as well as the more general "Bacterial enteritis" (008.5) and "Intestinal infection" (008). Parent phecodes are three digits. Digits may be added to a parent code following a decimal point. Each additional digit trailing indicates a subset of the ICD codes of the parent code. A child code like 008.52 implies 008.5 as well as 008 (Figure 1).

Dealing with Noise

Researchers who use ICD diagnosis codes from the EHR must be mindful of potential inaccuracies. Accuracy of ICD codes varies widely across diagnoses and cohorts, with studies finding accuracies ranging from less than 50% (18) to nearly 100% (19).

There are multiple reasons why an ICD code may not reflect a patient's true underlying disease (20). Some errors are simple typos (21).[2] In other cases, an ICD that is more familiar may be incorrectly used for a related condition (22). Because ICD codes are used for reimbursement, some errors can be traced back to financial incentives. Upcoding is the practice of inappropriately assigning a code with a higher reimbursement. Upcoding is illegal, and insurance companies and other regulatory bodies attempt to prevent it, but it still happens (23). The proportion of upcoded ICD codes has been shown to vary between institutions and across time (24).

Sometimes the use of an ICD code is technically correct in that it adheres to CMS coding guidelines (25) but is incorrect with respect to the truth about the patient. As O'Mallay et al. have described in a marvelous review of ICD accuracy, a diagnosis is an expression of probability, not a black-and-white statement of fact (26). Diagnoses can evolve and change as more information is gathered. Along this pathway, patients may acquire codes for conditions that they do not actually have.

[2]Typos are common in ICD-9 coding for T1D and T2D due to the fact that the codes for the two are interlaced (e.g., codes 250.01, 250.03, 250.11, 250.13, etc., are used for T1D, and 250.02, 250.04, 250.12, and 250.14 are used for T2D).

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

The Rule of Two

Because of these inaccuracies, defining cases of a given disease or condition based on a patient's being coded for a single phecode can result in false positives. One simple strategy to mitigate this problem is to require a patient to be coded for a phecode on multiple unique encounter dates to count as a case. In a study of 6,005 patients, the performance of four phecodes was tested against a manually reviewed gold standard (27) (Figure 2). The F_1 score (a measure of a classification's precision and recall) was calculated as a function of the number of codes required for defining a case. The study concluded that the maximum average F_1 score was achieved when cases were defined as having two or more phecodes. Based on this and subsequent work, the so-called rule of two was defined (28, 29). Phecode-based analyses like PheWAS commonly use this rule, and it is the default setting for the PheWAS R package (30).

The rule of two is not ideal in all circumstances and phenotypes. Indeed, even in the initial small study (27), using two or more phecodes only maximized the F_1 score for one of the four phenotypes tested. Requiring two codes may induce high false negative rates, particularly for acute and time-limited conditions. The ideal code count threshold will also vary based on characteristics of the EHR cohort. For example, cohorts of patients with records that span many years may improve phecode accuracy using a higher minimum code count requirement. PheProb is a method that obviates the need to define a minimum code count threshold by calculating the probability an individual has a phenotype based on the number of diagnosis codes in their EHR. In their study, the authors show that PheProb increases the statistical power for replicating known genetic associations (31). Such a method may be useful when a strict case/control status is not necessary.

Exclude Ranges and Defining Controls

Each phecode has an exclude range that can be used to filter out controls with related conditions. For example, phecode 555.1 (Crohn's disease) has an exclude range of 555.00–564.99, which en-compasses a group of phecodes for noninfective gastrointestinal disorders. Exclude ranges were modeled after exclusion criteria commonly used in case/control studies (32). Phecode version 1.2 currently defines 232 unique exclude ranges and each phecode is assigned an exclude range.

The primary rationale for exclude ranges was to mitigate the noisiness of ICD coding, where less specific or related codes may be used for a specific condition. For example, a specific underlying condition (e.g., Basal cell carcinoma) may only be recognized as a nonspecific entity (e.g., "Neoplasm of uncertain behavior of skin" in ICD-9) pending a further diagnostic workup.

While exclude ranges are a logical response to the noisiness of ICD codes, they have the potential to introduce bias. In any case/control study, exclusion criteria can compromise generalizability (33, 34). Applying phecode exclude ranges subtly changes the hypothesis in a way that is hard to understand, and so they should be used with caution. Exclude ranges also induce a high rate of missingness, which can be problematic for methods that require

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

every individual to be classified for a particular disease. Future study could help improve the function of exclude ranges and clarify their effects.

Sex-Specific Codes

Each phecode is labeled by sex specificity. Overall, 2% of phecodes are male specific and 7% are female specific. For codes that are sex specific, the sex of the patient is used as an inclusion criterion. Therefore, the phecode for prostate cancer only includes males as either cases and controls; all others are labeled as missing. This functionality is built into the PheWAS R package and can also be implemented using the sex column in the phecode definition file (https://phewascatalog.org/files/phecode_definitions1.2.csv).

Phecode Chapters

Each phecode belongs to one of 18 chapters representing broad organ systems or categories. These categories are based on ICD-9 chapters. Phecode chapters are used to visualize data in PheWAS plots, where phecodes from the same chapter are plotted proximally along the *x*-axis and are assigned a distinct color. They can also be used to restrict a PheWAS to a particular domain of interest.

While PheWAS is typically used as a hypothesis-free scan of all available phenotypes, sometimes a researcher is only interested in phenotypes from a particular domain. Chapters can be used to focus a PheWAS on phenotypes of interest, excluding tests that are not relevant to the study and easing the multiple comparison correction by reducing the overall number of tests. This technique was used in an analysis of ABO blood type and cancer risks, where only phecodes from the neoplasm chapter were tested (35). To discourage data snooping, researchers must make the decision to restrict a PheWAS by chapter prior to running the analysis rather than applying this filtering after the full PheWAS is generated.

The number of phecodes and the degree of compression (i.e., the number of ICD codes per phecode) across chapters are highly variable (Figure 3). The differences reflect, in part, properties intrinsic to the ICD coding structure, as well as the focus of phecodes on capturing common adult diseases with potentially genetic etiologies. There are many ICD codes relating to injuries and these are highly compressed in phecodes. In contrast, phecodes in the endocrine/metabolic chapter maintain much of the granularity available in the ICD coding structure itself. Future versions of phecodes may expand the granularity of phenotypes in chapters like “congenital anomalies,” which currently groups many specific ICD codes into a single nonspecific phecode.

Validating Phecodes

Phecodes are the product of a manual curation, and the map is the culmination of a series of subjective decisions regarding which phenotypes to define and how to define them. Since their creation lacks a formal theory, efforts to validate phecodes were empirically critical to demonstrating their utility.

EHR phenotypes can be validated in several ways. One method is to use a gold standard based on manual chart review. Performance metrics like positive predictive value (PPV)

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

can be calculated to compare phecode case/control assignments to a gold standard. This approach was used in the eMERGE (Electronic Medical Records and Genomics) Network to test and validate EHR phenotype algorithms (36). Studies comparing phecodes against manually reviewed gold standards have found a wide range of PPVs (37, 38). The reasons for the heterogeneity of phecode performance is likely multifactorial. Performance depends on the phenotype tested, as well as the average record length, data density, and patient mix of the EHR cohort. The influence of these factors makes it difficult to generalize on the results of any one study.

When available, genetic data can be used as an alternative approach to validate EHR phenotypes. In this case, the performance of a phenotype is based on its ability to replicate known phenotype/SNP associations (32, 39). A benefit of genetic validation is that it does not require manual chart review and is more easily applied to multiple cohorts.

A 2013 study developed a framework to do genetic validation at scale using the NHGRI-EBI (National Human Genome Research Institute–European Bioinformatics Institute) GWAS Catalog as the source of replication candidates. The GWAS Catalog is a curated resource of statistically significant associations between traits and SNPs found in published GWAS (40). Traits in the GWAS Catalog were mapped to 86 corresponding phecodes (Figure 4). Through this linkage, researchers created a SNP/phecode map (SPM) that related phecodes to an associated SNP. Each SNP/phecode pair was annotated with the ancestries of the discovery cohort, and the odds ratios from the original study were used to calculate the minimum number of cases required to replicate the association at a nominal $p < 0.05$. The study replicated 51 out of 77 sufficiently powered associations at $p < 0.05$ across a diverse set of phenotypes (41).

In this and other phecode replication studies, the odds ratios from the phecode-based associations are generally closer to 1 than the original GWAS finding (32). This is likely due to the noisiness of phecodes relative to the phenotypes used in the initial study. The winner’s curse also likely plays a role, as described in an interesting paper by Palmer & Pe’er about internal replication rates of loci in the GWAS Catalog (42).

Replication for Calibration and Quality Control

Genetic replication rates can be used to compare phenotyping methods applied to the same dataset. Hughey et al. used the SPM to demonstrate the relative superiority of using Cox regression over logistic regression to detect known associations (43). An interesting alternative method for genetic validation is to use a genome-wide heritability estimate, as was done in a study comparing different EHR-based bipolar disorder algorithms (44).

Genetic validation may also be used as a quality control measure on EHR datasets that are linked to genetic data. The SPM can be used to calculate a replication rate for any cohort with EHR and genetic data. If a dataset were somehow corrupted (for example, if the genotype data were linked to the wrong patients), then the replication rate would be suppressed. A formal study of replication rates has not been conducted and would be useful for enhancing this technique as a quality control measure.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

The current version of the SPM comprises 588 catalog traits mapped to 163 phecodes (Table 1); the map is available for download at https://www.phewascatalog.org/refmap. Because associations have a low replication rate across ancestries (42), subsets of the SPM were created to represent three genetic ancestry categories. [The relative dearth of SNP/phenotype associations in Asian and African ancestries is due to the unfortunate lack of genetic studies on individuals of non-European ancestries (45).] When calculating a replication rate, researchers should use the SPM that matches the ancestry of their cohort.

Other High-Throughput Coding Systems

Phecodes are not the only comprehensive mapping of ICD diagnosis codes used for research. Some researchers condense ICD codes into parent codes consisting of the first three digits/characters of the code (e.g., ICD-10 codes G61.0 through G61.9 are collapsed into G61). The Neale lab used this process in their massive catalog of UK Biobank GWAS results (46). The Agency for Health Research and Quality developed a manually curated grouping of ICD-9-CM and ICD-10-CM codes called CCS (47, 48). A study comparing phenotypes from phecodes, CCS, and ICD-9-CM top codes found that phecodes were more likely to replicate known genetic associations in the GWAS Catalog (49).

Rasmy et al. tested the suitability of five different ICD mappings to produce phenotypes for predictive modeling. Despite having far fewer unique codes, the UMLS (Unified Medical Language System) and phecode mapping performed similarly and were superior to other mappings like CCS (50). Their results suggest that phecodes represent an efficient compression of ICD-based data.

Zhang et al. explored a data-driven approach to grouping both ICD-9 and ICD-10 codes to address coding heterogeneity across health systems and reduce the bias inherent in manual groupings (51). Data-driven approaches may be used on their own or as a way of informing future manual efforts in grouping ICD codes.

PHECODE APPLICATION 1: PHEWAS

Phecodes were created to enable PheWAS, a method that requires a phenome-wide characterization of a cohort.[3] PheWAS has been reviewed extensively elsewhere (52, 53), so this section will briefly describe recent developments in PheWAS as they relate to phecodes and high-throughput phenotyping.

PheWAS was initially used to scan for associations between the phenome and a SNP. The results of a PheWAS analysis on a genetic variant can reveal pleiotropy—a phenomenon where a single genetic variant is associated with multiple phenotypic traits (54). In this capacity, PheWAS is often used as a way to learn more about genetic associations newly discovered using EHR data. First, an EHR-based phenotype algorithm is developed,

[3]"Phenome-wide association studies" is a play on the term "genome-wide association studies," but the analogy is imperfect. A GWAS samples loci of human genetic variation across a large by finite string of base pairs. But a PheWAS is a not sampling of human variation from head to toe, and no such standardized nomenclature exists to define the full phenome. The term "PheWAS" is used to describe an association study that loops through many unrelated human phenotypes, and there is no consensus on how many phenotypes must be tested to constitute a PheWAS.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

validated, and used to conduct a GWAS. Then PheWAS is applied to the significantly associated SNPs to scan for additional phenotypic associations (55–59).

PheWAS has been applied to genetic data beyond SNPs, including human leukocyte antigen types (60), gene expression levels (61–63), functional genetic variants (64), and genetic instrument variables from Mendelian randomization (65–67). PheWAS has also been used to explore the phenome of nongenetic variables to study sleep quality (68), race (69), disease comorbidities (70), healthcare costs (71), COVID-19 outcomes (72), and even the health and wellness of musicians (73).

Not every PheWAS uses phecodes, either by necessity or by design. Some studies use raw ICD codes instead of condensing them into phecodes (74, 75). Other studies are based on non-ICD sources like imaging data (76), birth defect registries (77), or lab results (78). For cross-cohort studies, a hybrid approach may be necessary. Diogo et al. conducted a study on four cohorts with different phenotypic data (79). A cross site meta-analysis was conducted using a manual map to harmonize phenotypes, from phecode, medical interviews, and surveys.

Recent work has suggested methodological improvements of the PheWAS method by incorporating temporal or contextual information associated with ICD codes. One study showed that the ability to replicate known genetic associations with phecodes is improved by using a Cox regression, which takes advantage of the temporal information encoded in longitudinal EHR data (43). Another PheWAS found that phecodes were strongly correlated with the types of clinics patients visited, suggesting that visit type might be taken into account to improve the quality of phecode phenotypes (80).

HIGH-THROUGHPUT PHENOTYPING AND MENDELIAN GENETICS

The study of complex diseases and the study of Mendelian diseases represent two separate domains of human genetics research, each with its own methods, tools, and tendencies (Table 2). Thus far, EHR-linked biobanks have primarily been used to study complex diseases (81, 82). As the cost of genetic sequencing has decreased over the past decade, resources linking EHR data to whole-exome or whole-genome sequencing are beginning to proliferate (83–85). In response to this exciting development, new phenotyping tools and methods must be developed that reconcile two very different approaches to studying human genetics.

At first glance, phecodes may seem ill-suited to contribute to the study of Mendelian diseases. Mendelian geneticists describe phenotypes with exquisite precision, noting an upturn of the nose, rotation of the ears, or curve of the fourth toe. Phecode-based phenotypes are, in comparison, very crude. Phecodes were developed for high-throughput methods applied to large cohorts, sacrificing accuracy for speed, scalability, and breadth. They are the factory-made version of the Mendelian geneticist’s bespoke creations.

Nuanced phenotyping is a cornerstone of Mendelian genetics. Indeed, the founders of this field made incredible advances through careful observation of patients’ traits combined with the logic of Mendelian inheritance. In 1966, Victor McKusick published a catalog of

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

heritable diseases in a resource that is now known as the Online Mendelian Inheritance in Man (OMIM) (86). By the time the *CFTR* gene was identified in 1989, OMIM had more than 4,000 disease entries, each with its own detailed description of clinical manifestations (87). Advances in gene mapping technology led to the rapid discovery of genes underlying the conditions described in OMIM. By the year 2000, OMIM contained entries for over 1,000 diseases linked to specific genes (88). Today, that number has grown more than fivefold (89).

The study of Mendelian diseases led to enormous advances in the understanding of many heritable conditions. However, its explanatory power only extends to a relatively small number of people. Doubts about the potential for Mendelian genetics to help the average person have been around for a while. A 1968 news article likened the medical geneticists at their annual meeting in Bar Harbor to medieval scholars debating how many angels could dance on the head of a pin: “This provided theologians with lasting employment, but did little to help the common man enter the kingdom of heaven.… [G]eneticists must be capable of more significant research” (90). While the proportion of people affected by monogenic conditions has increased in recent years—with some estimates as high as 10%—the majority of human disease cannot be attributed to rare monogenic mutations (91).

Common diseases like diabetes and heart disease are often referred to as complex. In contrast to monogenic conditions, complex diseases are due to the confluence of polygenic and environmental risks (92). The genetic study of complex diseases took off following the first successful GWAS in 2002 (93). In contrast to the study of Mendelian genetics, research in complex genetics was high throughput and hypothesis free, leveraging data from large populations. While Mendelian disease research identified causal variants, the associations discovered with GWAS were statistical in nature, typically with small effect sizes, and not well understood at the mechanistic level (94, 95). In short, Mendelian genetics explains a lot about a small number of patients, and complex genetics explains a little bit about everyone.

However, as several researchers have pointed out, the dichotomy of monogenic versus polygenic disease is artificial. Complex and Mendelian diseases actually exist on a continuum, with full-penetrance Mendelian diseases caused by rare variants on one end of the spectrum and complex disease risk caused by common variants on the other (96, 97).

To begin to bridge the gap, researchers have begun developing methods and resources that borrow the principles, knowledge, and methods from both domains. The Human Phenotype Ontology (HPO) was developed as a controlled vocabulary to annotate clinical descriptions in OMIM (98). Researchers have used HPO to find phenotypic commonalities in genes linked to both Mendelian and complex diseases (99) and to facilitate the development of population-based methodologies for the study of monogenic disease (100). A map that links HPO terms to phecodes was created to search for Mendelian disease patterns in the EHR at scale using a method called the phenotype risk score (PheRS).

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

PHECODE APPLICATION 2: PHENOTYPE RISK SCORES

PheRS is a method developed to leverage EHR data for the study of Mendelian diseases. It is a continuous measurement of the similarity between an individual and the clinical description of a Mendelian disease. The disease patterns are defined using OMIM's clinical descriptions of diseases, which have been mapped to HPO terms. HPO terms are linked to phecodes, so any given Mendelian disease in OMIM can be automatically described in terms of phenotypes that are easily extracted from the EHR (101) (Figure 5). Features are weighted based on their prevalence in a study population, such that unusual phenotypes contribute more to the score than common ones.

Many HPO terms do not have an exactly matching phecode. In these cases, a broader phecode is used. Thus, a specific feature like "exocrine pancreatic insufficiency" is mapped to the more general phecode 577 ("diseases of the pancreas"). The phecode 577 groups together the nonspecific ICD-9 code with several other conditions of the pancreas like acute pancreatitis. Including a broader phecode in a PheRS will decreases the specificity, while leaving it out would sacrifice sensitivity. Decisions about how to strike a balance between these two criteria need to be made based on the application of the PheRS. Future work might further refine phenotypic inputs to increase performance overall.

In a proof-of-concept study, the PheRS of patients diagnosed with a particular genetic disease were elevated compared with controls, suggesting that PheRS can be used to differentiate individuals affected by a Mendelian disease without using the disease label itself (101).

As a continuous score, PheRS can be elevated for a patient who only partially matches a clinical description. This is important for two reasons. First, genetic diseases vary in terms of their expressivity. *CFTR*, the gene that causes cystic fibrosis, is a case in point. Some *CFTR* mutations cause classical cystic fibrosis, while other mutations only affect the lungs or pancreas (102). Variable expressivity can even occur between individuals with the same mutation (103). An algorithmic approach to identifying individuals with *CFTR* mutations must allow for partial phenotypic overlap. Second, EHRs are known to have a substantial amount of missing information, which is a major challenge for using them in research (104, 105). Thus, methods that use EHR data must be tolerant of missing data.

While the examples given thus far have focused on cystic fibrosis, a PheRS can be created from any one of the thousands of diseases described in OMIM, enabling a high-throughput scan of rare genetic variants. In one study, a PheRS was calculated for 1,204 diseases to scan 6,000 rare genetic variants, identifying associations with PheRS and rare genetic variants that were replicated in external cohorts (101). Other researchers have created PheRS by combining the phenotypic features of multiple Mendelian diseases. Ye et al. found that patients with a high PheRS for monogenic aortopathy were more likely to have adverse outcomes and more likely to have pathogenic mutations found in genetic testing for aortopathy (106).

Applications of PheRS have extended beyond Mendelian diseases and rare variants. Zhong et al. found that the genetically predicted expression of *CFTR* was associated with PheRS

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

for cystic fibrosis (107). PheRS has also been used to characterize complex non-Mendelian phenotypes such as pancreatic cancer (108) and major depressive disorder (109).

PheRS might be used to identify undiagnosed patients using EHR data. In a study of ten patients who were diagnosed with cystic fibrosis as adults, eight had a PheRS above the ninety-fifth percentile prior to diagnosis (110). Figure 6 illustrates the PheRS trajectory for one of these individuals over time. While these results are encouraging, achieving the specificity necessary to find ultrarare patients remains a challenge. Future developments of the method may incorporate more elements of EHRs (e.g., clinical notes or labs) to enhance the specificity of PheRS.

TOWARD SYNTHESIS: USING HYBRID PHENOTYPING MODELS

ICD-based phenotyping is powerful because it utilizes data that are ubiquitous, easy to manipulate, and relatively standardized. While there is undeniable heterogeneity in the way ICD codes are assigned across time and place, concerted efforts are made to standardize their use. And while ICD codes only capture a finite number of phenotypes represented in the EHR, they do represent many of the diagnoses and symptoms needed for research.

However, the EHR is so much more than just ICD codes. Information encoded in lab results, clinical notes, pathology reports, and the like can be leveraged to increase the accuracy and granularity of phenotypes. Several methods have been developed using machine learning techniques to integrate multiple EHR data types, including PheNorm, MAP (multimodal automated phenotyping), and PheCAP (111–113). These methods have been shown to increase phenotype accuracy compared to ICD or phecodes alone, although implementing them involves significant effort upfront.

The last decade has seen amazing progress in the development of EHR-based phenotyping methods, but there is still a lot of work to be done. Future development on phenotype methods should prioritize scalability, portability, and interpretability, alongside accuracy, in order to more fully realize the research potential of EHRs.

ACKNOWLEDGMENTS

I would like to thank Adam Lewis for his suggestions on this work. This work was supported by the grant R01-LM010685 from the National Library of Medicine.

LITERATURE CITED

1. Leader JB, Pendergrass SA, Verma A, Carey DJ, Hartzel DN, et al. 2015. Contrasting association results between existing PheWAS phenotype definition methods and five validated electronic phenotypes. AMIA Annu. Symp. Proc 2015:824–32 [PubMed: 26958218]
2. WHO (World Health Organ.). 2020. International Classification of Diseases (ICD) information sheet. Fact Sheet, World Health Organ. https://www.who.int/standards/classifications/classification-of-diseases
3. Beck DE, Margolin DA. 2007. Physician coding and reimbursement. Ochsner. J 7(1):8–15 [PubMed: 21603473]
4. WHO (World Health Organ.). 2020. History of the development of the ICD. Fact Sheet, World Health Organ. https://www.who.int/classifications/icd/en/HistoryOfICD.pdf

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

5. Hirsch JA, Nicola G, McGinty G, Liu RW, Barr RM, et al. 2016. ICD-10: history and context. Am. J. Neuroradiol 37(4):596–99 [PubMed: 26822730]
6. NCHS (Natl. Cent. Health Stat.). 2015. International Classification of Diseases, Ninth Revision, Clinical Modification (ICD-9-CM). Web Resour., Natl. Cent. Health. Stat., Hyattsville, MD. https://www.cdc.gov/nchs/icd/icd9cm.htm
7. Manchikanti L, Kaye AD, Singh V, Boswell MV. 2015. The tragedy of the implementation of ICD-10-CM as ICD-10: Is the cart before the horse or is there a tragic paradox of misinformation and ignorance? Pain Physician 18(4):E485–95 [PubMed: 26218946]
8. NCHS (Natl. Cent. Health Stat.). 2020. International Classification of Diseases, Tenth Revision, Clinical Modification (ICD-10-CM). Web Resour., Natl. Cent. Health. Stat., Hyattsville, MD. https://www.cdc.gov/nchs/icd/icd10cm.htm
9. Topaz M, Shafran-Topaz L, Bowles KH. 2013. ICD-9 to ICD-10: evolution, revolution, and current debates in the United States. Perspect. Health Inf. Manag 10(Spring):1d
10. CMS (Cent. Medicare Medicaid Serv.). 2011. ICD-10-CM/PCS to ICD-9-CM reimbursement mappings. User Guide, Cent. Medicare Medicaid Serv., Baltimore, MD. https://www.cms.gov/Medicare/Coding/ICD10/downloads/2011_Reimbursement_Mapping_User_Guide.pdf
11. Iezzoni LI. 1990. Using administrative diagnostic data to assess the quality of hospital care: pitfalls and potential of ICD-9-CM. Int. J. Technol. Assess. Health Care 6(2):272–81 [PubMed: 2203703]
12. Jencks SF. 1992. Accuracy in recorded diagnoses. JAMA 267(16):2238–39 [PubMed: 1556801]
13. Cherkin DC, Deyo RA, Volinn E, Loeser JD. 1992. Use of the International Classification of Diseases (ICD-9-CM) to identify hospitalizations for mechanical low back problems in administrative databases. Spine 17(7):817–25 [PubMed: 1386943]
14. Deyo RA, Cherkin DC, Ciol MA. 1992. Adapting a clinical comorbidity index for use with ICD-9-CM administrative databases. J. Clin. Epidemiol 45(6):613–19 [PubMed: 1607900]
15. Alessandrini EA, Alpern ER, Chamberlain JM, Shea JA, Gorelick MH. 2010. A new diagnosis grouping system for child emergency department visits. Acad. Emerg. Med 17(2):204–13 [PubMed: 20370751]
16. Rassekh SR, Lorenzi M, Lee L, Devji S, McBride M, Goddard K. 2010. Reclassification of ICD-9 codes into meaningful categories for oncology survivorship research. J. Cancer Epidemiol 2010:569517 [PubMed: 21234317]
17. Wu P, Gifford A, Meng X, Li X, Campbell H, et al. 2019. Mapping ICD-10 and ICD-10-CM codes to phecodes: workflow development and initial evaluation. JMIR Med. Inform 7(4):e14325 [PubMed: 31553307]
18. Cipparone CW, Withiam-Leitch M, Kimminau KS, Fox CH, Singh R, Kahn L. 2015. Inaccuracy of ICD-9 codes for chronic kidney disease: a study from two practice-based research networks (PBRNs). J. Am. Board Fam. Med 28(5):678–82 [PubMed: 26355142]
19. Khurshid S, Keaney J, Ellinor PT, Lubitz SA. 2016. A simple and portable algorithm for identifying atrial fibrillation in the electronic medical record. Am. J. Cardiol 117(2):221–25 [PubMed: 26684516]
20. Lloyd SS, Rissing JP. 1985. Physician and coding errors in patient records. JAMA 254(10):1330–36 [PubMed: 3927014]
21. Klompas M, Eggleston E, McVetta J, Lazarus R, Li L, Platt R. 2013. Automated detection and classification of type 1 versus type 2 diabetes using electronic health record data. Diabetes Care 36(4):914–21 [PubMed: 23193215]
22. Rhodes ET, Laffel LMB, Gonzalez TV, Ludwig DS. 2007. Accuracy of administrative coding for type 2 diabetes in children, adolescents, and young adults. Diabetes Care 30(1):141–43 [PubMed: 17192348]
23. Sheshasayee A, Thomas SS. 2017. Implementation of data mining techniques in upcoding fraud detection in the monetary domains. In 2017 International Conference on Innovative Mechanisms for Industry Applications (ICIMIA), pp. 730–34. New York: IEEE
24. Silverman E, Skinner J. 2004. Medicare upcoding and hospital ownership. J. Health Econ 23(2):369–89 [PubMed: 15019762]

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

25. NCHS (Natl. Cent. Health Stat.). 2020. ICD-10-CM official guidelines for coding and reporting: FY 2020. Report. Guidel., Natl. Cent. Health. Stat., Hyattsville, MD. https://www.cdc.gov/nchs/data/icd/10cmguidelines-FY2020_final.pdf
26. O'Malley KJ, Cook KF, Price MD, Wildes KR, Hurdle JF, Ashton CM. 2005. Measuring diagnoses: ICD code accuracy. Health Serv. Res 40(5 Pt. 2):1620–39 [PubMed: 16178999]
27. Bastarache L, Denny JC. 2011. The use of ICD-9 codes in genetic association studies. AMIA Annu. Symp. Proc 2011:1738
28. Ye Z, Mayer J, Ivacic L, Zhou Z, He M, et al. 2015. Phenome-wide association studies (PheWASs) for functional variants. Eur. J. Hum. Genet 23(4):523–29 [PubMed: 25074467]
29. Verma A, Ritchie MD. 2017. Current scope and challenges in phenome-wide association studies. Curr. Epidemiol. Rep 4(4):321–29 [PubMed: 29545989]
30. Carroll RJ, Bastarache L, Denny JC. 2014. R PheWAS: data analysis and plotting tools for phenome-wide association studies in the R environment. Bioinformatics 30(16):2375–76 [PubMed: 24733291]
31. Sinnott JA, Cai F, Yu S, Hejblum BP, Hong C, et al. 2018. PheProb: probabilistic phenotyping using diagnosis codes to improve power for genetic association studies. J. Am. Med. Inform. Assoc 25(10):1359–65 [PubMed: 29788308]
32. Denny JC, Ritchie MD, Basford MA, Pulley JM, Bastarache L, et al. 2010. PheWAS: demonstrating the feasibility of a phenome-wide scan to discover gene-disease associations. Bioinformatics 26(9):1205–10 [PubMed: 20335276]
33. Pearce N, Checkoway H. 1988. Case-control studies using other diseases as controls: problems of excluding exposure-related diseases. Am. J. Epidemiol 127(4):851–56 [PubMed: 3281448]
34. Velentgas P, Dreyer NA, Nourjah P, Smith SR, Torchia MM, eds. 2013. Developing a Protocol for Observational Comparative Effectiveness Research: A User's Guide. Rockville, MD: Agency Healthc. Res. Quality
35. Beeghly-Fadiel A, Giri A, Bastarache L, Pully J, Warner J, Denny J. 2017. ABO blood type and cancer risk: preliminary findings from a phenome analysis. Proc. AACR Annu. Meet 77(13):1293 (Abstr.)
36. Newton KM, Peissig PL, Kho AN, Bielinski SJ, Berg RL, et al. 2013. Validation of electronic medical record-based phenotyping algorithms: results and lessons learned from the eMERGE network. J. Am. Med. Inform. Assoc 20(e1):e147–54 [PubMed: 23531748]
37. Teixeira PL. 2015. Computational phenotyping and phenome-wide association studies: leveraging machine learning and natural language processing to understand electronic health record data. PhD Thesis, Vanderbilt Univ., Nashville, TN
38. Liao KP, Sparks JA, Hejblum BP, Kuo I, Cui J, et al. 2017. Phenome-wide association study of autoantibodies to citrullinated and noncitrullinated epitopes in rheumatoid arthritis. Arthritis Rheumatol. 69(4):742–49 [PubMed: 27792870]
39. Ritchie MD, Denny JC, Crawford DC, Ramirez AH, Weiner JB, et al. 2010. Robust replication of genotype-phenotype associations across multiple diseases in an electronic medical record. Am. J. Hum. Genet 86(4):560–72 [PubMed: 20362271]
40. Welter D, MacArthur J, Morales J, Burdett T, Hall P, et al. 2014. The NHGRI GWAS Catalog, a curated resource of SNP-trait associations. Nucleic Acids Res. 42(D1):D1001–6 [PubMed: 24316577]
41. Denny JC, Bastarache L, Ritchie MD, Carroll RJ, Zink R, et al. 2013. Systematic comparison of phenome-wide association study of electronic medical record data and genome-wide association study data. Nat. Biotechnol 31(12):1102–11 [PubMed: 24270849]
42. Palmer C, Pe'er I. 2017. Statistical correction of the Winner's Curse explains replication variability in quantitative trait genome-wide association studies. PLOS Genet. 13(7):e1006916 [PubMed: 28715421]
43. Hughey JJ, Rhoades SD, Fu DY, Bastarache L, Denny JC, Chen Q. 2019. Cox regression increases power to detect genotype-phenotype associations in genomic studies using the electronic health record. BMC Genom. 20:805

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

44. Chen C-Y, Lee PH, Castro VM, Minnier J, Charney AW, et al. 2018. Genetic validation of bipolar disorder identified by automated phenotyping using electronic health records. Transl. Psychiatry 8:86 [PubMed: 29666432]
45. Morales J, Welter D, Bowler EH, Cerezo M, Harris LW, et al. 2018. A standardized framework for representation of ancestry data in genomics studies, with application to the NHGRI-EBI GWAS Catalog. Genome Biol. 19:21 [PubMed: 29448949]
46. Neale Lab. 2020. UK Biobank results. Web Resour, Neale Lab., Cambridge, MA. https://www.nealelab.is
47. AHRQ (Agency Healthc. Res. Qual.). 2019. Clinical classifications software (CCS) for ICD-10-PCS (beta version). Web Resour., Agency Healthc. Res. Qual., Rockville, MD. https://www.hcup-us.ahrq.gov/toolssoftware/ccs10/ccs10.jsp
48. Lu T-H, Jen I, Chou Y-J, Chang H-J. 2005. Evaluating the comparability of different grouping schemes for mortality and morbidity. Health Policy 71(2):151–59 [PubMed: 15607378]
49. Wei W-Q, Bastarache LA, Carroll RJ, Marlo JE, Osterman TJ, et al. 2017. Evaluating phecodes, clinical classification software, and ICD-9-CM codes for phenome-wide association studies in the electronic health record. PLOS ONE 12(7):e0175508 [PubMed: 28686612]
50. Rasmy L, Tiryaki F, Zhou Y, Xiang Y, Tao C, et al. 2020. Representation of EHR data for predictive modeling: a comparison between UMLS and other terminologies. J. Am. Med. Inform. Assoc 27(10):1593–99 [PubMed: 32930711]
51. Zhang L, Zhang Y, Cai T, Ahuja Y, He Z, et al. 2019. Automated grouping of medical codes via multiview banded spectral clustering. J. Biomed. Inform 100:103322 [PubMed: 31672532]
52. Denny JC, Bastarache L, Roden DM. 2016. Phenome-wide association studies as a tool to advance precision medicine. Annu. Rev. Genom. Hum. Genet 17:353–73
53. Pendergrass SA, Brown-Gentry K, Dudek SM, Torstenson ES, Ambite JL, et al. 2011. The use of phenome-wide association studies (PheWAS) for exploration of novel genotype-phenotype relationships and pleiotropy discovery. Genet. Epidemiol 35(5):410–22 [PubMed: 21594894]
54. Safarova MS, Satterfield BA, Fan X, Austin EE, Ye Z, et al. 2019. A phenome-wide association study to discover pleiotropic effects of *PCSK9*, *APOB*, and *LDLR*. NPJ Genom. Med 4:3 [PubMed: 30774981]
55. Denny JC, Crawford DC, Ritchie MD, Bielinski SJ, Basford MA, et al. 2011. Variants near *FOXE1* are associated with hypothyroidism and other thyroid conditions: using electronic medical records for genome- and phenome-wide studies. Am. J. Hum. Genet 89(4):529–42 [PubMed: 21981779]
56. Namjou B, Lingren T, Huang Y, Parameswaran S, Cobb BL, et al. 2019. GWAS and enrichment analyses of non-alcoholic fatty liver disease identify new trait-associated genes and pathways across eMERGE Network. BMC Med. 17:135 [PubMed: 31311600]
57. Veatch OJ, Bauer CR, Keenan BT, Josyula NS, Mazzotti DR, et al. 2020. Characterization of genetic and phenotypic heterogeneity of obstructive sleep apnea using electronic health records. BMC Med. Genom 13:105
58. Namjou B, Stanaway IB, Lingren T, Mentch FD, Benoit B, et al. 2020. Evaluation of the *MC4R* gene across eMERGE network identifies many unreported obesity-associated variants. Int. J. Obes 45:155–69
59. Klarin D, Verma SS, Judy R, Dikilitas O, Wolford BN, et al. 2020. Genetic architecture of abdominal aortic aneurysm in the Million Veteran Program. Circulation 142:1633–46 [PubMed: 32981348]
60. Karnes JH, Bastarache L, Shaffer CM, Gaudieri S, Xu Y, et al. 2017. Phenome-wide scanning identifies multiple diseases and disease severity phenotypes associated with HLA variants. Sci. Transl. Med 9(389):eaai8708 [PubMed: 28490672]
61. Unlu G, Gamazon ER, Qi X, Levic DS, Bastarache L, et al. 2019. *GRIK5* genetically regulated expression associated with eye and vascular phenomes: discovery through iteration among biobanks, electronic health records, and zebrafish. Am. J. Hum. Genet 104(3):503–19 [PubMed: 30827500]
62. Unlu G, Qi X, Gamazon ER, Melville DB, Patel N, et al. 2020. Phenome-based approach identifies *RIC1*-linked Mendelian syndrome through zebrafish models, biobank associations and clinical studies. Nat. Med 26:98–109 [PubMed: 31932796]

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

63. Roden DM. 2017. Phenome-wide association studies: a new method for functional genomics in humans. J. Physiol 595(12):4109–15 [PubMed: 28229460]
64. Emdin CA, Khera AV, Chaffin M, Klarin D, Natarajan P, et al. 2018. Analysis of predicted loss-of-function variants in UK Biobank identifies variants protective for disease. Nat. Commun 9:1613 [PubMed: 29691411]
65. Millard LAC, Davies NM, Timpson NJ, Tilling K, Flach PA, Smith GD. 2015. MR-PheWAS: hypothesis prioritization among potential causal effects of body mass index on many outcomes, using Mendelian randomization. Sci. Rep 5:16645 [PubMed: 26568383]
66. Robinson JR, Carroll RJ, Bastarache L, Chen Q, Mou Z, et al. 2020. Association of genetic risk of obesity with postoperative complications using Mendelian randomization. World J. Surg 44:84–94 [PubMed: 31605180]
67. Rosa M, Chignon A, Li Z, Boulanger M-C, Arsenault BJ, et al. 2019. A Mendelian randomization study of IL6 signaling in cardiovascular diseases, immune-related disorders and longevity. NPJ Genom. Med 4:1–10 [PubMed: 30675382]
68. Dashti HS, Cade BE, Stutaite G, Saxena R, Redline S, Karlson EW. 2020. Sleep health, diseases, and pain syndromes: findings from an electronic health record biobank. Sleep 2020:zsaa189
69. Pulley JM, Jerome RN, Bernard GR, Shirey-Rice JK, Xu Y, Wilkins CH. 2021. The astounding breadth of health disparity: phenome-wide effects of race on disease risk. J. Natl. Med. Assoc 113:187–94 [PubMed: 32958289]
70. Zhang T, Goodman M, Zhu F, Healy B, Carruthers R, et al. 2020. Phenome-wide examination of comorbidity burden and multiple sclerosis disease severity. Neurol. Neuroimmunol. Neuroinflamm 7(6):e864 [PubMed: 32817202]
71. Cai W, Cagan A, He Z, Ananthakrishnan AN. 2021. A phenome-wide analysis of healthcare costs associated with inflammatory bowel diseases. Dig. Dis. Sci 66:760–67 [PubMed: 32436120]
72. Salvatore M, Gu T, Mack JA, Prabhu Sankar S, Patil S, et al. 2020. A phenome-wide association study (PheWAS) of COVID-19 outcomes by race using the electronic health records data in Michigan Medicine. medRxiv 2020.06.29.20141564. 10.1101/2020.06.29.20141564
73. Niarchou M, Lin G, Lense MD, Gordon RL, Davis LK. 2020. The medical signature of Nashville musicians: a phenome-wide association study using Vanderbilt's electronic health record database. medRxiv 2020.08.14.20175109. 10.1101/2020.08.14.20175109
74. Hebbring SJ, Schrodi SJ, Ye Z, Zhou Z, Page D, Brilliant MH. 2013. A PheWAS approach in studying *HLA-DRB1*1501*. Genes Immunity 14(3):187–91 [PubMed: 23392276]
75. Verma A, Lucas A, Verma SS, Zhang Y, Josyula N, et al. 2018. PheWAS and beyond: the landscape of associations with medical diagnoses and clinical measures across 38,662 individuals from Geisinger. Am. J. Hum. Genet 102(4):592–608 [PubMed: 29606303]
76. Zhao L, Batta I, Matloff W, O'Driscoll C, Hobel S, Toga AW. 2021. Neuroimaging PheWAS (phenome-wide association study): a free cloud-computing platform for big-data, brain-wide imaging association studies. Neuroinformatics 19:285–303 [PubMed: 32822005]
77. Schraw JM, Langlois PH, Lupo PJ. 2020. Comprehensive assessment of the associations between maternal diabetes and structural birth defects in offspring: a phenome-wide association study. Ann. Epidemiol 53:14–20.e8 [PubMed: 32920098]
78. Goldstein JA, Weinstock JS, Bastarache LA, Larach DB, Fritsche LG, et al. 2020. LabWAS: novel findings and study design recommendations from a meta-analysis of clinical labs in two independent biobanks. PLOS Genet. 16(11):e1009077 [PubMed: 33175840]
79. Diogo D, Tian C, Franklin CS, Alanne-Kinnunen M, March M, et al. 2018. Phenome-wide association studies across large population cohorts support drug target validation. Nat. Commun 9:4285 [PubMed: 30327483]
80. Boland MR, Alur-Gupta S, Levine L, Gabriel P, Gonzalez-Hernandez G. 2019. Disease associations depend on visit type: results from a visit-wide association study. BioData Min. 12:15 [PubMed: 31338127]
81. McCarty CA, Chisholm RL, Chute CG, Kullo IJ, Jarvik GP, et al. 2011. The eMERGE Network: A consortium of biorepositories linked to electronic medical records data for conducting genomic studies. BMC Med. Genom 4:13

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

82. McCarty CA, Wilke RA, Giampietro PF, Wesbrook SD, Caldwell MD. 2005. Marshfield Clinic Personalized Medicine Research Project (PMRP): design, methods and recruitment for a large population-based biobank. Pers. Med 2(1):49–79
83. Dewey FE, Murray MF, Overton JD, Habegger L, Leader JB, et al. 2016. Distribution and clinical impact of functional variants in 50,726 whole-exome sequences from the DiscovEHR study. Science 354(6319):aaf6814 [PubMed: 28008009]
84. All Us Res. Prog. Investig. 2019. The "All of Us" Research Program. N. Engl. J. Med 381(7):668–76 [PubMed: 31412182]
85. Zouk H, Venner E, Lennon NJ, Muzny DM, Abrams D, et al. 2019. Harmonizing clinical sequencing and interpretation for the eMERGE III Network. Am. J. Hum. Genet 105(3):588–605 [PubMed: 31447099]
86. Amberger J, Bocchini CA, Scott AF, Hamosh A. 2009. McKusick's Online Mendelian Inheritance in Man (OMIM®). Nucleic Acids Res. 37(Suppl. 1):D793–96 [PubMed: 18842627]
87. Hamosh A, Scott AF, Amberger JS, Bocchini CA, McKusick VA. 2005. Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders. Nucleic Acids Res. 33(Suppl. 1):D514–17 [PubMed: 15608251]
88. Antonarakis SE, McKusick VA. 2000. OMIM passes the 1,000-disease-gene mark. Nat. Genet 25:11 [PubMed: 10802643]
89. Amberger JS, Bocchini CA, Scott AF, Hamosh A. 2019. OMIM.org: leveraging knowledge across phenotype–gene relationships. Nucleic Acids Res. 47(D1):D1038–43 [PubMed: 30445645]
90. Randal J 1962. For basic look at heredity. Newark Evening News, Aug. 15
91. Rehm HL. 2017. The MedSeq and BabySeq studies: integrating genomics into the practice of medicine. Pathology 49:S32
92. Motulsky AG. 2006. Genetics of complex diseases. J. Zhejiang Univ. Sci. B 7(2):167–68 [PubMed: 16421979]
93. Ikegawa S 2012. A short history of the genome-wide association study: where we were and where we are going. Genom. Inform 10(4):220–25
94. Gallagher MD, Chen-Plotkin AS. 2018. The post-GWAS era: from association to function. Am. J. Hum. Genet 102(5):717–30 [PubMed: 29727686]
95. Frayling T 2014. Genome-wide association studies: the good, the bad and the ugly. Clin. Med 14(4):428–31
96. Katsanis N 2016. The continuum of causality in human genetic disorders. Genome Biol. 17:233 [PubMed: 27855690]
97. Blair DR, Lyttle CS, Mortensen JM, Bearden CF, Jensen AB, et al. 2013. A nondegenerate code of deleterious variants in Mendelian loci contributes to complex disease risk. Cell 155(1):70–80 [PubMed: 24074861]
98. Robinson PN, Mundlos S. 2010. The Human Phenotype Ontology. Clin. Genet 77(6):525–34 [PubMed: 20412080]
99. Groza T, Köhler S, Moldenhauer D, Vasilevsky N, Baynam G, et al. 2015. The Human Phenotype Ontology: semantic unification of common and rare disease. Am. J. Hum. Genet 97(1):111–24 [PubMed: 26119816]
100. Köhler S, Vasilevsky NA, Engelstad M, Foster E, McMurry J, et al. 2017. The Human Phenotype Ontology in 2017. Nucleic Acids Res. 45(D1):D865–76 [PubMed: 27899602]
101. Bastarache L, Hughey JJ, Hebbring S, Marlo J, Zhao W, et al. 2018. Phenotype risk scores identify patients with unrecognized Mendelian disease patterns. Science 359(6381):1233–39 [PubMed: 29590070]
102. Wallis C 1997. Diagnosing cystic fibrosis: blood, sweat, and tears. Arch. Dis. Child 76(2):85–88 [PubMed: 9068292]
103. Schacherer J 2016. Beyond the simplicity of Mendelian inheritance. C. R. Biol 339(7):284–88 [PubMed: 27344551]
104. Wells BJ, Chagin KM, Nowacki AS, Kattan MW. 2013. Strategies for handling missing data in electronic health record derived data. eGEMs 1(3):7

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

105. Haneuse S, Bogart A, Jazic I, Westbrook EO, Boudreau D, et al. 2016. Learning about missing data mechanisms in electronic health records-based research: a survey-based approach. Epidemiology 27(1):82–90 [PubMed: 26484425]

106. Ye Zi, Kullo Iftikhar J 2018. A phenotype risk score for monogenic aortopathy is associated with dilatation of the thoracic aorta and risk of adverse aortic events. Circulation 138(Suppl. 1):A12500 (Abstr.)

107. Zhong X, Yin Z, Jia G, Zhou D, Wei Q, et al. 2020. Electronic health record phenotypes associated with genetically regulated expression of *CFTR* and application to cystic fibrosis. Genet. Med 22(7):1191–200 [PubMed: 32296164]

108. Salvatore M, Beesley LJ, Fritsche LG, Hanauer D, Shi X, et al. 2020. Phenotype risk scores (PheRS) for pancreatic cancer using time-stamped electronic health record data: discovery and validation in two large biobanks. J. Biomed. Inform 113:103652 [PubMed: 33279681]

109. Lebovitch D, Johnson J, Duenas H, Stahl E, Charney A, Huckins L. 2019. Construction of a phenotype risk score for MDD. Eur. Neuropsychopharmacol 29:S138

110. Bastarache L, Hughey JJ, Goldstein JA, Bastraache JA, Das S, et al. 2019. Improving the phenotype risk score as a scalable approach to identifying patients with Mendelian disease. J. Am. Med. Inform. Assoc 26(12):1437–47 [PubMed: 31609419]

111. Yu S, Ma Y, Gronsbell J, Cai T, Ananthakrishnan AN, et al. 2018. Enabling phenotypic big data with PheNorm. J. Am. Med. Inform. Assoc 25(1):54–60 [PubMed: 29126253]

112. Liao KP, Sun J, Cai TA, Link N, Hong C, et al. 2019. High-throughput multimodal automated phenotyping (MAP) with application to PheWAS. J. Am. Med. Inform. Assoc 26(11):1255–62 [PubMed: 31613361]

113. Zhang Y, Cai T, Yu S, Cho K, Hong C, et al. 2019. High-throughput phenotyping with electronic medical record data using a common semi-supervised approach (PheCAP). Nat. Protoc 14(12):3426–44 [PubMed: 31748751]

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

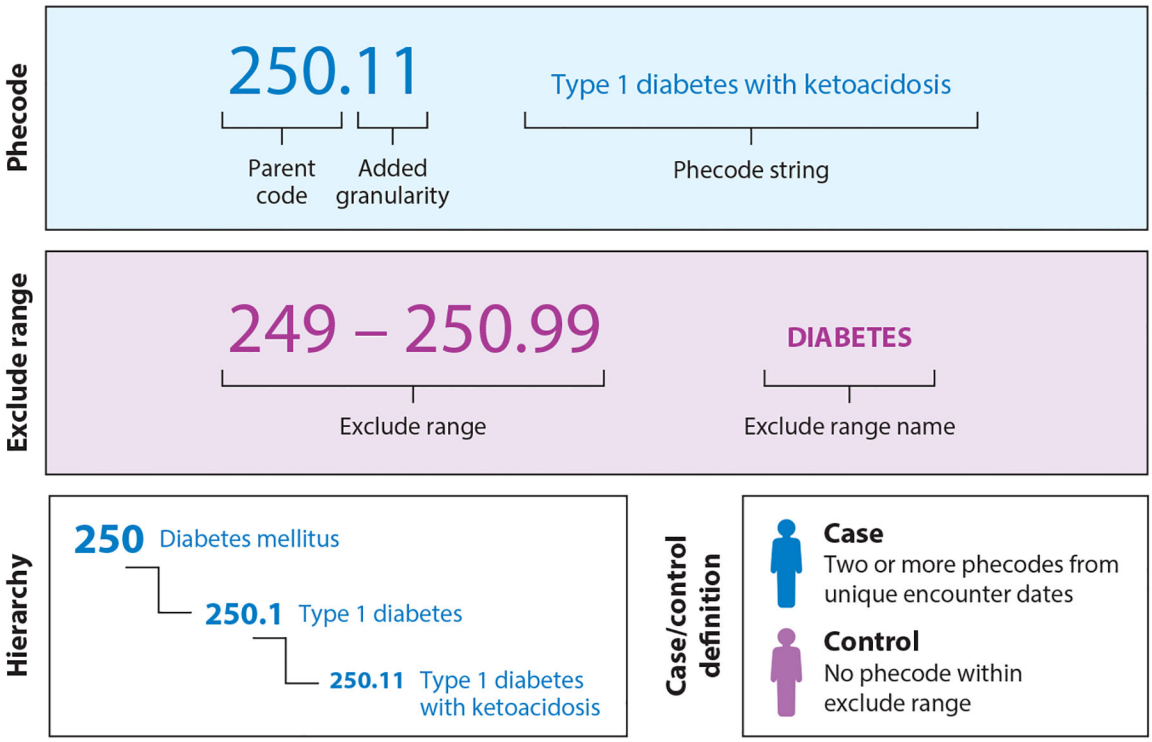
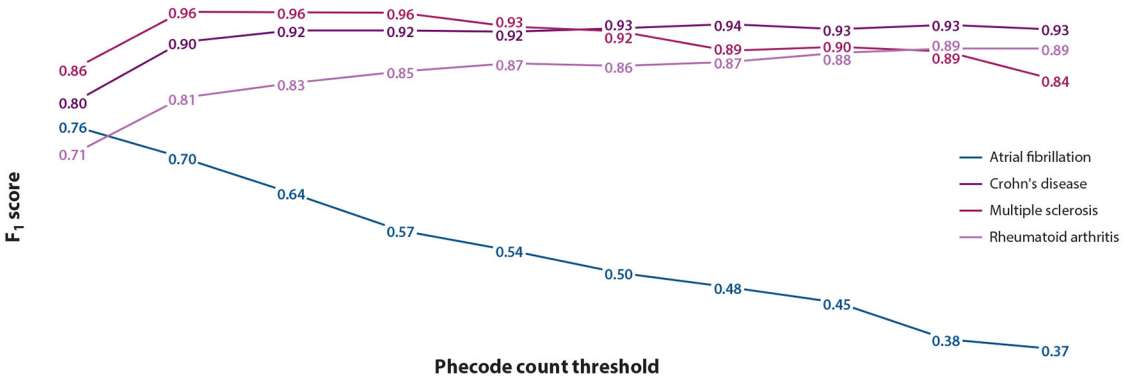


Figure 1.

The anatomy of a phecode. A phecode is a three-digit parent code with optional digits following a decimal point. Numbers after the decimal point indicate a hierarchical relationship. Phecodes without subordinate codes are called leaf codes. Each phecode has a string label and is linked to an exclude range. Cases are often defined as individuals with two or more unique phecodes, and controls are defined as individuals who do not have any code within the exclude range.



	1+	2+	3+	4+	5+	6+	7+	8+	9+	10+
Mean F_1	0.779	0.842	0.838	0.827	0.815	0.803	0.792	0.789	0.774	0.756

Figure 2.

Phecode performance based on the minimum code count required for a patient to count as a case. Requiring two or more phecodes on unique dates to define a case resulted in the highest mean F_1 score across four phenotypes.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

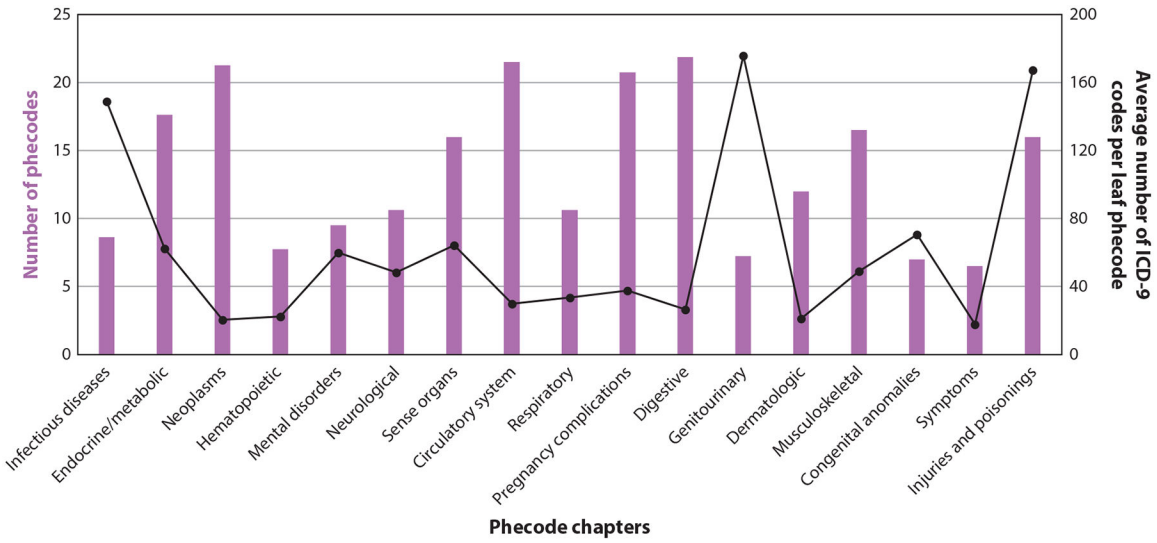


Figure 3.

Phecode statistics by chapter. Across phecode chapters, there is variability in the total number of phecodes (*purple bars*; *left axis*) as well as the average number of ICD-9 (International Classification of Diseases, Ninth Revision) codes that define each leaf phecode (*black points*; *right axis*).

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

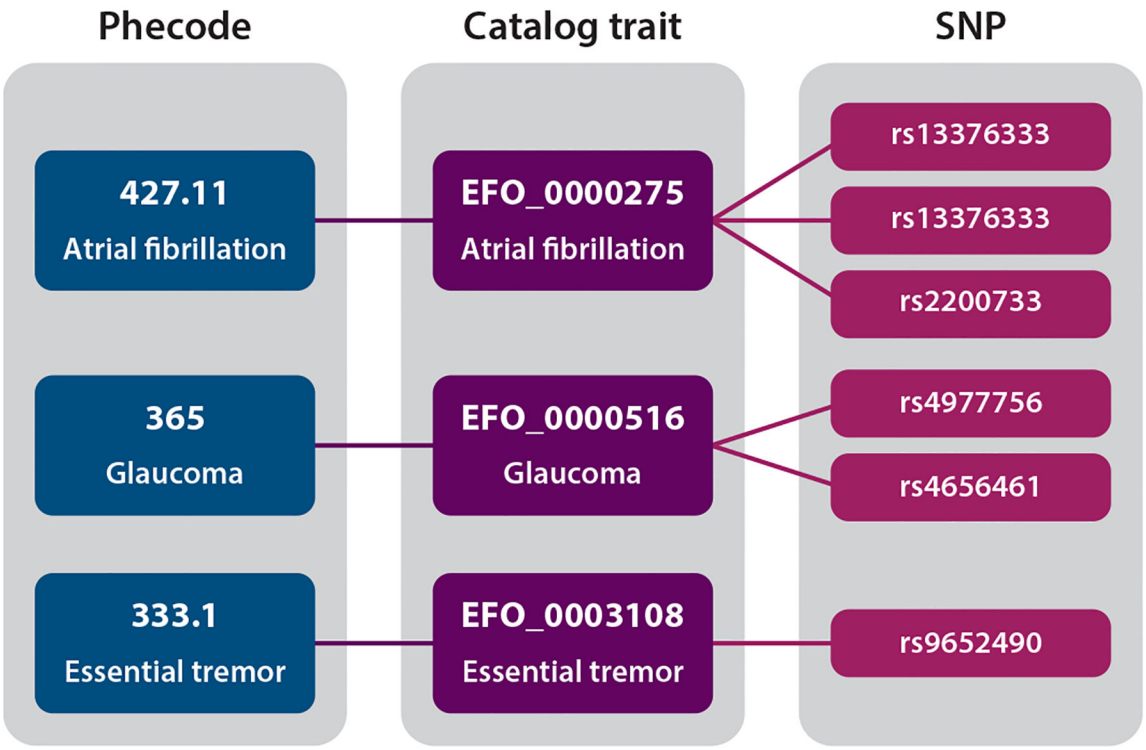


Figure 4.

Linking phecodes to single-nucleotide polymorphisms (SNPs). The GWAS (genome-wide association study) Catalog reports SNP–trait associations found in previous studies. Catalog traits are annotated with the Experimental Factory Ontology (EFO). Phecodes are linked to SNPs through a phecode/EFO map; three examples are shown here.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

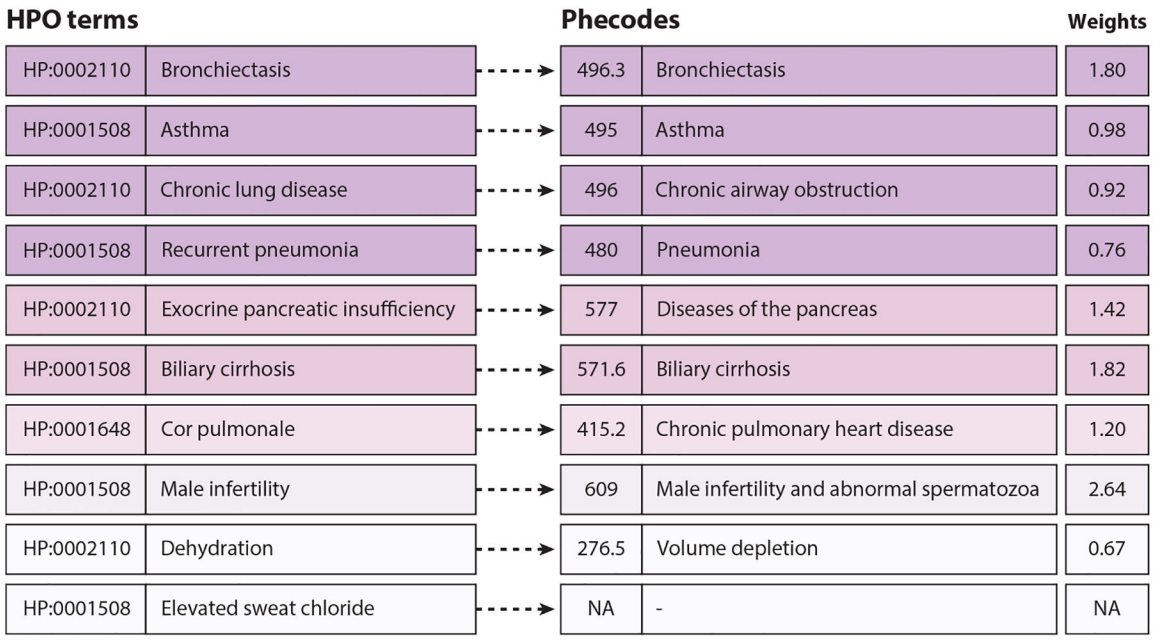


Figure 5.

Mapping from The Human Phenotype Ontology (HPO) to phecodes. Select features from Online Mendelian Inheritance in Man (OMIM) clinical description of cystic fibrosis are shown as HPO terms (*left*), along with their mapping to phecodes (*right*).

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

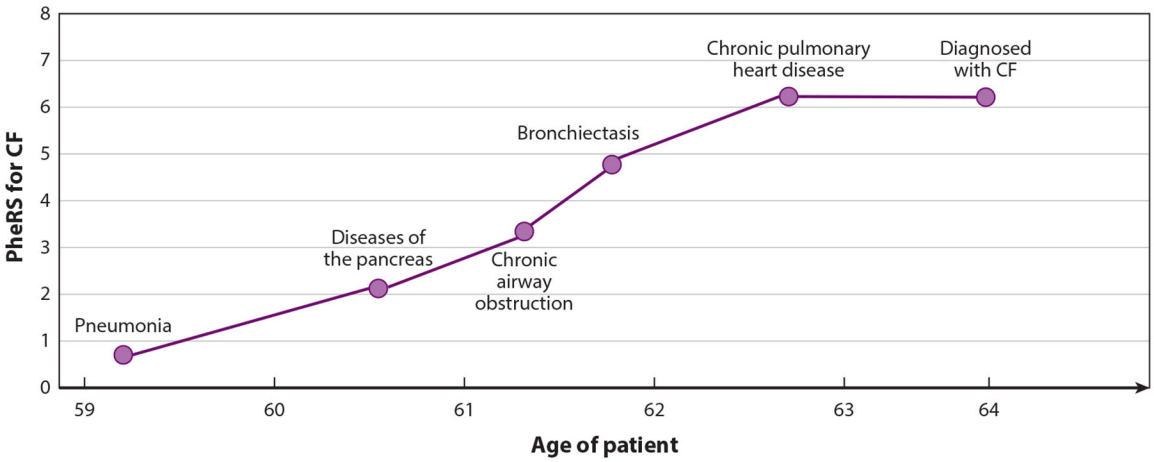


Figure 6.

Phenotype risk score (PheRS) for cystic fibrosis (CF) of a patient diagnosed late in life. The PheRS for CF score rises over time as the patient acquires more diagnoses that overlap with the disease profile. By the time this patient was diagnosed with CF, their PheRS was higher than that of 99% of patients.

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

Table 1

Summary counts of the current phecode/SNP map, based on the GWAS Catalog

	Phecode/SNP pairs	Unique SNPs	Unique phecodes	GWAS studies
European	8,600	6,302	141	575
Asian	1,238	1,052	66	198
African	226	224	30	46
All[a]	11,462	8,121	163	902

[a]Includes studies with multiple or unspecified ancestries.

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

Table 2

Contrasting attributes of Mendelian versus complex diseases

	Mendelian diseases	Complex diseases
Definition	Heritable conditions that run in families and are caused by mutations in a single gene	Diseases that arise from the interaction of multiple genetic and environmental factors
Field of study	Genetics	Genomics
Genetic architecture	Monogenic (single gene)	Polygenic (many genes)
Subject of study	Individuals/families	Populations
Genetic variant type	Rare, often exonic and nonsynonymous	Common, often noncoding or intergenic
Genotype/phenotype relationship	Causal; high or full penetrance	Statistical, often small effect size
Method	Family-based studies	GWAS
Genotyping data	Sequence data	SNP arrays
Resource catalog	OMIM (Online Mendelian Inheritance in Man)	GWAS Catalog