Genetics in Medicine (2023) 25, 100921

ELSEVIER

Genetics in Medicine
An Official Journal of the ACMG

www.journals.elsevier.com/genetics-in-medicine

ARTICLE

Phenotypic presentation of Mendelian disease across the diagnostic trajectory in electronic health records

Rory J. Tinker[1], Josh Peterson[2,3], Lisa Bastarache[3,*]

[1]Division of Medical Genetics and Genomic Medicine, Vanderbilt University Medical Center, Nashville, TN; [2]Vanderbilt University Medical Center, Department of Medicine, Nashville, TN; [3]Vanderbilt University Medical Center, Department of Biomedical Informatics, Nashville, TN

ARTICLE INFO

Article history:
Received 14 February 2023
Received in revised form
12 June 2023
Accepted 13 June 2023
Available online 17 June 2023

Keywords:
Diagnostic Convergence
Diagnostic delay
EHR
Mendelian genetic disorders
Rare disease

ABSTRACT

Purpose: To investigate the phenotypic presentation of Mendelian disease across the diagnostic trajectory in the electronic health record (EHR).
Methods: We applied a conceptual model to delineate the diagnostic trajectory of Mendelian disease to the EHRs of patients affected by 1 of 9 Mendelian diseases. We assessed data availability and phenotype ascertainment across the diagnostic trajectory using phenotype risk scores and validated our findings via chart review of patients with hereditary connective tissue disorders.
Results: We identified 896 individuals with genetically confirmed diagnoses, 216 (24%) of whom had fully ascertained diagnostic trajectories. Phenotype risk scores increased following clinical suspicion and diagnosis ($P < 1 \times 10^{-4}$, Wilcoxon rank sum test). We found that of all International Classification of Disease–based phenotypes in the EHR, 66% were recorded after clinical suspicion, and manual chart review yielded consistent results.
Conclusion: Using a novel conceptual model to study the diagnostic trajectory of genetic disease in the EHR, we demonstrated that phenotype ascertainment is, in large part, driven by the clinical examinations and studies prompted by clinical suspicion of a genetic disease, a process we term diagnostic convergence. Algorithms designed to detect undiagnosed genetic disease should consider censoring EHR data at the first date of clinical suspicion to avoid data leakage.

Introduction

The study of Mendelian disorders has yielded a great body of knowledge about the phenotypic manifestations of Mendelian genetic disease (hereafter referred to as genetic disease).[1-4] Knowledge of phenotype/genotype correlation is essential for the recognition and diagnosis of genetic disease and is summarized in resources such as the Online Mendelian Inheritance in Man (OMIM) and Orphanet.[5,6] Despite the availability of sophisticated knowledgebases, recognizing genetic disease can still be a great challenge for clinicians, leading to prolonged diagnostic delay.[7-9] Diagnostic delay continues to be a problem for a variety of genetic diseases, which has not consistently improved with

The Article Publishing Charge (APC) for this article was paid by Lisa Bastarache.
*Correspondence and requests for materials should be addressed to Lisa Bastarache, 2525 West End Ave, Suite 1500, Nashville, TN 37203. *Email address:* Lisa.bastarache@vumc.org

doi: https://doi.org/10.1016/j.gim.2023.100921

time.[10-14] Current evidence suggests that neither increased availability of genetic testing nor targeted campaigns aimed at increasing awareness of rare diseases are sufficient to fully address diagnostic delay.[10-12] The persistence of diagnostic delay has led to an interest in diagnosis support systems to identify undiagnosed patients using data from the electronic health record (EHR).[15-19]

To effectively identify undiagnosed patients, EHR-based algorithms must "recognize" the clinical patterns of the genetic disease, just as clinicians do. But unlike clinicians, these algorithms cannot examine patients. Rather, they must make use of phenotypic clues that are available in real-world clinical data. As such, they must address 2 key challenges of EHR data.[20] First, EHRs are often missing data. A leading cause of incompleteness is information fragmentation, which occurs when patients receive care at multiple care sites.[21] EHR chart fragmentation may be particularly problematic for patients with rare genetic disease, who are more likely to seek care at disparate referral centers. For algorithms that seek to identify undiagnosed disease, fragmented EHRs may be inappropriate for training and testing. Second, EHRs are affected by ascertainment bias.[22] EHRs contain myriad observations, measurements, and diagnoses that simultaneously reflect the physiology of a patient, as well as the process of health care itself. True facts about a patient are often missing from a record until they become clinically relevant. Consequently, the way that genetic disease "presents" in the EHR may differ from clinical descriptions in resources such as OMIM, the latter of which are based on highly detailed physical examinations of patients known to be affected by a genetic condition.

In this paper, we develop a conceptual model to study the phenotypic manifestations of genetic disease from the perspective of the EHR. A recent scoping review noted that there is no systematic framework to test, train, and evaluate EHR-based diagnosis support systems in rare disease, hindering the ability to realistically assess and compare different algorithms whilst increasing the risk of data leakage.[23] Our conceptual model seeks to describe the diagnostic trajectory as reflected in the EHR. The notion of a diagnostic trajectory is analogous to a disease trajectory, in which the progression of a disease is tracked over time.[24] The model delineates longitudinal data into pre-suspicion, pre-diagnosis, and post-diagnosis intervals, enabling the assessment of data availability and phenotype ascertainment at different phases of the diagnostic trajectory.

We applied our framework to the EHRs of patients diagnosed with 1 of 9 genetic diseases. We (1) measure the degree of EHR fragmentation in our cohort, (2) assess EHR data availability within each diagnostic time interval defined by clinical suspicion or diagnosis, and (3) characterize the way ascertainment bias influences how phenotypes are documented in the EHR data at different points. We found that EHR fragmentation is a major barrier to ascertaining undiagnosed patients before clinical suspicion for disease. And we note that the ascertainment of phenotypes characteristic of Mendelian disease often occurs only after clinical suspicion and/or diagnosis.

Material and Methods

Defining the conceptual model and key terms

We defined a conceptual model (Figure 1) that describes key events of the diagnostic process as reflected in EHR data, including the following: first encounter within the health system, first clinical suspicion (the date when a disease is first mentioned in clinical notes), diagnosis (the date when a clinical diagnosis is established), and final encounter. These events delineate the following time intervals: pre-ascertainment, pre-suspicion, pre-diagnosis, suspicion to diagnosis, and post-diagnosis. Within these intervals, we characterize data availability by both quantity (number of encounters) and EHR temporal length (length of time in days). A patient has a fully ascertained diagnostic trajectory if they have encounters in the pre-suspicion, suspicion to diagnosis, and post-diagnosis intervals. A full glossary of the definition of terms can be found in Table 1.

Data source

Our cohort comprised all individuals with at least 3 encounters within Vanderbilt University Medical Center (VUMC) between January 1, 2002, and January 1, 2022. Data were drawn from the Research Derivative, a copy of VUMC's EHR stored in the Observational Medical Outcomes Partnership common data model that includes demographics, clinical notes, International Classification of Disease (ICD) and Current Procedural Terminology codes used in this project.[25] All EHR clinical data were available for extraction in the current project and was not filtered by if it was related to the diagnosis or not.

Applying the conceptual model to genetic diseases

We elected to study genetic diseases with prominent multisystem phenotypes (caused by diverse biological etiologies) that have reports of diagnostic delay in the literature and vary by age of onset (Supplemental Table S1).[8,9,12,26-29] These included 5 hereditary connective tissue diseases (HCTDs)—Marfan syndrome (MFS), Loeys-Dietz syndrome (LDS), Stickler syndrome (STL), classical Ehlers-Danlos syndrome type 1 and 2 (cEDS), and vascular Ehlers-Danlos syndrome (vEDS)—as well as hereditary hemorrhagic telangiectasia (HHT), hypophosphatasia (HPP), Noonan syndrome (NS), and cystic fibrosis (CF). We restricted our cohort to adults diagnosed with CF (age $\geq$ 18) to exclude patients diagnosed from newborn screening. We used OMIM to generate a list of causal genes in the phenotypic series for these disorders and indexed all clinical notes for gene names. Because of

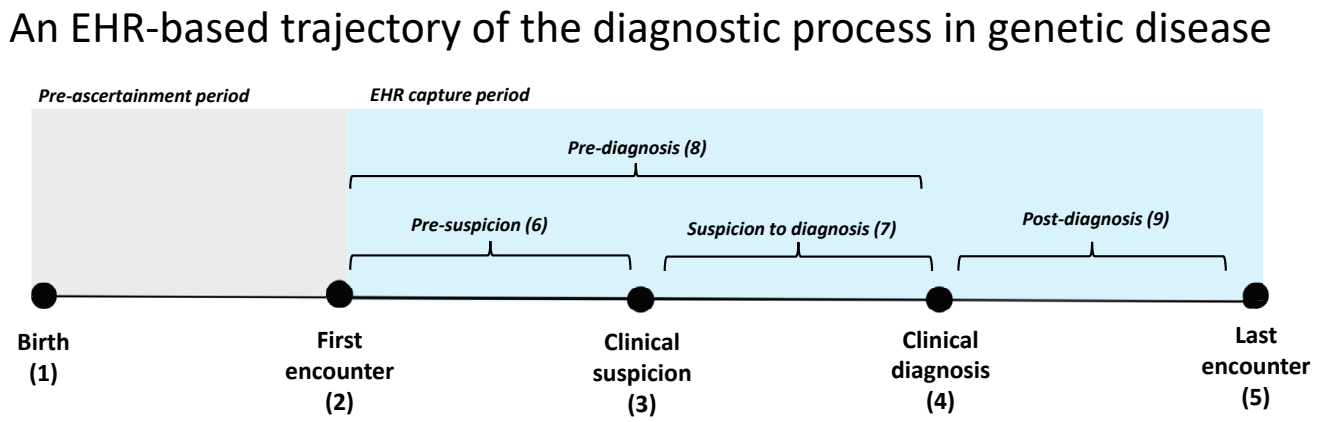

Figure 1 A graphical representation of the diagnostic trajectory as represented in EHR data. This model is intended to help assemble and code EHR data to both study the diagnostic process and test/train models to identify undiagnosed patients. It includes the following temporal events: (1) date of birth, (2) date at first encounter, (3) date of first clinical suspicion, (4) date of clinical diagnosis, and (5) date at last encounter. These time points define the following EHR temporal length intervals: (6) pre-suspicion interval (first encounter to 1 week before clinical suspicion), (7) suspicion to diagnosis interval (date of suspicion to 1 week before diagnosis), (8) pre-diagnosis interval (first encounter to 1 week before diagnosis) and (9) post-diagnosis interval (date of diagnosis to final encounter). Two periods were also defined (gray) pre-ascertainment period (birth to first encounter) and (blue) EHR capture period (date of clinical date of first encounter to date of last encounter). Key time points are indicated by black circles, key time intervals are denoted by brackets, and key periods are indicated in color. EHR, electronic health record.

clinical heterogeneity of the various forms for NS, we restricted our search to *PTPN11*, *SOS1*, and *RAF1*, 3 of the most common genetic causes of NS.[30] The chart of every individual with a mention of a gene name in their record was manually reviewed, and individuals with a pathogenic, diagnostic variant were flagged for further review. Because of the many matches for the *ENG* gene, we restricted our review to patients with a gene mentioned plus at least 1 ICD code for HHT (448.0 or I78.0). For adult CF cases, we identified potential cases using ICD codes requiring 2 or more ICD codes (277* or E84*) with the first occurrence at age 18 or older.[31] Previous studies have demonstrated that ICD codes have a high sensitivity for CF.[32]

To define the diagnostic trajectory of each genetic disease patient, we extracted key time points from the EHR (Figure 1). Date of birth was derived from the demographics table, and the first and last encounters were defined using the first and last ICD billing code dates. Initial suspicion was defined as the first mention of an individual genetic disease in clinical notes, either by a specific disease name (eg, MS) or disease class (eg, connective tissue disease). Supplemental Table 2 indicates keywords used to search across the clinical notes during manual review. The date of diagnosis was defined as the first date when a treating physician states that the patient is clinically diagnosed with the condition, also defined via manual chart review. We calculated patient age at each of the key time points and computed EHR temporal length (length of time in days) and quantity within each interval. Encounters were defined using ICD billing dates.

Phenotype risk score analysis

The phenotype risk score (PheRS) is a measurement of the similarity of a patient's EHR data and the clinical presentation of a disease as described in OMIM. PheRS represents characteristic phenotypes of genetic disease with phecodes: ICD based high-throughput phenotypes. PheRS has been used to assess the pathogenicity of genetic variants, as well as to detect undiagnosed individuals using EHR data.[31,33] Here, we tested the ability of PheRS to distinguish between cases diagnosed with Mendelian disease and unaffected controls at multiple time points along the diagnostic trajectory.

As OMIM has multiple clinical synopses for similar genetic diseases, we merged the synopses into a single feature set for cEDS, NS, LDS, STL, and HHT. We applied PheRS for each of the 9 genetic diseases to our entire cohort of 1.8 million individuals. A linear regression was used to normalize the PheRS using sex (assigned within the EHR by the clinician at birth), age, and record length as covariates, producing a "residualized PheRS" (rPheRS). This method used the EHR records ($N = 1.8$ million) of individuals in our cohort as controls to ascertain a baseline rPheRS for unaffected individuals. We compared rPheRS in cases with controls using the Wilcox rank sum test. We also counted the number of cases with "highly elevated" scores (rPheRS $>$ 4 standard deviations above the median).

We repeated the above analysis after censoring genetic disease data at different time points of the diagnostic trajectory, using ICDs from the pre-suspicion interval, the pre-diagnosis interval, and the post-diagnosis interval. For each analysis, a new cohort was generated in which the data from the individual genetic disease patients were censored to match the target interval, using covariates that reflected the censored data. Finally, we used a Wilcoxon signed-rank test to compare rPheRS of the same individuals at different time intervals, comparing scores from pre-suspicion, pre-diagnosis, and post-diagnosis time intervals.

Table 1 A glossary of key terms used in the current study

Term	Definition	Note
Date of birth	Date of patient's birth as listed in the demographics table	—
Date at first encounter	The first encounter date recorded in an EHR	Encounter dates can be defined in several ways (eg, billing codes or clinic notes). For a predictive algorithm, encounter should be defined as dates where new phenotypic information is ascertained (ie, an ascertainment event). For algorithms that use claims data to define features, the ascertainment event may be defined as billing days. For those that use clinical notes, a clinical encounter may be used.
Date of first clinical suspicion	The first date a clinician documents a suspicion of a disease in clinical notes, whether specific (eg, "Marfan syndrome") or more general (eg, "connective tissue disease")	—
Date of diagnosis	The first date when a treating physician states that the patient is clinically diagnosed with the condition. This declaration is often preceded by statements that the diagnosis is "probable" or "likely." Such conditional statements do not qualify as a diagnosis	—
Date at last encounter	Last encounter date in the EHR	—
Pre-ascertainment period	Birth to first encounter in calendar days	—
EHR capture period	Date of clinical date of first encounter to date of last encounter	
Pre-suspicion interval	First encounter to 1 week before clinical suspicion in calendar days	—
Suspicion to diagnosis interval	Date of suspicion to 1 week before diagnosis	
Pre-diagnosis interval	First encounter to 1 week before diagnosis in calendar days	—
Post-diagnosis interval	Date of diagnosis to final encounter in calendar days	—
Data quantity	The number of unique encounter dates	The pre-suspicion data quantity indicates the amount of information available on which to make a prediction
EHR temporal length	Length of time interval in days between 2 events within the EHR. This could be a billing event or alternatively a clinical event (eg, suspicion or diagnosis).	The longitudinally of the pre-suspicion interval defines a theoretical maximum of how much earlier a diagnosis could have been made.

EHR, electronic health record.

Individual phenotype analysis

We explored the phenotypes driving the changes in rPheRS across the diagnostic trajectory by analyzing when individual phenotypes appear in the EHR. For each genetic disease clinical feature, we counted the number of individuals who had a phecode dated prior to suspicion, between suspicion and diagnosis, following diagnosis, and ever. We also counted the first occurrence of each phenotype.

Targeted ascertainment analysis

We hypothesized that, from the perspective of the EHR, clinical suspicion of a specific Mendelian disorder prompts the clinician to look for and ascertain signs and symptoms of the disease. To test this hypothesis, we analyzed the timing of procedure codes for diagnostic examinations related to HCTDs. We extracted Current Procedural Terminology codes and dates for the following procedures: transthoracic echocardiography (93303, 93304, 93306, 93307, and 93308), ophthalmological examination and evaluation (92002, 92004, 92012, 92014, 92015, 92018, and 92019), and complete blood count (CBC) (85027 and 85025); the latter served as a control test, unrelated to HCTD. Among patients with HCTDs, we counted the number of individuals within each code group, as well as the total number of procedure codes prior to suspicion.

ICD billing code chart review

We assessed the possibility that clinical notes contained relevant information that was not captured in ICD codes.

Table 2 EHR data availability in our genetic disease cohort

Disease Name	Gene(s)	Abbreviation	Total	Diagnosed Before First Visit *N* (%)	Suspicion On First Encounter *N* (%)	Fully Ascertained Trajectory *N* (%)
Marfan syndrome	*FBN1*	MFS	145	55 (37.9)	57 (39.3)	33 (22.8)
Ehlers Danlos, Classic	*COL5A1/2*	cEDS	9	2 (22.2)	1 (11.1)	6 (66.6)
Ehlers Danlos, Vascular	*COL3A1*	vEDS	27	5 (20.8)	9 (33.3)	13 (48.1)
Loeys-Dietz syndrome	*TGFBR1/2, TGFB2, SMAD2/3*	LDS	32	7 (21.9)	8 (25.0)	17 (53.1)
Stickler syndrome	*COL2A1, COL11A1, COL9A1, COL9A3*	STL	40	8 (20.0)	14 (35.0)	18 (45.0)
Hereditary Hemorrhagic Telangiectasia	*ACVRL1, ENG*	HHT	79	28 (35.4)	19 (24.1)	32 (40.5)
Hypophosphatasia	*ALPL*	HPP	93	53 (57.0)	10 (10.8)	30 (32.3)
Noonan syndrome	*PTPN11, SOS1, RAF1*	NS	92	17 (18.5)	26 (28.3)	49 (53.3)
Cystic Fibrosis	*CFTR*	CF	379	353 (93.1)	7 (1.85)	18 (4.75)
All	—	All	896	528 (60.8)	151 (16.9)	216 (24.1)

We conducted a chart review of patients with HCTDs (MFS, LDS, STL, cEDS, and vEDS). This review was conducted by a clinical geneticist who was blinded to all data used in the prior analyses except for the target condition. Briefly, we reviewed the consensus guidelines, OMIM, and Gene Reviews and selected the 3 cardinal features most reported in these databases. We chose to do this to note the most specific common features that a nongeneticist would likely note outside of OMIM. We noted the date that these phenotypic features were first noted in the chart and determined if this was before or after clinical suspicion.

Statistical tools

All statistical analyses were conducted in R. We used the PheRS R package to generate scores.[34] For the residual scores (rPheRS), we included covariates for sex, age at first visit (days), age at last visit (days), and total number of ascertainment events (unique dates with ICD codes). PhecodeX was used to translate ICDs into phecodes.[35]

Results

We indexed the records of 1.8 million patients for 21 genes associated with our 9 individual genetic diseases. A total of 4711 records included a mention of 1 of our target genes, 896 (19%) of which indicated a positive genetic testing that confirmed a diagnosis for a target condition (Supplemental Table S3).

Description of diagnostic delay in cohort

Among our cohort of patients diagnosed with genetic disease, 528 (59%) were diagnosed before their first visit to VUMC (Table 2). For another 151 individuals (17%), the date of clinical suspicion occurred within a week of their first visit. In total, 216 individuals (24%) were found to have fully ascertained diagnostic trajectories, including 33 MFS, 6 cEDS, 13 vEDS, 17 LDS, 18 STL, 32 HHT, 30 HPP, 49 NS, and 18 adult CF cases. Among individuals with fully ascertained trajectories, the age of diagnosis in years varied widely, with a median of 12 (interquartile range [IQR]: 3.5-36); 84 patients (38.8%) were diagnosed as adults, which is suggestive of a long diagnostic delay (Table 3, Figure 2A).

The impact of data fragmentation on the ascertainment of a complete diagnostic trajectory

An algorithm designed to shorten diagnostic delay must rely on data from before the date of diagnosis or, better yet, clinical suspicion. Thus, the measures of EHR temporal length and quantity indicate how much information is available for predictive algorithms. The median time from first visit to clinical suspicion was 3.23 years (IQR: 153 days-6.4 years), indicating a theoretical opportunity to shorten the time to diagnosis. Within that interval, a median of 6 visits (IQR: 2-20) occurred (Figure 2B). EHR temporal length and quantity were significantly correlated ($R = 0.43$; $P \leq .001$), though there were many instances with high data quantity and low EHR temporal length (Supplemental Figure 1). The median time from clinical suspicion to diagnosis was 71 days (IQR: 20.8-235), indicating that, for most patients, diagnosis occurred soon after a genetic disease was suspected.

PheRS signal increases over the diagnostic trajectory

The PheRS feature definition included 91 phenotypes for the 9 diseases tested (Supplemental Table S4). The median

Table 3 The key temporal points, data availability, and quantity of our genetic disease cohort individuals with fully ascertained trajectories

			HCTD					Non-HCTD				
	Disease Name	MFS	cEDS	vEDS	LDS	STL	HHT	HPP	NS	CF	ALL
All (Complete Trajectory)		33	6	13	17	18	32	30	49	18	216
Ages (years)	First visit	3.6 [0.2-12.5]	3.8 [2.3-16.2]	23.1 [3.5-29.9]	2.6 [0.1-25.3]	0.4 [0.1-7.9]	8.7 [2-35.1]	32.6 [3-51.1]	0.1 [0-1.3]	49.9 [26.6-58.3]	3.7 [0.1-29.8]
	Suspicion	8 [2.8-16.7]	11.5 [7-21.7]	23.6 [16-41.5]	7.8 [2.6-35.9]	6.3 [0.9-15.7]	12.8 [6.7-41.8]	35.1 [8-61.8]	1 [0.4-4.2]	52.3 [29-62.6]	9.5 [1.9-34.1]
	Diagnosis	12.1 [4.5-18.8]	11.5 [7.3-22.2]	23.6 [16.4-41.6]	14.4 [7-35.9]	7.1 [0.9-15.8]	14.8 [7.4-43.4]	35.9 [8.1-62.2]	1.8 [0.7-7.2]	54.2 [29.8-63.9]	12 [3.5-35.6]
	Last visit	16.9 [9.3-29.5]	17.6 [14-26.2]	33.2 [22.1-45.7]	19.7 [8.4-38.5]	11.8 [9.6-19.2]	21.3 [11.5-44.2]	39.6 [15.8-66.9]	9.9 [5.9-14.2]	55.8 [37.3-65.8]	18.2 [9.2-41.9]
EHR temporal length (days)	First visit to suspicion	1301 [347-2300]	1653 [660-3761]	1096 [155-4390]	1492 [394-2462]	550.5 [142.5-2393.5]	1288.5 [204-2571.5]	1488.5 [212.5-3842]	219 [96-751]	899 [56.5-1818]	845.5 [152.5-2332.5]
	Suspicion to diagnosis	84 [28-364]	88.5 [2-220.8]	29 [21-118]	42 [0-1659]	36.5 [17-114.2]	72 [21.8-323]	52.5 [18.5-157.8]	78 [35-151]	100 [49.5-720.5]	71 [20.8-235]
	Post-diagnosis	1293 [531-3584]	1675.5 [1179-2178.8]	2659 [781-4926]	1833 [602-1995]	1172.5 [760-3351.2]	799 [72.5-2087.5]	1199 [305-1842.2]	1825 [450-3227]	1170 [901-2001.2]	1267 [557.5-2726.8]
Data quantity (visits)	First visit to suspicion	4 [2-7]	4 [3-9.5]	4 [1-42]	4 [3-14]	10 [2-19.5]	5.5 [2-10.5]	7.5 [3.2-43]	13 [3-21]	3 [1-12.8]	6 [2-20]
	Suspicion to diagnosis	2 [1-6]	0.5 [0-3.2]	1 [1-3]	2 [0-9]	2 [1-3]	2 [1-7.2]	2 [1-8.8]	4 [1-10]	4 [1-23]	2 [1-8]
	Post-diagnosis	23 [8-41]	9 [6.2-23.8]	16 [15-52]	20 [9-35]	24.5 [13.5-35.5]	11.5 [2-20.5]	21 [5.5-58.8]	28 [11-62]	78 [26.8-116.5]	22.5 [9-50.2]
Adult diagnosis (age ≥18)		10	2	9	6	3	14	20	2	18	84

CF, Cystic Fibrosis; *cEDS*, classical Ehlers-Danlos syndrome; *EHR*, electronic health record; *HCTD*, hereditary connective tissue diseases; *HHT*, Hemorrhagic Telangiectasia; *HPP*, Hypophosphatasia; *LDS*, Loeys-Dietz syndrome; *MFS*, Marfan syndrome; *NS*, Noonan syndrome; *STL*, Stickler syndrome; *vEDS*, vascular Ehlers-Danlos syndromes.

rPheRS for each disease was lowest in the pre-suspicion period, followed by the pre-diagnosis period, and the post-diagnosis period (Figure 3). We found that cases affected by genetic disease had higher rPheRS compared with controls when using all available data and when using data with each time interval tested (pre-suspicion, pre-diagnosis, and post-diagnosis). The difference between cases and controls was statistically significant ($P < .05$) for each disease, except for vEDS, which was not significant for any time interval, and HPP, which was not significant in the pre-suspicion interval. The difference in location between cases and controls for rPheRS was 0.56 (95% CI, 0.45-0.71) in the pre-suspicion interval, 1.59 (95% CI, 1.32-1.85) in the pre-diagnosis interval, and 2.43 (95% CI, 2.16-2.70) in the post diagnosis period (Supplemental Table S5). In the paired analysis, we found that the PheRS increased significantly across each time interval ($P < .05$), including from the pre-suspicion and post-diagnosis intervals (1.56 [95% CI, 1.15-1.97]; $P = 5.2 \times 10^{-21}$) (Supplemental Table S6). Additionally, the number of individuals with highly elevated scores increased across the diagnostic trajectory, from 22 (10%) in the pre-suspicion interval, to 48 (22%) in the pre-diagnosis interval, to 64 (29%) in the post-diagnosis interval, and 90 (42%) when all available data were used (Supplemental Table S7).

Phenotypic features of genetic disease are more likely to be present in the EHR after clinical suspicion

Only a minority of genetic disease-related phenotypes were first ascertained before clinical suspicion (Supplemental Tables S8-16) (Figure 4). This pattern was true for most individual phenotypes. For example, for MFS patients only 3 features were most likely to be ascertained before suspicion (tall stature; congenital deformities of skull, face, and jaw; and hammer toe), whereas 11 were more common after clinical suspicion, including aortic aneurysm and ectasia, mitral valve disorders, pectus excavatum, and lens dislocation. Similar patterns were found for other genetic diseases (Supplemental Tables S8-16). We call this phenomenon diagnostic convergence: wherein the key features of a Mendelian disease are ascertained in the EHR only after the disease is suspected.

Investigating ascertainment procedures and phenotyping error in HCTD diagnosis

Diagnostic convergence is consistent with good clinical practice; a clinician who suspects a genetic disease should conduct targeted additional examinations and tests that might increase or decrease confidence in the diagnosis. In this process, phenotypes that may have been present for a long time might first be observed and noted in the EHR. Indeed, we found that orders for transthoracic echocardiogram and ophthalmology examinations—2 procedures

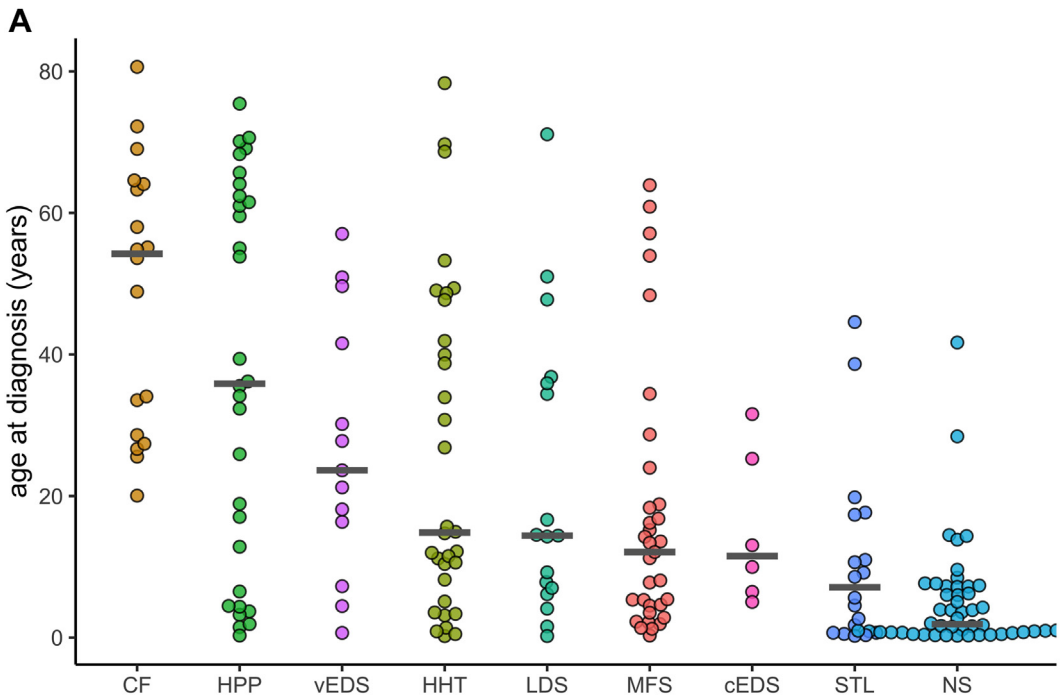

Figure 2 **A. Age of diagnosis for all individuals in the genetic disease cohort.** Each dot represents an individual patient, color coded by genetic disease. Line in figure represents median. B. EHR temporal length at different time intervals. Each patient is represented as a bar. Time before diagnosis is plotted in the negative direction and post-diagnosis in the positive direction. Time intervals of the diagnostic trajectory are represented in different colors. CF, Cystic Fibrosis; cEDS, classical Ehlers-Danlos syndrome; HHT, Hemorrhagic Telangiectasia; HPP, Hypophosphatasia; LDS, Loeys-Dietz syndrome; MFS, Marfan syndrome; NS, Noonan syndrome; STL, Stickler syndrome; vEDS, vascular Ehlers-Danlos syndromes.

used to ascertain key features HCTD—were more likely to be ordered after clinical suspicion of an HCTD.[36,37] Among the 87 HCTD patients, 67 (89%) and 47 (54%) received a transthoracic echocardiogram and ophthalmology examination, respectively. However, only a minority of transthoracic echocardiograms 28 of 308 (18%) and ophthalmology examinations 46 of 278 (17%) occurred before suspicion, in contrast with CBCs (one of the most common clinical investigations and not specific the HCTD), which were just as likely to be ordered before suspicion as they were after (Supplemental Table S17).

However, because our analysis relied on ICD billing codes, it is possible that genetic disease related phenotypes were noted in the clinical text but not billed for before clinical suspicion or diagnosis. To ensure that phenotyping error did not occur in ICD codes capturing the phenotypic features from the individuals EHR a further chart review of the cardinal features of HCTD's was conducted. Of the individuals with HCTD who had a fully ascertained diagnostic trajectory, there were minimal phenotypic features in the patients notes outside of ICD codes on chart review before clinical suspicion (N = 7/87, 8%) with the majority of features present post suspicion (N = 56/87, 64%) (Supplemental Table S18).

Discussion

The clinical presentation of genetic disease has been studied extensively, but less attention has been paid to the way these diseases are reflected within EHR data. Understanding the phenotypic expression in the EHR is key to developing algorithms to detect undiagnosed patients. In this study, we develop a conceptual model to quantify data availability and the phenotypic manifestations of genetic disease from the perspective of the EHR. We focused on 9 Mendelian diseases, chosen because of their multisystem phenotypes, varied reported age of onset, and evidence in the literature of significant diagnostic delay. We demonstrated the relevancy of EHR fragmentation—a problem that may be particularly acute in tertiary care centers like VUMC—and found that the majority of patients with genetically confirmed diagnosis were already diagnosed before their first visit to VUMC.

Our results demonstrated the relevance of ascertainment bias; analyzing the phenotypic signal throughout the ascertainment process, we found that many of the key phenotypes indicative of a genetic disease were ascertained only after a clinician suspected the disease. Thus, the EHR phenotypes of patients come to resemble the classical picture of the disease during the diagnostic trajectory in a phenomenon we named diagnostic convergence. Although diagnostic convergence is

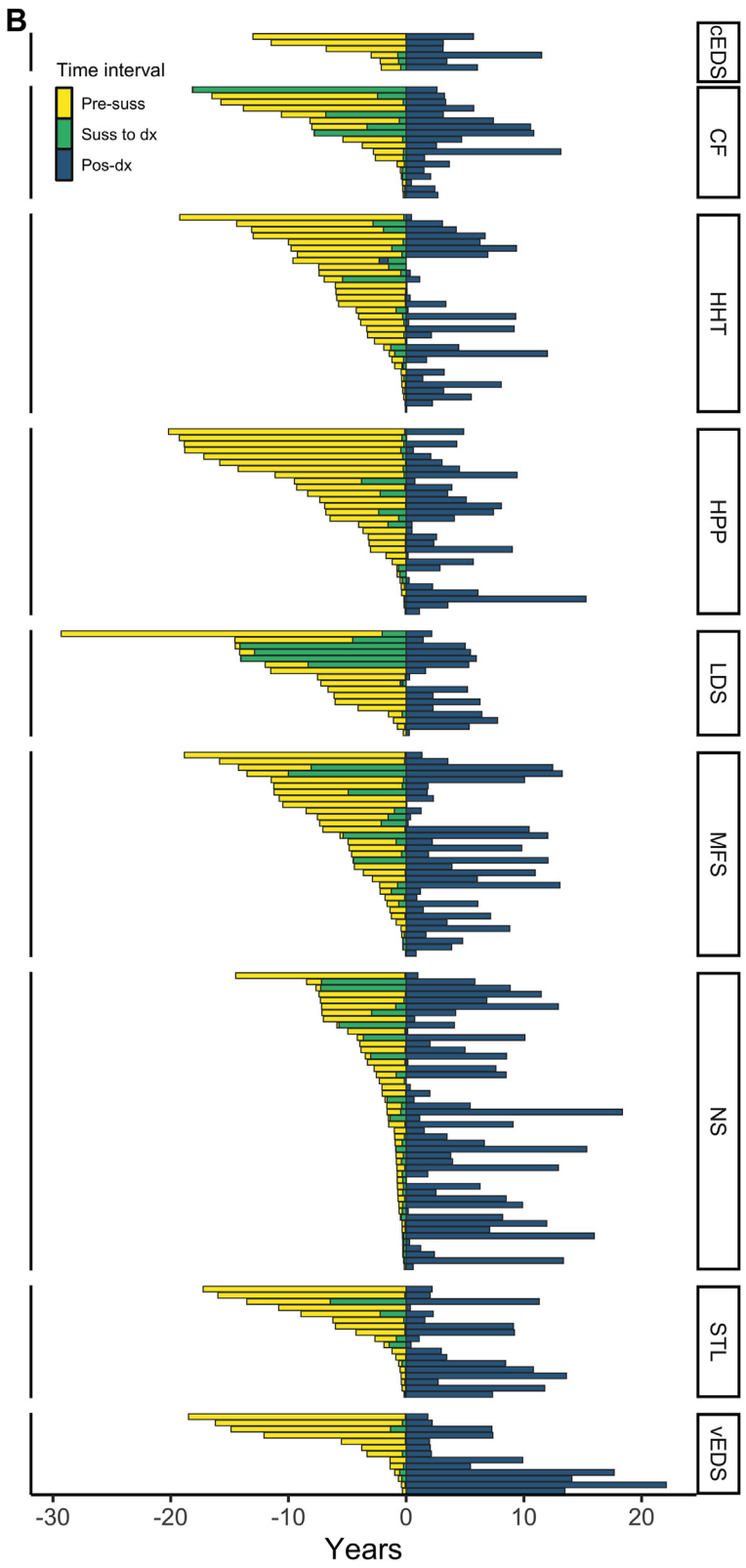

Figure 2 continued

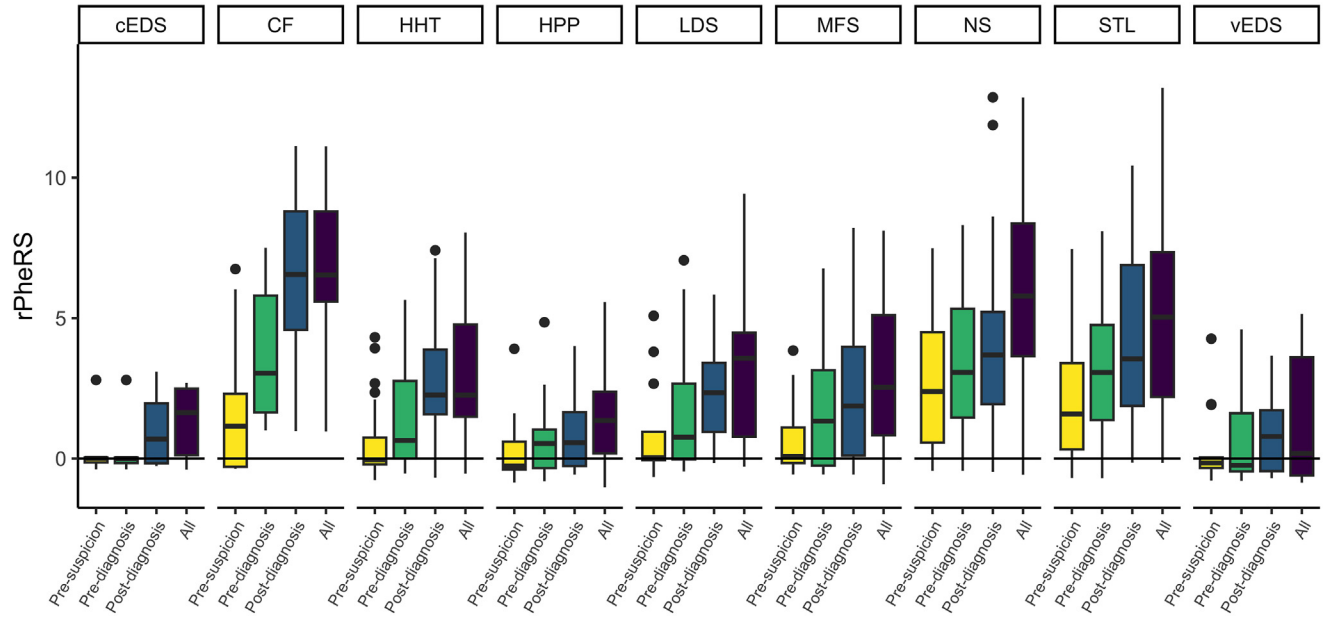

Figure 3 Changes in the PheRS throughout the diagnostic process. Dots represent individual patients with outlying scores that were included in our analysis. CF, Cystic Fibrosis; cEDS, classical Ehlers-Danlos syndrome; HHT, Hemorrhagic Telangiectasia; HPP, Hypophosphatasia; LDS, Loeys-Dietz syndrome; MFS, Marfan syndrome; NS, Noonan syndrome; STL, Stickler syndrome; vEDS, vascular Ehlers-Danlos syndromes.

consistent with good clinical care, it poses a challenge for designing EHR-based algorithms to detect undiagnosed disease. Using all available EHR data, PheRS effectively distinguished individuals with genetic disease versus unaffected controls. However, most genetic disease features were ascertained after clinical suspicion and diagnosis. The PheRS increased after every event in the diagnostic trajectory resulting in a gradual convergence on the classical phenotype known in the literature. Even phenotypic attributes that were likely to have been present since birth (eg, congenital heart defects and musculoskeletal deformities) or represent long term chronic consequences of individual genetic disease (eg, arteriovenous malformation in HHT or bronchiectasis in CF) were often mentioned only after clinical suspicion.

A deeper look at the diagnostic process revealed evidence that diagnostic convergence in our cohort was largely driven by the diagnostic process itself. Clinical suspicion prompted the clinician to look for additional signs of the disease, and diagnosis leading to heightened surveillance for particular phenotypes. Both transthoracic echocardiograms and ophthalmology examinations were far more likely to occur after clinical suspicion than before, whereas CBC, a test unrelated to the specific genetic disease, were just as likely to be ordered before suspicion as after. In an independent manual chart review of HCTD patients, we observed the same phenomenon, indicating that diagnostic convergence was not driven by billing practices alone.

Our findings have multiple implications for those who seek to develop algorithms that detect undiagnosed genetic disease. We demonstrate the challenge presented by EHR fragmentation, a well-known phenomenon that may be particularly problematic for the study of rare and difficult to diagnose conditions. We also demonstrate the importance of censoring data before suspicion to avoid data leakage wherein information only available after the prediction time point is used in the predictive model. Our results suggest that the concern of leakage is not merely theoretical. The diagnostic process itself elicits key phenotypes so that using data after suspicion may bias algorithm performance. Moreover, this process starts once a disease is suspected, indicating that censoring data on the diagnosis date may not be sufficient to prevent leakage. Our conceptual model may be useful in preparing data sets to train or test algorithms that seek to identify undiagnosed patients. As more researchers are working on computational solutions to address the problem of diagnostic delay, our model may help researchers organize their data for testing.

Our study has some limitations. First, our decision to only include individuals with a genetic and clinical confirmation of genetic disease means our cohort is rigorous and of high quality despite its size. However, this does have some implications for our analysis. Furthermore, some individuals in our control population would have reached a clinical diagnosis of MS (based on the application of the Ghent criteria) without a positive genetic test. However, given the size of our control cohort (1.8 million individuals used in our regression model), this likely would have had minimal impact on our results. Second, our cohort of genetic diseases is small and based on a single tertiary medical center, which is based in a major metropolitan area with primary and secondary care inpatient and outpatient facilities. This prevents us being able to conclude if diagnostic convergence is a global phenomenon or one restricted to a particular medical context. Third, we based our analysis on a subset of genetic diseases. Therefore,

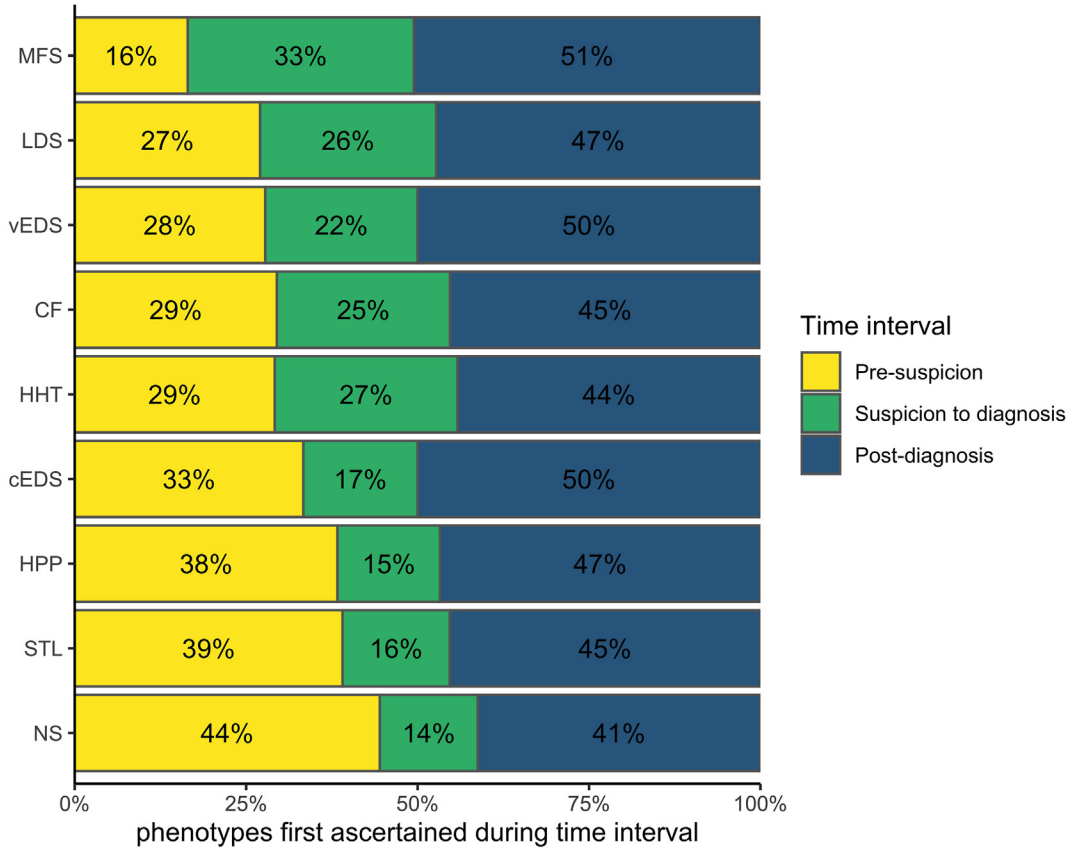

Figure 4 A figure showing the % of phenotypic features ascertained during the specific time intervals within the diagnostic trajectory. CF, Cystic Fibrosis; cEDS, classical Ehlers-Danlos syndrome; HHT, Hemorrhagic Telangiectasia; HPP, Hypophosphatasia; LDS, Loeys-Dietz syndrome; MFS, Marfan syndrome; NS, Noonan syndrome; STL, Stickler syndrome; vEDS, vascular Ehlers-Danlos syndromes.

future work is needed to establish whether diagnostic convergence is a phenomenon seen in other genetic or non-genetic diseases. Finally, our analysis was limited to only what was recorded in the EHR. As a result, we cannot be certain of the exact time a clinician became suspicious for a genetic disease or when a phenotype was noticed. Clinicians may not document all of their clinical assessments in the EHR for various practical reasons (eg, not wanting to stigmatize patients or affect their future insurability) or because a particular finding does not seem medically relevant. However, this constraint will likely be shared by researchers who wish to use EHRs to identify undiagnosed patients.

Conclusion

Here, we present a conceptual model to evaluate the diagnostic trajectory of Mendelian genetic diseases in EHR data. Using this model, we found that EHR fragmentation is a challenge for observing the diagnostic trajectory of individuals affected by genetic disease. We also observed that characteristic Mendelian disease phenotypes were more likely to be ascertained after clinical suspicion and diagnosis in a process we term diagnostic convergence. Our conceptual model may help in the design of algorithmic solutions that seek to shorten diagnostic delay.

Data Availability

Available upon request.

Acknowledgments

We thank the Potocsnak family foundation for their generous support in funding rare disease research at VUMC.

Funding

This work was supported by the National Library of Medicine (R01LM010685) and the National Human Genome Research Institute (R01HG012657).

Author Information

Conceptualization: R.J.T., L.B.; Data curation: L.B.; Formal analysis: R.J.T., L.B.; Writing-original draft: R.J.T., L.B., J.P.; Writing-review and editing: R.J.T., L.B., J.P.

Ethics Declaration

The Vanderbilt University Medical Center Approved the study (IRB # 221124). Informed consent was not required because the study was purely retrospective and did not require any identifiable data as per the IRB. The study adhered to the principles set out in the Declaration of Helsinki.

Conflict of Interest

Lisa Bastarache is a consultant for Galatea Bio. Josh Peterson is a consultant for Natera. Rory J. Tinker has provided ad hoc consulting to Gerson Lehrman Group.

Additional Information

The online version of this article (https://doi.org/10.1016/j.gim.2023.100921) contains supplemental material, which is available to authorized users.

References

1. Amberger JS, Bocchini CA, Schiettecatte F, Scott AF, Hamosh A. OMIM.org: Online Mendelian Inheritance in Man (OMIM®), an online catalog of human genes and genetic disorders. *Nucleic Acids Res*. 2015;43(Database issue):D789-D798. http://doi.org/10.1093/nar/gku1205
2. Ferreira CR. The burden of rare diseases. *Am J Med Genet A*. 2019;179(6):885-892. http://doi.org/10.1002/ajmg.a.61124
3. 100,000 Genomes Project Pilot Investigators, Smedley D, Smith KR, Martin A, et al. 100,000 genomes pilot on rare-disease diagnosis in health care – preliminary report. *N Engl J Med*. 2021;385(20):1868-1880. http://doi.org/10.1056/NEJMoa2035790
4. Tinker RJ, Guess T, Rinker DC, et al. A novel, likely pathogenic variant in UBTF-related neurodegeneration with brain atrophy is associated with a severe divergent neurodevelopmental phenotype. *Mol Genet Genomic Med*. 2022;10(12):e2054. http://doi.org/10.1002/mgg3.2054
5. Hamosh A, Scott AF, Amberger JS, Bocchini CA, McKusick VA. Online Mendelian Inheritance in Man (OMIM), a KnowledgeBase of human genes and genetic disorders. *Nucleic Acids Res*. 2005;33(Database issue):D514-D517. http://doi.org/10.1093/nar/gki033
6. Nguengang Wakap S, Lambert DM, Olry A, et al. Estimating cumulative point prevalence of rare diseases: analysis of the Orphanet database. *Eur J Hum Genet*. 2020;28(2):165-173. http://doi.org/10.1038/s41431-019-0508-0
7. Kuiper GA, Meijer OLM, Langereis EJ, Wijburg FA. Failure to shorten the diagnostic delay in two ultra-orphan diseases (mucopolysaccharidosis types I and III): potential causes and implications. *Orphanet J Rare Dis*. 2018;13(1):2. http://doi.org/10.1186/s13023-017-0733-y
8. Högler W, Langman C, Gomes da Silva H, et al. Diagnostic delay is common among patients with hypophosphatasia: initial findings from a longitudinal, prospective, global registry. *BMC Musculoskelet Disord*. 2019;20(1):80. http://doi.org/10.1186/s12891-019-2420-8
9. Pierucci P, Lenato GM, Suppressa P, et al. A long diagnostic delay in patients with hereditary haemorrhagic telangiectasia: a questionnaire-based retrospective study. *Orphanet J Rare Dis*. 2012;7:33. http://doi.org/10.1186/1750-1172-7-33
10. Indelicato E, Nachbauer W, Eigentler A, et al. Onset features and time to diagnosis in Friedreich's ataxia. *Orphanet J Rare Dis*. 2020;15(1):198. http://doi.org/10.1186/s13023-020-01475-9
11. El-Helou SM, Biegner AK, Bode S, et al. The German national registry of primary immunodeficiencies (2012-2017). *Front Immunol*. 2019;10:1272. http://doi.org/10.3389/fimmu.2019.01272
12. Groth KA, Hove H, Kyhl K, et al. Prevalence, incidence, and age at diagnosis in Marfan syndrome. *Orphanet J Rare Dis*. 2015;10:153. http://doi.org/10.1186/s13023-015-0369-8
13. Pera MC, Coratti G, Berti B, et al. Diagnostic journey in Spinal Muscular Atrophy: is it still an odyssey? *PLoS ONE*. 2020;15(3):e0230677. http://doi.org/10.1371/journal.pone.0230677
14. Löscher WN, Huemer M, Stulnig TM, et al. Pompe disease in Austria: clinical, genetic and epidemiological aspects. *J Neurol*. 2018;265(1):159-164. http://doi.org/10.1007/s00415-017-8686-6
15. Movaghar A, Page D, Brilliant M, Mailick M. Advancing artificial intelligence-assisted pre-screening for fragile X syndrome. *BMC Med Inform Decis Mak*. 2022;22(1):152. http://doi.org/10.1186/s12911-022-01896-5
16. Yang Z, Shikany A, Ni Y, Zhang G, Weaver KN, Chen J. Using deep learning and electronic health records to detect Noonan syndrome in pediatric patients. *Genet Med*. 2022;24(11):2329-2337.S1098-3600(22)00893-0. http://doi.org/10.1016/j.gim.2022.08.002
17. Kothari C, Srivastava S, Kousa Y, et al. Validation of a computational phenotype for finding patients eligible for genetic testing for pathogenic PTEN variants across three centers. *J Neurodev Disord*. 2022;14(1):24. http://doi.org/10.1186/s11689-022-09434-0
18. Morley TJ, Han L, Castro VM, et al. Phenotypic signatures in clinical data enable systematic identification of patients for genetic testing. *Nat Med*. 2021;27(6):1097-1104. http://doi.org/10.1038/s41591-021-01356-z
19. Bastarache L, Hughey JJ, Hebbring S, et al. Phenotype risk scores identify patients with unrecognized Mendelian disease patterns. *Science*. 2018;359(6381):1233-1239. http://doi.org/10.1126/science.aal4043
20. Hripcsak G, Albers DJ. Next-generation phenotyping of electronic health records. *J Am Med Inform Assoc*. 2013;20(1):117-121. http://doi.org/10.1136/amiajnl-2012-001145
21. Botsis T, Hartvigsen G, Chen F, Weng C. Secondary use of EHR: data quality issues and informatics opportunities. *Summit Transl Bioinform*. 2010;2010:1-5.
22. Bastarache L, Brown JS, Cimino JJ, et al. Developing real-world evidence from real-world data: transforming raw data into analytical datasets. *Learn Health Syst*. 2022;6(1):e10293. http://doi.org/10.1002/lrh2.10293
23. Faviez C, Chen X, Garcelon N, et al. Diagnosis support systems for rare diseases: a scoping review. *Orphanet J Rare Dis*. 2020;15(1):94. http://doi.org/10.1186/s13023-020-01374-z
24. Jensen AB, Moseley PL, Oprea TI, et al. Temporal disease trajectories condensed from population-wide registry data covering 6.2 million patients. *Nat Commun*. 2014;5:4022. http://doi.org/10.1038/ncomms5022
25. Makadia R, Ryan PB. Transforming the premier Perspective® hospital database into the Observational Medical Outcomes Partnership (OMOP) common data model. *eGEMs (Wash DC)*. 2014;2(1):1110. http://doi.org/10.13063/2327-9214.1110
26. Steinraths M, Vallance HD, Davidson AGF. Delays in diagnosing cystic fibrosis: can we find ways to diagnose it earlier? *Can Fam Phys Med Fam Can*. 2008;54(6):877-883.
27. Boothe M, Morris R, Robin N. Stickler syndrome: a review of clinical manifestations and the genetics evaluation. *J Pers Med*. 2020;10(3):105. http://doi.org/10.3390/jpm10030105
28. Malfait F, Wenstrup R, Paepe AD. *Classic Ehlers-Danlos Syndrome*. 2018. Accessed March 1, 2023. https://www.ncbi.nlm.nih.gov/books/NBK1244/
29. Zenker M, Edouard T, Blair JC, Cappa M. Noonan syndrome: improving recognition and diagnosis. *Arch Dis Child*. 2022;107(12):1073-1078. http://doi.org/10.1136/archdischild-2021-322858
30. Roberts AE. Noonan syndrome. In: Adam MP, Mirzaa GM, Pagon RA, et al., eds. *GeneReviews®*. 1993-2023. Accessed March 1, 2023. https://www.ncbi.nlm.nih.gov/books/NBK1124/

31. Bastarache L, Hughey JJ, Goldstein JA, et al. Improving the phenotype risk score as a scalable approach to identifying patients with Mendelian disease. *J Am Med Inform Assoc*. 2019;26(12):1437-1447. http://doi.org/10.1093/jamia/ocz179
32. Zhong X, Yin Z, Jia G, et al. EHR phenotypes associated with genetically regulated expression of CFTR and application to cystic fibrosis. *Genet Med Off J Am Coll Med Genet*. 2020;22:1191-1200. http://doi.org/10.1038/s41436-020-0786-5
33. Bastarache L. Using phecodes for research with the electronic health record: from PheWAS to PheRS. *Annu Rev Biomed Data Sci*. 2021;4:1-19. http://doi.org/10.1146/annurev-biodatasci-122320-112352
34. Aref L, Bastarache L, Hughey JJ. The phers R package: using phenotype risk scores based on electronic health records to study Mendelian disease and rare genetic variants. *Bioinformatics*. 2022;38(21):4972-4974. http://doi.org/10.1093/bioinformatics/btac619
35. Phe WAS. *Phenome Wide Association Studies*. 2023. Accessed March 1, 2023. https://phewascatalog.org/phecode_x
36. Dietz H. *FBN1-Related Marfan Syndrome*. 2022. Accessed March 1, 2023. https://www.ncbi.nlm.nih.gov/books/NBK1335/
37. Wheeler JB, Ikonomidis JS, Jones JA. Connective Tissue Disorders and cardiovascular Complications: the indomitable role of transforming growth factor-beta signaling. *Adv Exp Med Biol*. 2014;802:107-127. http://doi.org/10.1007/978-94-007-7893-1_8