BioData Mining

RESEARCH Open Access

Enriched phenotypes in rare variant carriers suggest pathogenic mechanisms in rare disease patients

Lane Fitzsimmons[1,2], Undiagnosed Diseases Network, Brett Beaulieu-Jones[1,3*] and Shilpa Nadimpalli Kobren[1*]

*Correspondence:
beaulieujones@uchicago.edu;
shilpa_kobren@hms.harvard.edu

[1] Department of Biomedical Informatics, Harvard Medical School, Boston, MA 02115, USA
[2] Renaissance School of Medicine, Stony Brook University, Stony Brook, NY 11794, USA
[3] Department of Medicine, University of Chicago, Chicago, IL 60615, USA

Abstract

Background: The mechanistic pathways that give rise to the extreme symptoms exhibited by rare disease patients are complex, heterogeneous, and difficult to discern. Understanding these mechanisms is critical for developing treatments that address the underlying causes of diseases rather than merely the presenting symptoms. Moreover, the same dysfunctional series of interrelated symptoms implicated in rare recessive diseases may also lead to milder and potentially preventable symptoms in carriers in the general population. Seizures are a common and extreme phenotype that can result from diverse and often elusive pathways in patients with ultrarare or undiagnosed disorders.

Methods: In this pilot study, we present an approach to understand the underlying pathways leading to seizures in patients from the Undiagnosed Diseases Network (UDN) by analyzing aggregated genotype and phenotype data from the UK Biobank (UKB). Specifically, we look for enriched phenotypes across UKB participants who harbor rare variants in the same gene known or suspected to be causally implicated in a UDN patient's recessively manifesting disorder. Analyzing these milder but related associated phenotypes in UKB participants can provide insight into the disease-causing mechanisms at play in rare disease UDN patients.

Results: We present six vignettes of undiagnosed patients experiencing seizures as part of their recessive genetic condition. For each patient, we analyze a gene of interest: *MPO*, *P2RX7*, *SQSTM1*, *COL27A1*, *PIGQ*, or *CACNA2D2*, and find relevant symptoms associated with UKB participants. We discuss the potential mechanisms by which the digestive, skeletal, circulatory, and immune system abnormalities found in the UKB patients may contribute to the severe presentations exhibited by UDN patients. We find that in our set of rare disease patients, seizures may result from diverse, multi-step pathways that involve multiple body systems.

Conclusions: Analyses of large-scale population cohorts such as the UKB can be a critical tool to further our understanding of rare diseases in general. Continued research in this area could lead to more precise diagnostics and personalized treatment strategies for patients with rare and undiagnosed conditions.

Keywords: Seizures, Compound heterozygous, Variant carriers, Recessive conditions, Rare diseases

Background

Collectively, as many as 446 million people, or nearly 6% of the worldwide population, are afflicted with rare diseases [1]. However, any specific rare disease may manifest in as few as a handful of patients across the world. Rare diseases are difficult to diagnose, and patients often endure years of arduous, inconclusive testing and misdiagnoses. Even after diagnosis, treatments indicated for these specific diseases are typically limited [2]. Successful repurposing of other drugs for off-label rare disease treatment is also limited because existing usage guidelines, efficacy, and adverse event measurements from historical clinical studies may not apply to rare disease cases [3]. Despite the critical need for diagnostic and therapeutic improvements for rare disease patients, funders have difficulty justifying resource-intensive research and development when only a small number of patients will benefit. Additionally, support from insurers is often absent or insufficient [2, 4].

Rare diseases are largely genetic in nature. Eighty-five percent are estimated to have genetic etiologies, with the majority classified as recessive conditions [5]. The manifestation of recessive conditions requires that both copies of relevant autosomal gene(s) or single copies of relevant X-linked gene(s) have been rendered dysfunctional. This occurs when a patient inherits the same deleterious variant from both parents, resulting in a homozygous recessive variant [6]. Sickle-cell anemia, for instance, manifests when both copies of a patient's HBB gene harbor the same A > T base pair mutation at a critical position [7]. Alternatively, a recessive condition can arise when a patient inherits two different deleterious variants from each parent, called compound heterozygous variants. Cystic fibrosis, for example, arises when both copies of a patient's CFTR gene are compromised due to potentially different variants inherited from each parent [8]. In both of these types of recessive cases, affected individuals have seemingly unaffected parents but a potentially significant family history of disease, such as a similarly affected grandparent or sibling [6].

In addition to known pathogenic variants, other variants in the same disease-causing genes can result in a spectrum of symptoms ranging from benign to pathogenic. In the CFTR gene, for example, nearly 1700 different mutations lead to varying manifestations of cystic fibrosis with differing prognoses and comorbidities [9]. Intriguingly, recent studies have demonstrated that seemingly healthy individuals in the general population who carry one normally functioning and one compromised copy of the CFTR gene may exhibit related yet less severe symptoms typical of cystic fibrosis as well [10]. Patients with rare genetic disorders often present on the extreme side of the phenotypic spectrum, with debilitating symptoms that severely restrict normal physical and cognitive functioning [11]. The less severe phenotypes exhibited by carriers of variants in the same genes can provide insight into the underlying mechanisms at play in recessive disorders [12, 13]. Increased understanding of this disease process could have the potential to both help rare disease patients and potentially lead to preventative treatments in larger "carrier" populations where less severe phenotypes build up over time. For example, if variants in a gene causatively linked to neurodegeneration in a rare disease patient were associated with increased inflammation in the general population, carriers might be at increased risk of developing neurodegenerative conditions as well and could be candidates for clinical studies of early interventions.

It can be extremely difficult to gather information and draw conclusions about rare disorders that individually manifest in such small numbers. For suspected genetic disorders where unrelated but similarly-presenting individuals are found, disease-associated genes or genic regions can be identified statistically through case–control analyses. However, for undiagnosed, "N-of-1" cases where no similarly affected individuals are known and no matching disorders have been described in the literature, clinical researchers must rely on external functional information to determine whether and how candidate genes harboring intriguing variants are concordant with a patient's symptoms [14, 15]. Even after a diagnostic gene is identified, the perturbed mechanisms leading to symptom onset and progression can remain elusive. In these cases, milder disease-related symptoms observed across individuals with only one damaging variant in the same gene may shed light on the underlying process implicated in a rare disease patient's extreme manifestation of symptoms of the same nature.

The Undiagnosed Diseases Network (UDN) was established in 2014 to serve patients with extremely rare and difficult-to-diagnose conditions. These patients represent "N-of-1" cases, and their disorders or specific phenotypic presentation manifest in so few patients worldwide that it is nearly impossible to first identify similarly presenting patients and then to search for recurrently-implicated, significantly associated genomic loci. The majority of patients who enroll in the UDN suffer from sporadic or suspected recessive conditions that manifest across multiple body systems. Examining these rare and complicated conditions in the context of healthy individuals who carry similar variants has the potential to benefit patients directly by identifying root causes of certain symptoms and to more broadly enhance understandings of recessive illnesses. Moreover, establishing connections between rare disorders and potentially actionable symptoms in the general population may further incentivize the development of improved diagnostics and therapeutics for these patients.

Here, we present a framework and pilot study to examine genes likely or known to be implicated in extremely rare disorders through the lens of cohort-level phenotype enrichment. We include in our pilot a subset of UDN patients with recessive disorders who experience seizures. Seizures do not have a single cause, but instead may be the eventual result of various biological and pathophysiological pathways. The specific pathways leading to seizure onset per patient are also often elusive; as many as half of global epilepsy cases have an unknown cause [16]. Though antiepileptic drugs can reduce the seizure burden in 60–70% of patients, these treatments minimize symptoms while rarely addressing the root biological cause [17]. The overall phenotypic heterogeneity of cases within the UDN and enrichment of patients with seizures within this cohort provides the opportunity to explore the differing potential mechanistic pathways that may predispose individuals to experience seizures.

For each UDN patient in our pilot with an identified strong candidate or diagnostic gene listed, we define "genotypically-similar" individuals as those who carry at least one rare variant in the same gene. We select cohorts of genotypically-similar patients for each UDN patient from the UK Biobank (UKB), a control dataset that contains both genetic and phenotypic patient information from 500,000 people between 40 and 69 years old from across the UK. From this cohort of adult patients with heterozygous variants, we expect milder symptoms that relate to and inform understandings of severe

pediatric homozygous cases. These identified related mechanisms aim to direct further exploration in cases where limited alternative avenues exist. We then examine the significant phenotype enrichments within each cohort to draw potential connections between the implicated body systems and mechanistic pathways underlying the UDN patients' conditions.

Methods

Phenotype and genotype information for UDN patients

Comprehensive and relevant phenotype and genotype data for all patients enrolled in the UDN are uploaded to the central data repository managed by the UDN Data Management and Coordinating Center and made available to all UDN researchers through the UDN Gateway. These data often include standardized phenotype terms from the Human Phenotype Ontology (HPO) and candidate genes and variants that had been or are currently under investigation by the clinical teams in charge of each case.

The genetic variants shared in the UDN Gateway could have been identified on a clinical sequencing report provided by the Clinical Laboratory Improvement Amendments (CLIA)-certified UDN Sequencing Core, through prior analyses or sequencing as documented in a patient's medical records, or through internal site-specific protocols and pipelines. Genetic variants are marked as "candidate", "solved" or "rejected."

Curating a set of recessive genes in seizure patients

We first filtered the 2366 patients enrolled in the UDN as of June 2022 to the subset of 622 patients annotated with at least one codified HPO seizure term. Our set of seizure terms was reviewed by a neurologist and consisted of 339 terms that are descendants of the seizure phenotype (HP:0001250) in the HPO ontology (Supplemental Table 1) [18]. We further filtered this list of patients to those with genes and variants listed in the UDN Gateway that the clinical team had classified as either candidate or confirmed causal. Some diagnoses in sporadic cases (i.e., two unaffected parents and no other affected family members) are due to single *de novo* variants that likely operate in a dominant manner. To consider only recessive disease etiologies, we further filtered our set of patients to those with a recessive combination of putatively or known causal variants in the same gene, including homozygous alternate variants or compound heterozygous variant pairs. Corresponding patient and gene counts for each filtering step are depicted in Fig. 1.

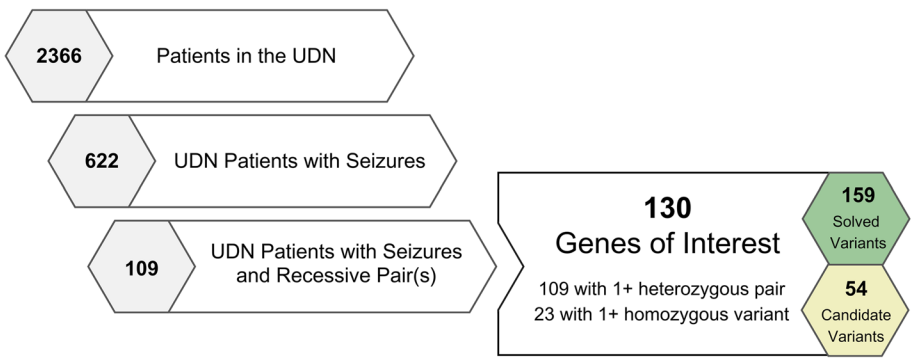

Fig. 1 Selection process to identify genes of interest

Genes that appeared in more than one category are included in both counts. Note that in some recessively manifesting cases, only one of the requisite two causal variants may have been identified and listed in the UDN Gateway; these cases were excluded in our analysis. The filtering steps resulted in a set of 130 unique genes with either homozygous or compound heterozygous variants that were implicated in at least one UDN patient experiencing seizures.

Identifying phenotypes enriched across genotypically-similar individuals

For each gene of interest in a UDN patient, we considered the corresponding subset of genotypically-similar individuals from the UKB who had whole-exome sequencing and at least one rare variant in the same gene. We then queried Genebass, a resource that aggregates phenotypic information from 281,850 individuals with whole-exome sequencing in the UKB, to extract all phenotypes significantly associated with the aggregated rare variants per gene [19]. Genebass phenotype association scores were calculated using SKAT-O, a method for associating rare variants and phenotypes. To increase the statistical power to detect significant associations for rare variants that are harbored by only a few individuals each, SKAT-O assigns weights to rare variants based on their observed frequencies within the population and effect sizes. The variants are then grouped within a gene and considered together via a weighted burden test to compute phenotypic associations [20]. Using SAIGE-GENE for multiple hypothesis correction, we considered gene-based phenotype associations below the adjusted 2.5e-10 threshold significant [21].

Assessing significance between enriched phenotypes and UDN patient presentation

We compared the enriched phenotypes found in genotypically-similar UKB patients to each starting UDN patient. We met with the clinicians and genetic counselors overseeing each patient's case to discuss potential connections between the enriched phenotypes and the mechanisms implicated in the UDN patients. The overall workflow is depicted in Fig. 2.

Results

We first sought to identify genes where high impact variants may recessively manifest in rare disorders and seizures. To this end, we considered the subset of patients enrolled in the UDN who experienced seizures and had one or more listed genes harboring a

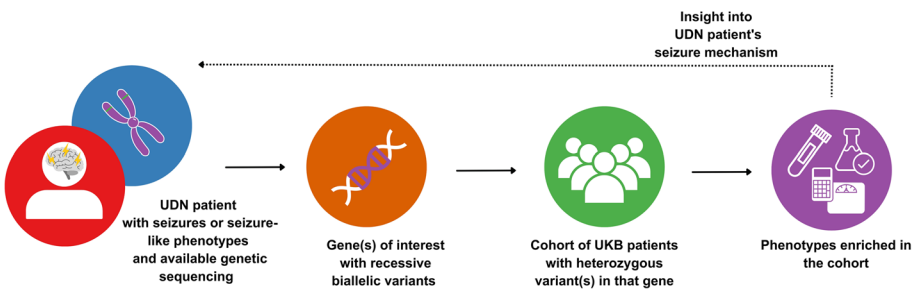

Fig. 2 Workflow for deriving insight into rare disease patient's seizure mechanism through cohort analysis

recessive variant pair, occurring either as compound heterozygous variants or a homozygous variant. This search resulted in a set of 238 unique variants falling into 130 genes across 109 patients. Variants are categorized by clinical teams into one of three causality statuses: candidate ($n = 159$), solved ($n = 54$), or rejected ($n = 25$). Most variants are initially categorized as "candidate" and then recategorized as "solved" if experimental studies confirm the variant's causal impact on an observed phenotype or additional patients with the same variant and compelling phenotypic overlap are identified through variant matchmaking services such as Matchmaker Exchange. Conversely, candidate variants are recategorized as "rejected" if functional studies demonstrate no phenotypic impact or additional individuals are found with the same variants but no overlapping symptoms. Computational tools can predict variant pathogenicity in the absence of experimental studies by considering protein structure, evolutionary conservation, and additional functionality measures (e.g., CADD, Polyphen, SIFT). As expected, we found that pathogenicity scores were significantly more deleterious for solved variants compared to candidate variants ($p < 0.001$ across all three pathogenicity predictors, KS-test) and compared to rejected variants ($p < 0.001$, KS-test, Fig. 3). This suggests that the solved variants in our rare disease cohort likely correspond to strongly deleterious molecular impacts, and even carriers of one copy of these or similar variants may exhibit some disease-relevant symptoms.

We next evaluated phenotypic enrichments across cohorts of individuals who are heterozygous for these specific variants or similar variants within each gene. For each of the 130 genes of interest in our UDN cohort, we considered individuals from the UKB with one or more variants in the same gene. We considered association values computed using SKAT-O per gene and provided through Genebass [19]. Eight of the 130 genes tested had one or more phenotypes enriched below our significance threshold (Table 1). Two of these genes were removed from subsequent analyses upon further case investigation and discussions with the corresponding clinical teams. Specifically, *DENND2C* was determined to be unlikely to be causal by the UDN clinical team and the patient with variants in *ATP7B* was determined to not experience seizures.

The phenotypes associated with the six remaining genes of interest—*MPO*, *PIGQ*, *P2RX7*, *COL27A1*, *CACNA2D2*, and *SQSTM1*—suggest plausible and potentially interesting mechanistic pathways that could be relevant for each UDN patient's condition. Intriguingly, these associated phenotypes primarily impacted the skeletal, circulatory, digestive, or immune systems rather than the nervous system (Fig. 4). This result is in line with previous suggestions that seizures as a symptom may be an eventual and extreme result of a wide range of complex and heterogenous underlying biological causes that are not necessarily primarily neurological in nature. Indeed, the phenotypes associated with these candidate genes were not immediate and obvious precursors to seizures, but instead suggested multi-step pathways which could lead to seizures in our rare disease cases. Below we present six vignettes discussing phenotypes associated with each candidate gene in genotypically-similar cohorts from the UKB and how these phenotypes are related to the initial rare disease patient case.

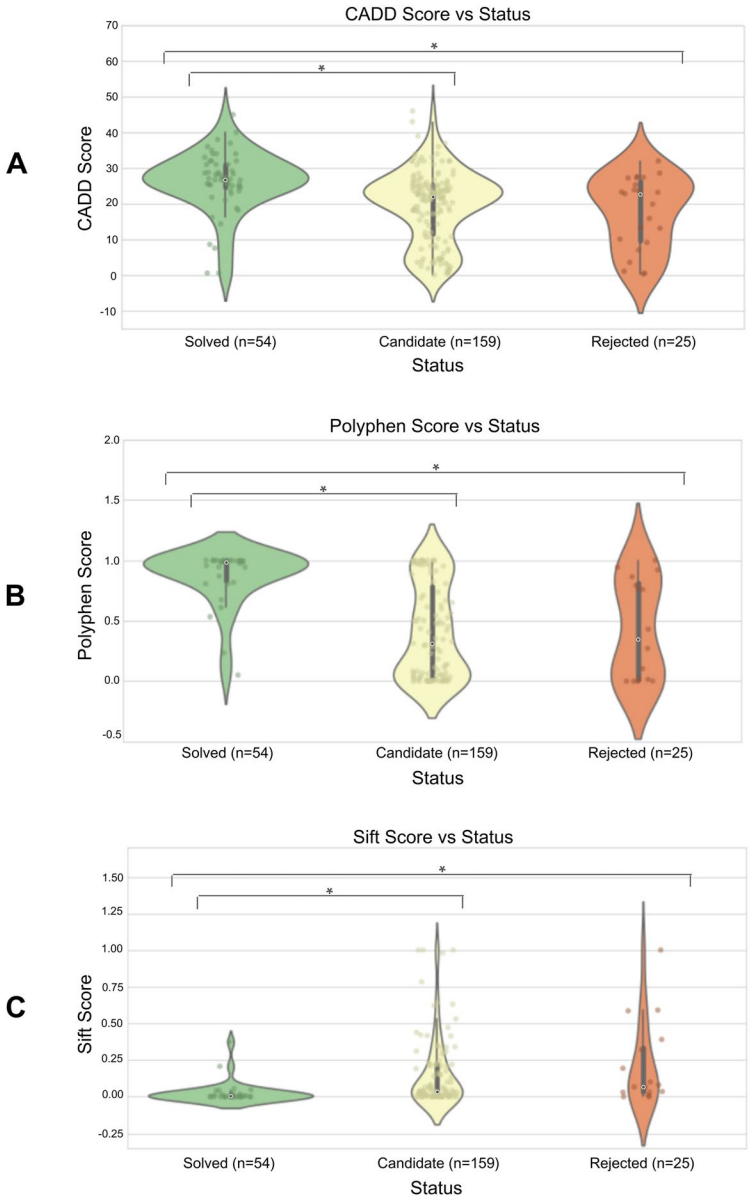

Fig. 3 Deleteriousness vs Causality Status. **A** CADD Score vs Status. **B** Polyphen Score vs Status. **C** Sift Score vs Status

MPO

The phenotypes significantly associated with UKB participants with rare variants in the MPO gene were abnormal white blood cell (leukocyte) count, monocyte count, neutrophil count, basophil count, lymphocyte percentage, neutrophil percentage, and basophil percentage: all classic indicators of infection. The UDN patient with compound heterozygous variants in the MPO gene is highly sensitive to recurrent infections—including yeast infections and recurrent vaginal infections. These frequent infections are often accompanied by intense painful spells that result in insomnia. The clinical team assigned to this patient believes that this patient's seizures are triggered by her insomnia.

Table 1 Enriched phenotypes for UK Biobank cohorts

Gene	UKB Enriched Phenotype(s)	UKB Data Field
ATP7B	Mean reticulocyte volume	30,260
	Mean sphered cell volume	30,272
	Sex hormone binding globulin	30,830
CACNA2D2	High light scatter reticulocyte percentage	30,290
	High light scatter reticulocyte count	30,300
COL27A1	Whole body fat free mass adjusted for body mass index	23,101
DENND2C	Platelet count	30,080
	Mean platelet (thrombocyte) volume	30,100
	Platelet distribution width	30,110
MPO	White blood cell (leukocyte) count	30,000
	Monocyte count	30,130
	Neutrophil count	30,140
	Basophil count	30,160
	Lymphocyte percentage	30,180
	Neutrophil percentage	30,200
	Basophil percentage	30,220
PIGQ	Trunk fat-free mass	23,129
	Trunk predicted mass	23,128
P2RX7	Mean platelet (thrombocyte) volume	30,100
	C-reactive protein	301,710
SQSTM1	Date M88 first reported (Paget's disease of bone)	131,978

Sleep deprivation is thought to promote seizure-causing interictal epileptiform discharges, or abnormal electrical activities that occur when a person is not actively experiencing a seizure [22]. It has also been shown that increasing the duration a person sleeps at night decreases the likelihood that they will experience a subsequent seizure [23]. As such, a connection between the patient's insomnia and seizures is reasonable and suggests the patient's seizures might be at least partially addressed by treating either the pain or insomnia. Although the pathogenic *MPO* variants are not necessarily the root cause of the patient's susceptibility to infection, her case is illustrative of how seizures can be a consequence of an indirect pathway involving non-nervous body systems. The insomnia and seizures represent only a portion of the patient's complex condition, but this connection suggests that individuals who are genetically predisposed to abnormal infection response due to rare variants in the MPO gene may benefit from stricter or preemptive management of pain and sleep loss.

P2RX7

An adolescent patient in the UDN suffers from bilateral tonic-clonic seizures and harbors a homozygous recessive variant in the P2RX7 gene. The phenotypes enriched among individuals with rare variants in *P2RX7* include abnormal C reactive protein level and mean platelet volume. C reactive protein regulates platelet activation as part of the body's inflammation response, and the rate of platelet production dictates mean platelet volume. Consequently, changes in these phenotypes are indicative of inflammation [24, 25]. Increased levels of C reactive protein have been found to be associated with epilepsy, suggesting a connection between inflammation and seizures [26–28]. It is, however, difficult to use retrospective data to determine whether the inflammation is contributing to the seizures or is a result of recurrent seizures.

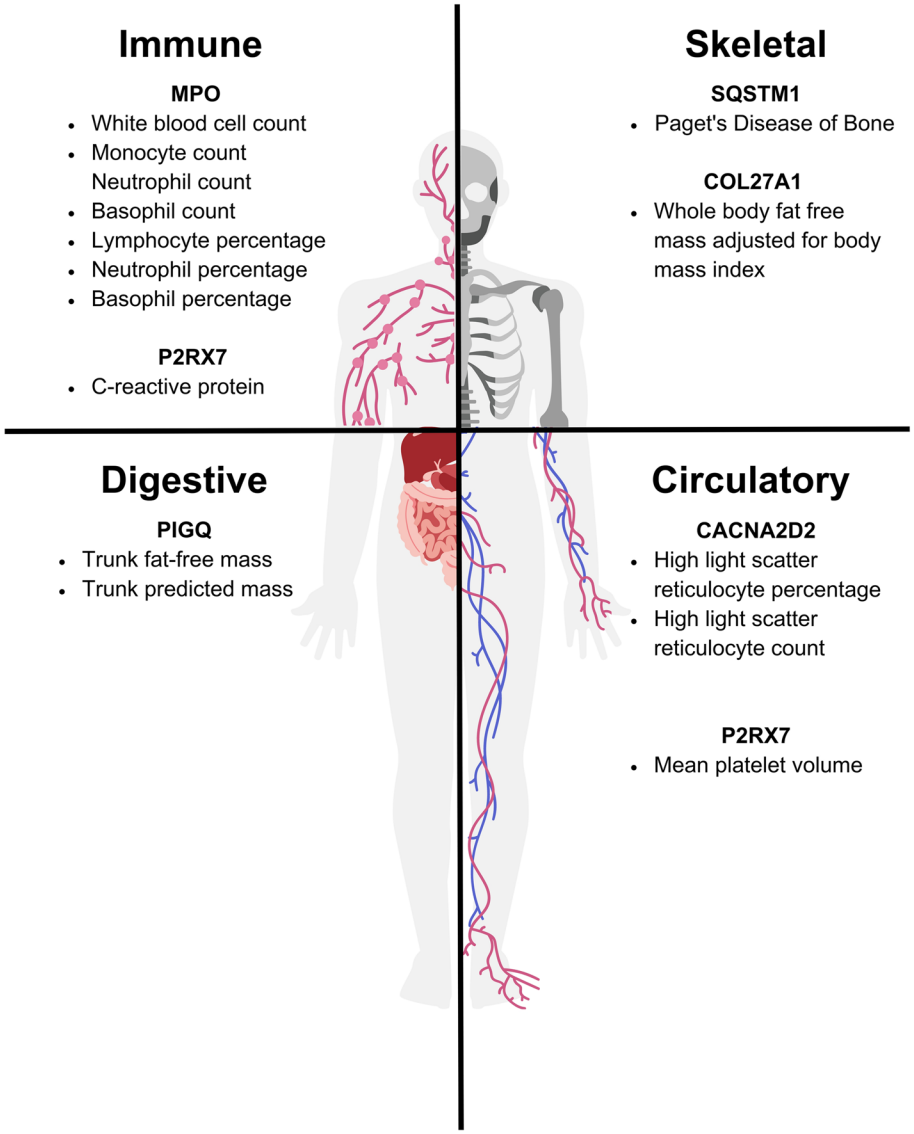

Fig. 4 Enriched phenotypes and associated body systems involved

Both C reactive protein and mean platelet volume have also been associated with increased risk of thrombosis, or blood clots that obstruct normal blood flow [24, 25]. Abnormally increased platelet production can lead to blood clots and an increased risk for stroke [29]. Patients who experience strokes are at heightened risk for seizures, as injuries to the brain can result in scar tissue and disruption to the brain's electrical activity [30]. Although dysregulation of platelet volume or C reactive protein levels may be related to seizures in general, there was limited evidence demonstrating abnormal platelet or C reactive protein levels in this particular UDN patient. However, these levels may have entered a normal range as a result of drug treatments for their symptoms. A potential follow-up test for this patient could be measuring the erythrocyte sedimentation rate (ESR). ESR can be more useful for monitoring chronic inflammation, as C reactive protein levels rise and return to normal more quickly than ESR levels [31].

PIGQ

UKB participants with variants in the PIGQ gene were enriched for abnormal "trunk fat-free mass" and "trunk predicted mass", two phenotype terms that are rarely used outside of the UKB. Trunk fat-free mass indicates abnormality in non-fat tissues such as organs in the abdominal region. Trunk predicted mass refers to general abnormalities in mass in the abdomen. The UDN rare disease patient with compound heterozygous variants in *PIGQ* has symptoms consistent with what could be detected as abnormal mass in abdominal organs—including chronic constipation, bulging pouches in the intestines, volvulus and bowel obstruction, and abdominal distention. Intriguingly, similar digestive symptoms as well as epilepsy have been found in a cohort of patients with variants in *PIGQ* [32].

Although *PIGQ* has been determined to be seizure-associated, the mechanism underlying this connection, including any link between digestive and neurological symptoms, has not been well defined [33]. The PIGQ gene encodes for Phosphatidylinositol Glycan Anchor Biosynthesis Class Q, a protein that helps synthesize glycosylphosphatidylinositol (GPI) anchors. GPI anchors attach proteins to the outer surface of cell membranes and are generally involved in cell-cell adhesion and the transport of molecules through cell membranes [34]. Though mutations in genes coding for GPI anchors have largely been associated with neurodevelopmental conditions, nervous system abnormalities are likely to impact the digestive system as well via the gut-brain axis. The gut-brain axis tightly links the nervous and digestive systems via neural, hormonal and immunological pathways to enable constant bidirectional communication between the gut and brain [32]. Gastrointestinal disorders have been found to occur at higher rates in patients with epilepsy in general, and constipation is the second most common comorbidity in children with epilepsy [35, 36]. Indeed, chronic constipation results in a buildup of toxins that may be seizure-triggering [36]. Variants in *PIGQ* could therefore contribute to epilepsy and indirectly to gastrointestinal abnormalities through a multi-step pathway.

SQSTM1

UKB participants with variants in *SQSTM1* were enriched for "Paget's Disease of Bone", a disorder characterized by joint stiffness, weakness, and abnormal bone growth resulting from new bone failing to effectively replace old bone [37]. The association between autosomal dominant *SQSTM1* mutations and Paget's Disease of Bone is well-established and consistent with this identified enrichment [38]. Paget's Disease of Bone does not classically present with seizures, but the disease's associated symptoms help identify the mechanistic pathways that are disrupted by *SQSTM1* abnormalities. It is not uncommon for genes to be associated with both autosomal dominant (potentially gain-of-function) and autosomal recessive (likely loss-of-function) disorders that involve related mechanisms but do not have phenotypically identical presentations [39].

SQSTM1 has been identified as a potentially seizure-related gene, as abnormal skull growth can cause cerebellar compression and increased intracranial pressure that results in seizures in the extreme case [33, 40]. The UDN patient with compound heterozygous variants impacting *SQSTM1* experiences abnormal movements and stiffness. Additionally, the patient's skull circumference is smaller than the third percentile and had been characterized as infantile-onset microcephaly during the first year of life. As a result, the

patient's skull may have been even more compressed, putting them at greater risk for seizure activity.

COL27A1

Individuals with variants in the COL27A1 gene were enriched for abnormal "whole body fat free mass adjusted for body mass index," a UKB-specific phenotype term referring to abnormal bone or organ mass relative to body size. Consistently, the COL27A1 gene is associated with bone malformations, including hip dislocations, cleft palate, scoliosis, and craniofacial differences [41, 42]. The UDN patient with compound heterozygous variants in *COL27A1* has presented with each of these phenotypes. To address the patient's craniosynostosis, the premature fusing of some cranial sutures in the infant's skull, surgery was required. Surgery to implant a shunt was then required to address the patient's subsequent hydrocephalus, or the accumulation of excess fluid within the brain cavities. The connection between patients with shunted hydrocephalus and seizures is well documented, with 20–50% of children with shunts also developing epilepsy [43, 44]. Shunt infections, one of the UDN patient's symptoms, have been found to double the risk of developing epilepsy [45]. The clinical team overseeing this case also feels that the patient's epilepsy is likely a symptom resulting from impaired brain development. Overall, the compound heterozygous variants in *COL27A1* are plausibly linked to the early-onset skeletal and skull abnormalities experienced by the UDN patient. These skeletal abnormalities may have led to seizure onset through this multi-step pathway, suggesting another mechanism potentially underlying seizures.

CACNA2D2

The cohort of carriers for rare variants in *CACNA2D2* were significantly associated with abnormal "high light scatter reticulocyte percentage" and "high light scatter reticulocyte count." These two measurements are an indicator of a greater proportion of immature red blood cells than would be normally expected. Several of the UDN patient's symptoms—looking spacy, seizures with cyanosis (bluish or purplish discoloration of the skin and mucous membranes), pallor (abnormally pale or whitish skin appearance), dark circles under the eyes, and lethargy—are all consistent with a lack of mature red blood cells [46]. Obtaining complete blood count information, including a reticulocyte count or index test specifically, would be an important next step to indicate abnormalities in red blood cell production in this patient.

Although *CACNA2D2* has independently been identified as a potentially seizure-related gene, any mechanism by which it is causatively related to seizures is unknown [33]. Potential seizure-causing mechanisms related to a dysregulation in red blood cell production include oxygen deprivation to the brain, reduction of inhibitory neurotransmitters, and changes in the metabolism of neurons [47].

Discussion

Despite considerable scientific efforts, rare diseases remain exceedingly difficult to diagnose and understand mechanistically. Moreover, they are seldom treated effectively, and only ~ 6% of rare diseases have a treatment available at all [48]. Interpreting findings in

ultrarare disease patients whose conditions are not represented in published literature and who are too unique for traditional case-control statistical analyses requires unique approaches. Developing additional angles by which to assess these cases will expedite and enhance the diagnosis, understanding, and treatment of rare diseases. The availability of resources like the UK Biobank provides aggregated genotype and phenotype data at scale. Phenotypes observed across cohorts of "genotypically similar" individuals from these population-scale cohorts provide context to better understand small sets of unique patients and can illuminate new directions for future exploration, especially in complex cases where progress has otherwise stalled.

Over a quarter of all UDN patients have experienced seizures, underscoring the heterogeneity of mechanisms that may lead to seizure onset. In our pilot study presented here, we found phenotypic associations in UKB participants across four different body systems. These phenotypes suggested multistep pathways that might reasonably culminate in the occurrence of seizures. To eventually find disease modifying therapies, understandings of the upstream seizure-causing processes are crucial.

Insights into milder or later-onset manifestations of disease from the UKB can inform therapeutic development and interventions from which both rare disease patients and individuals with disease-associated variants in the general population may benefit. Even when the disease-causing mechanisms in rare disease patients are understood, the development of treatment options for "N-of-1" patients is often financially infeasible given the limited number of patients who may benefit. However, the same disease mechanisms implicated in ultrarare disorders may lead to similar, treatable symptoms in the general population, incentivizing the development of improved therapeutics. For example, individuals in the broader population who harbor rare variants in disease-causing genes may experience milder or later-onset symptoms related to the severe symptoms that manifest in rare disease cases. Finding these links can inform positive interventions for both rare disease patients and for heterozygous variant carriers. Additionally, individuals who don't fully express disease phenotypes could offer insights into resilience factors and the role of potential genetic or environmental modifiers.

Given the unique and complex nature of each UDN case, further work is still necessary to conclusively validate any of these hypothesized pathways. In the six vignettes we presented here, the clinical teams overseeing each case accepted that the multistep pathways leading to seizure onset were plausible, but there would be no easy way to confirm them or provide further evidence in support of or against the associations. The benefits of any additional testing, such as blood work or research assays, must be carefully considered given the high burden of medical tests and interventions these patients already undergo. Additionally, we do not explore the potential underlying causes of dysfunction at the molecular level. Further study of the biological pathways involving the implicated genes could provide additional insight.

Another important consideration is that many patients in the UDN had multiple strong candidate diagnostic genes listed in the UDN Gateway. Although many disorders may be monogenic, some patients may instead have multiple genes implicated that together give rise to simultaneous conditions. Recessive variants in each patient may not fully explain a patient's entire syndromic presentation. Moreover, not all genes are equally strong candidates in any given case. We discarded genes from our analysis

if the clinical team determined that a UDN patient had a different gene that was more likely to be causative or relevant to the patient's seizures.

The aims of this pilot study were two-fold: to investigate the etiology of seizures, but also to inspire future studies to interpret rare disease cases by leveraging large-scale biobank data. There are many directions to extend the pilot project presented here. First, it would be worthwhile to repeat a similar approach with all UDN patients or a subset of patients with a different initial phenotype with diverse, complex, and unknown causes. Second, this study could be further refined by considering and differentiating based on variant functionality. In our work, UKB participants were considered genotypically similar if they harbored any rare variant in a gene of interest. However, variants at different positions in the same gene can have varying or even contradictory impacts, so defining genotypically similar individuals as only those harboring variants with functionally similar impacts as the pathogenic variants in a rare disease patient is reasonable. However, restricting the cohort of "genotypically similar" individuals available for analysis in this manner may also reduce the statistical power needed to detect significant associations. Third, incorporating candidate noncoding variants detectable from whole genome sequencing, rather than exome-only sequencing, would enable the consideration of many more variants and possible phenotypic associations. Although UDN patients have whole genome sequencing available, the UKB participants with precomputed phenotype associations available in Genebass only had exome data available.

A plethora of prior work has leveraged large-scale biobanks such as the UK Biobank to further our collective understanding of rare variants and their role in human disease. For instance, Patrick et al. identified patients within the UK Biobank diagnosed with rare diseases to identify new, potentially pathogenic variants for these diseases [49]. Wang et al. and Jurgens et al. introduced foundational approaches associating rare variants with common diseases and phenotypes [50, 51]. In our work, we begin with patients already diagnosed with ultrarare or unique genetic conditions through N-of-1 analyses, and then leverage biobank data to further understand their conditions in an effort to shorten the therapeutic odyssey that follows the diagnostic odyssey for many of these patients.

Conclusions

The UDN provides both unique challenges and exciting opportunities for study. UDN patients are truly one-of-a kind, making their conditions difficult to interpret and contextualize. Studying these complex conditions also highlights gaps in our overall biological understanding and pushes the boundaries of current medical expertise. Analyzing these rare disease patients in the context of a larger cohort of genotypically similar biobank patients is one way to address the challenge of the conditions' uniqueness and seizes the opportunity to better understand the underlying disease-causing mechanisms.

Abbreviations

UDN	Undiagnosed Diseases Network
UKB	United Kingdom Biobank
HPO	Human Phenotype Ontology
CADD	Combined Annotation Dependent Depletion
GPI	Glycosylphosphatidylinositol

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s13040-024-00418-5.

Supplementary Material 1.

Supplementary Material 2.

Acknowledgements
We would like to thank the UDN, all of the participants in the UDN, as well as the clinical teams—including Brian Corner, Dr. John Phillips, Rebecca Spillmann, Dr. David Sweetser, Lauren Briere, Dr. James Orengo, and Jill Mokry—for their valuable insights.
This research has been conducted using data from UK Biobank, a major biomedical database at www.ukbiobank.ac.uk.
Members of the Undiagnosed Diseases Network:
Maria T. Acosta, David R. Adams, Ben Afzali, Ali Al-Beshri, Eric Allenspach, Aimee Allworth, Raquel L. Alvarez, Justin Alvey, Ashley Andrews, Euan A. Ashley, Carlos A. Bacino, Guney Bademci, Ashok Balasubramanyam, Dustin Baldridge, Erin Baldwin, Jim Bale, Michael Bamshad, Deborah Barbouth, Pinar Bayrak-Toydemir, Anita Beck, Alan H. Beggs, Edward Behrens, Gill Bejerano, Hugo J. Bellen, Jimmy Bennett, Jonathan A. Bernstein, Gerard T. Berry, Anna Bican, Stephanie Bivona, Elizabeth Blue, John Bohnsack, Devon Bonner, Nicholas Borja, Lorenzo Botto, Lauren C. Briere, Elizabeth A. Burke, Lindsay C. Burrage, Manish J. Butte, Peter Byers, William E. Byrd, Kaitlin Callaway, John Carey, George Carvalho, Thomas Cassini, Sirisak Chanprasert, Hsiao-Tuan Chao, Ivan Chinn, Gary D. Clark, Terra R. Coakley, Laurel A. Cobban, Joy D. Cogan, Matthew Coggins, F. Sessions Cole, Brian Corner, Rosario I. Corona, William J. Craigen, Andrew B. Crouse, Vishnu Cuddapah, Precilla D'Souza, Hongzheng Dai, Kahlen Darr, Surendra Dasari, Joie Davis, Margaret Delgado, Esteban C. Dell'Angelica, Katrina Dipple, Daniel Doherty, Naghmeh Dorrani, Jessica Douglas, Emilie D. Douine, Dawn Earl, Lisa T. Emrick, Christine M. Eng, Cecilia Esteves, Kimberly Ezell, Elizabeth L. Fieg, Paul G. Fisher, Brent L. Fogel, Jiayu Fu, William A. Gahl, Rebecca Ganetzky, Emily Glanton, Ian Glass, Page C. Goddard, Joanna M. Gonzalez, Andrea Gropman, Meghan C. Halley, Rizwan Hamid, Neal Hanchard, Kelly Hassey, Nichole Hayes, Frances High, Anne Hing, Fuki M. Hisama, Ingrid A. Holm, Jason Hom, Martha Horike-Pyne, Alden Huang, Yan Huang, Anna Hurst, Wendy Introne, Gail P. Jarvik, Suman Jayadev, Orpa Jean-Marie, Vaidehi Jobanputra, Oguz Kanca, Yigit Karasozen, Shamika Ketkar, Dana Kiley, Gonench Kilich, Eric Klee, Shilpa N. Kobren, Isaac S. Kohane, Jennefer N. Kohler, Bruce Korf, Susan Korrick, Deborah Krakow, Elijah Kravets, Seema R. Lalani, Christina Lam, Brendan C. Lanpher, Ian R. Lanza, Kumarie Latchman, Kimberly LeBlanc, Brendan H. Lee, Kathleen A. Leppig, Richard A. Lewis, Pengfei Liu, Nicola Longo, Joseph Loscalzo, Richard L. Maas, Ellen F. Macnamara, Calum A. MacRae, Valerie V. Maduro, AudreyStephannie Maghiro, Rachel Mahoney, May Christine V. Malicdan, Rong Mao, Ronit Marom, Gabor Marth, Beth A. Martin, Martin G. Martin, Julian A. Martínez-Agosto, Shruti Marwaha, Allyn McConkie-Rosell, Ashley McMinn, Matthew Might, Mohamad Mikati, Danny Miller, Ghayda Mirzaa, Breanna Mitchell, Paolo Moretti, Marie Morimoto, John J. Mulvihill, Lindsay Mulvihill, Mariko Nakano-Okuno, Stanley F. Nelson, Serena Neumann, Dargie Nitsuh, Donna Novacic, Devin Oglesbee, James P. Orengo, Laura Pace, Stephen Pak, J. Carl Pallais, Neil H. Parker, LéShon Peart, Leoyklang Petcharet, John A. Phillips III, Filippo Pinto e Vairo, Jennifer E. Posey, Lorraine Potocki, Barbara N. Pusey Swerdzewski, Aaron Quinlan, Daniel J. Rader, Ramakrishnan Rajagopalan, Deepak A. Rao, Anna Raper, Wendy Raskind, Adriana Rebelo, Chloe M. Reuter, Lynette Rives, Lance H. Rodan, Martin Rodriguez, Jill A. Rosenfeld, Elizabeth Rosenthal, Francis Rossignol, Maura Ruzhnikov, Marla Sabaii, Jacinda B. Sampson, Timothy Schedl, Lisa Schimmenti, Kelly Schoch, Daryl A. Scott, Elaine Seto, Vandana Shashi, Emily Shelkowitz, Sam Sheppeard, Jimann Shin, Edwin K. Silverman, Giorgio Sirugo, Kathy Sisco, Tammi Skelton, Cara Skraban, Carson A. Smith, Kevin S. Smith, Lilianna Solnica-Krezel, Ben Solomon, Rebecca C. Spillmann, Andrew Stergachis, Joan M. Stoler, Kathleen Sullivan, Shamil R. Sunyaev, Shirley Sutton, David A. Sweetser, Virginia Sybert, Holly K. Tabor, Queenie Tan, Arjun Tarakad, Herman Taylor, Mustafa Tekin, Willa Thorson, Cynthia J. Tifft, Camilo Toro, Alyssa A. Tran, Rachel A. Ungar, Adeline Vanderver, Matt Velinder, Dave Viskochil, Tiphanie P. Vogel, Colleen E. Wahl, Melissa Walker, Nicole M. Walley, Jennifer Wambach, Michael F. Wangler, Patricia A. Ward, Daniel Wegner, Monika Weisz Hubshman, Mark Wener, Tara Wenger, Monte Westerfield, Matthew T. Wheeler, Jordan Whitlock, Lynne A. Wolfe, Heidi Wood, Kim Worley, Shinya Yamamoto, Zhe Zhang, Stephan Zuchner

Authors' contributions
LF contributed to methodology design, software development, data analysis, and manuscript writing. SNK contributed to literature review, utilization of UDN resources, results interpretation, and manuscript writing. BBJ contributed to question conceptualization, study design, software workflow, data interpretation, and manuscript revisions. All authors critically reviewed the manuscript and read and approved the final version.

Funding
Research reported in this manuscript was supported by the National Institutes of Health (NIH) Common Fund grant U01HG007530 and the NIH National Institute of Neurological Disorders and Stroke (NINDS) grants U2CNS132415 and R00NS114850. The content is solely the authors' responsibility and does not necessarily represent the official views of the National Institutes of Health.

Data availability
No datasets were generated during the current study.

Declarations

Ethics approval and consent to participate
The authors declare no competing interests. This work was performed in accordance with all ethical guidelines outlined in the NIH IRB #15HG0130. The study proposal and manuscript were approved by the Undiagnosed Diseases Network (UDN) Publications and Research Committee.

Consent for publication
All patients are consented as part of their enrollment into the UDN in accordance with the UDN manual of operations [52].

Competing interests
The authors declare no competing interests.

Received: 23 August 2024 Accepted: 23 December 2024
Published online: 17 January 2025

References

1. 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):165–73.
2. Crooke ST. A call to arms against ultra-rare diseases. Nat Biotechnol. 2021;39(6):671–7.
3. Shah S, Dooms MM, Amaral-Garcia S, Igoillo-Esteve M. Current drug repurposing strategies for rare neurodegenerative disorders. Front Pharmacol. 2021;12:768023.
4. Angelis A, Tordrup D, Kanavos P. Socio-economic burden of rare diseases: a systematic review of cost of illness evidence. Health Policy. 2015;119(7):964–79.
5. Gene map statistics - OMIM. https://www.omim.org/statistics/geneMap. Accessed 11 Jul 2024.
6. Gulani A, Weiler T. Genetics, autosomal recessive. In: StatPearls. Treasure Island: StatPearls Publishing; 2023.
7. Guo W, Yuan J, Dong Q, Wang E. Highly sequence-dependent formation of fluorescent silver nanoclusters in hybridized DNA duplexes for single nucleotide mutation identification. J Am Chem Soc. 2010;132(3):932–4.
8. Farrell PM, White TB, Ren CL, et al. Diagnosis of cystic fibrosis: consensus guidelines from the cystic fibrosis foundation. J Pediatr. 2017;181:S4-S15.e1.
9. Fajac I, De Boeck K. New horizons for cystic fibrosis treatment. Pharmacol Ther. 2017;170:205–11.
10. Miller AC, Comellas AP, Hornick DB, et al. Cystic fibrosis carriers are at increased risk for a wide range of cystic fibrosis-related conditions. Proc Natl Acad Sci U S A. 2020;117(3):1621–7.
11. Wright CF, FitzPatrick DR, Firth HV. Paediatric genomics: diagnosing rare disease in children. Nat Rev Genet. 2018;19(5):253–68.
12. Kalyta K, Stelmaszczyk W, Szczęśniak D, Kotuła L, Dobosz P, Mroczek M. The spectrum of the heterozygous effect in biallelic mendelian diseases-the symptomatic heterozygote issue. Genes (Basel). 2023;14(8):1562.
13. Watts JA, Morley M, Burdick JT, et al. Gene expression phenotype in heterozygous carriers of ataxia telangiectasia. Am J Hum Genet. 2002;71(4):791–800.
14. Kobren SN, Baldridge D, Velinder M, et al. Commonalities across computational workflows for uncovering explanatory variants in undiagnosed cases. Genet Med. 2021;23(6):1075–85.
15. Strande JL. Haploinsufficiency MYBPC3 mutations: another stress induced cardiomyopathy? Let's take a look! J Mol Cell Cardiol. 2015;79:284–6.
16. Beghi E. The epidemiology of epilepsy. Neuroepidemiology. 2020;54(2):185–91.
17. Elger CE, Schmidt D. Modern management of epilepsy: a practical approach. Epilepsy Behav. 2008;12(4):501–39.
18. Fitzsimmons L, Kobren SN, Beaulieu-Jones B. Supplemental Table 1: Seizure and seizure-related HPO terms. Figshare. 2023. https://doi.org/10.6084/m9.figshare.23818869.v1.
19. Karczewski KJ, Solomonson M, Chao KR, et al. Systematic single-variant and gene-based association testing of thousands of phenotypes in 394,841 UK Biobank exomes. Cell Genom. 2022;2(9):100168.
20. Lee S, Emond MJ, Bamshad MJ, et al. Optimal unified approach for rare-variant association testing with application to small-sample case-control whole-exome sequencing studies. Am J Hum Genet. 2012;91(2):224–37.
21. Zhou W, Zhao Z, Nielsen JB, et al. Scalable generalized linear mixed model for region-based association tests in large biobanks and cohorts. Nat Genet. 2020;52(6):634–9.
22. Malow BA. Sleep deprivation and epilepsy. Epilepsy Curr. 2004;4(5):193–5.
23. Dell KL, Payne DE, Kremen V, et al. Seizure likelihood varies with day-to-day variations in sleep duration in patients with refractory focal epilepsy: a longitudinal EEG investigation. EClinicalMedicine. 2021;4. https://doi.org/10.1101/2021.05.03.21256436.
24. Gasparyan AY, Ayvazyan L, Mikhailidis DP, Kitas GD. Mean platelet volume: a link between thrombosis and inflammation? Curr Pharm Des. 2011;17(1):47–58.
25. Fay WP. Linking inflammation and thrombosis: role of C-reactive protein. World J Cardiol. 2010;2(11):365–9.
26. Zhong R, Chen Q, Li M, Zhang X, Lin W. Elevated blood C-reactive protein levels in patients with epilepsy: a systematic review and meta-analysis. Front Neurol. 2019;10:974.
27. Sohn HS, Kim SK, Lee SY. Inflammatory markers associated with seizures. Epileptic Disord. 2016;18(1):51–7.
28. Millichap JG. C-reactive protein serum levels after seizures. Pediatr Neurol Briefs. 2012;26(6):41.
29. Ot S, Zafar L, Beg M, Siddiqui OA. Association of mean platelet volume with risk factors and functional outcome in acute ischemic stroke. J Neurosci Rural Pract. 2021;12(4):764–9.
30. Yang H, Rajah G, Guo A, Wang Y, Wang Q. Pathogenesis of epileptic seizures and epilepsy after stroke. Neurol Res. 2018;40(6):1–7.
31. Assasi N, Blackhouse G, Campbell K, et al. Comparative value of Erythrocyte Sedimentation Rate (ESR) and C-Reactive protein (CRP) testing in combination versus individually for the diagnosis of undifferentiated patients with suspected inflammatory disease or serious infection: a systematic review and economic analysis. Ottawa: Canadian Agency for Drugs and Technologies in Health; 2015.
32. Johnstone DL, Nguyen TTM, Zambonin J, et al. Early infantile epileptic encephalopathy due to biallelic pathogenic variants in PIGQ: report of seven new subjects and review of the literature. J Inherit Metab Dis. 2020;43(6):1321–32.
33. Wang J, Lin ZJ, Liu L, et al. Epilepsy-associated genes. Seizure. 2017;44:11–20.

34. Kinoshita T, Fujita M. Biosynthesis of GPI-anchored proteins: special emphasis on GPI lipid remodeling. J Lipid Res. 2016;57(1):6–24.
35. Avorio F, Cerulli Irelli E, Morano A, et al. Functional gastrointestinal disorders in patients with epilepsy: reciprocal influence and impact on seizure occurrence. Front Neurol. 2021;12:705126.
36. Moezi L, Pirsalami F, Inaloo S. Constipation enhances the propensity to seizure in pentylenetetrazole-induced seizure models of mice. Epilepsy Behav. 2015;44:200–6.
37. Nairn C, Ralston SH. Paget's disease of bone: when and why to refer to specialist care. Br J Gen Pract. 2020;70(700):561–2. https://doi.org/10.3399/bjgp20X713369.
38. Daroszewska A, Ralston SH. Mechanisms of disease: genetics of Paget's disease of bone and related disorders. Nat Clin Pract Rheumatol. 2006;2(5):270–7.
39. Chung HL, Wangler MF, Marcogliese PC, et al. Loss- or gain-of-function mutations in ACOX1 cause axonal loss via different mechanisms. Neuron. 2020;106(4):589-606.e6.
40. Poncelet A. The neurologic complications of paget's disease. J Bone Min Res. 1999;14(S2):88–91.
41. Gonzaga-Jauregui C, Gamble CN, Yuan B, et al. Mutations in COL27A1 cause Steel syndrome and suggest a founder mutation effect in the Puerto Rican population. Eur J Hum Genet. 2015;23(3):342–6.
42. Kim JS, Jeon H, Lee H, et al. Biallelic novel mutations of the COL27A1 gene in a patient with steel syndrome. Hum Genome Var. 2021;8(1):17. https://doi.org/10.1038/s41439-021-00149-7.
43. Sato O, Yamguchi T, Kittaka M, Toyama H. Hydrocephalus and epilepsy. Childs Nerv Syst. 2001;17(1–2):76–86.
44. Keene DL, Ventureyra EC. Hydrocephalus and epileptic seizures. Childs Nerv Syst. 1999;15(4):158–62.
45. Tully HM, Kukull WA, Mueller BA. Clinical and surgical factors associated with increased epilepsy risk in children with hydrocephalus. Pediatr Neurol. 2016;59:18–22.
46. Baldwin C, Pandey J, Olarewaju O. Hemolytic anemia. In: StatPearls. Treasure Island: StatPearls Publishing; 2023.
47. Padda J, Khalid K, Syam M, et al. Association of anemia with epilepsy and antiepileptic drugs. Cureus. 2021;13(11):e19334.
48. Dawkins HJS, Draghia-Akli R, Lasko P, et al. Progress in rare diseases research 2010–2016: an IRDiRC perspective. Clin Transl Sci. 2018;11(1):11–20.
49. Patrick MT, Bardhi R, Zhou W, Elder JT, Gudjonsson JE, Tsoi LC. Enhanced rare disease mapping for phenome-wide genetic association in the UK Biobank. Genome Med. 2022;14(1):85. https://doi.org/10.1186/s13073-022-01094-y.
50. Wang Q, Dhindsa RS, Carss K, et al. Rare variant contribution to human disease in 281,104 UK Biobank exomes. Nature. 2021;597(7877):527–32.
51. Jurgens SJ, Choi SH, Morrill VN, et al. Analysis of rare genetic variation underlying cardiometabolic diseases and traits among 200,000 individuals in the UK Biobank. Nat Genet. 2022;54(3):240–50.
52. Undiagnosed Diseases Network. Undiagnosed diseases network manual of operations. Boston: Undiagnosed Diseases Network; 2024. https://undiagnosed.hms.harvard.edu/research/udn-manual-of-operations/.

Publisher's Note