How can real-world data, registries, and databases address the challenges of rare diseases?

Sara E. Mole[1], **Emily Gardner**[1], **Heather L. Mason**[2]

[1] University College London, UCL Great Ormond Street Institute of Child Health

[2] Coufetery Comms, Medical Writing Services, France

doi: 10.56012/kwle5627

Correspondence to:
Sara Mole
s.mole@ucl.ac.uk

Abstract

A goal for all diseases is a treatment that works to prevent, halt, or reverse their effects – essentially, a cure. Achieving this requires early diagnosis, knowledge of disease mechanisms, and effective treatment. For rare diseases, each of these elements is a huge challenge. This perspective explores how real-world and registry data can address these obstacles and considers future developments with the potential for the most significant impact.

A goal for all diseases is a treatment that works to prevent, halt, or reverse their effects – essentially a cure. The quest to understand and treat rare diseases is among the most challenging and vital missions in healthcare today. In Europe, over 20 million people live with a rare disease, with about 75% affecting children under the age of 2 years, and more than 260 million people are affected worldwide, about 5% of the total population.[1] A disease is classed as rare if it affects less than 5 in 10,000 or 1 in 2,000 of the European population, or fewer than 200,000 people in the USA. There are over 6,000 rare diseases, many of which are life-limiting and lack effective preventative or curative treatment. These are often inherited metabolic diseases and so can affect several children within a family. For ultrarare diseases, only a few families may be diagnosed.

In this perspective, we highlight the need for real-world evidence and registries that capture patient data and summarise how these can address the specific challenges of rare diseases.

Core concepts of real-world evidence and registries in the context of rare diseases

For common diseases affecting many people, it is relatively easy to collect data about the disease and to find enough people willing to participate in clinical trials. However, this is not the case for rare diseases. Therefore, real-world data (RWD) is an essential source of information for rare diseases and can help with diagnosis, treatment development, clinical management, and research.[2] Essentially, RWD is a collection of patient health-related data. This is most useful when held in electronic format to allow processing by codifying to aid analysis. RWD can include data from patient registries and hospital records, including regular checkups and other sources such as wearable devices, smartphones, and information provided by patients or disease registries.

RWD can provide information on disease prevalence, incidence, and natural history and can be used for scientific health research and public health purposes. RWD can be either structured, such as laboratory orders, prescriptions, and lists of procedures; semi-structured, which includes digital images that contain structured attributes like device identification and DateTime stamp; or unstructured, such as clinical progress notes, pathology reports, radiology reports, patient correspondence, and insurance letters. Unstructured data has a lot of richness due to its diverse, variable, and sometimes unpredictable nature, but it is not easy to code and analyse.

RWD can provide information on disease prevalence, incidence, and natural history and can be used for scientific health research and public health purposes.

Codifying RWD into electronic format so that it can be analysed has huge potential. Machine learning with clinical narratives containing deep and detailed phenotypes can recognise new patterns across the whole group of patients and tie these to individual patients to estimate disease activity, including progression and remission and recognition of different disease subtypes.[3] This can give a clearer picture of a patient's history, provide more details about disease trajectory, and provide early warning signs for adapting care, an emerging critical clinical event, or even a new diagnosis. It can be powerful to link a problem needing a solution with real-world evidence (RWE) and artificial intelligence (AI).

Patient registries have been set up for many rare diseases to gather information required for treatment development in one place. There are different types of patient registries.[4] These can be based on a single or group of related diseases, assembled to gather data to test a new product in a clinical trial, or draw data from a particular population. Registries may be set up either by pharma developing a product and restricting the data for internal use or by patient organisations or clinical consortia, in which case the data may be available for others to use.

The scientific evidence derived from analysing RWD is called real world evidence (RWE). For example, gathering and analysing clinical evidence of the benefits or risks of a new medicinal product is RWE. It can be used to support regulatory purposes, such as the first applications for marketing authorisations for orphan medicines. In this way, RWD and RWE can be used to bring new therapies to patients.

Data from clinical trials are prospectively obtained with a predetermined purpose and often from a specially determined and limited group of similar patients. However, RWD is observational, can be large in size, and is frequently drawn from a variety of patient backgrounds. Therefore, RWD can be messy,

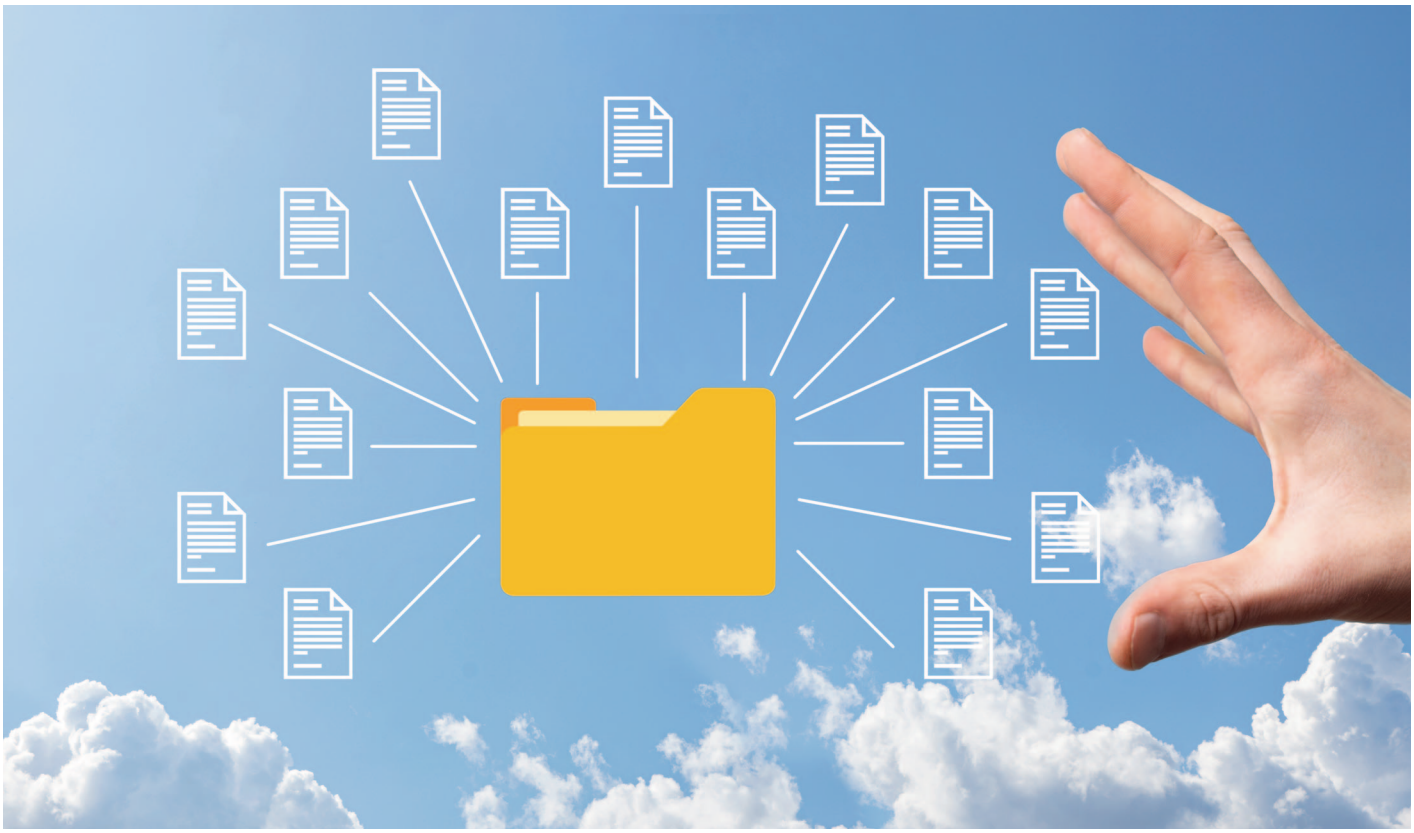

Photo: freepik

incomplete, and subject to bias. RWD complements traditional clinical research data. Equally, RWD can consolidate knowledge from data that may not be collected during clinical trials, such as the impacts of economic and social factors and the quality of life of patients with rare diseases. These additional data further enhance the evidence-based decisions made when bringing new medicines to patients, especially as waiting for the next trial may be too late for some.

Unique challenges of rare diseases

People with rare diseases are scattered across the globe, and so is their data. The collection of such data and its use is vital and challenging, partly because of its scarcity but also the heterogeneity of the patient population. Gathering RWD, especially those collected during daily life, may reduce the number of hospital visits and avoid the need to relocate during a clinical trial, which has massive implications on family life and resources. Designing a clinical trial to include RWD can benefit families, although there are concerns about the quality and comparability of RWD with randomised clinical trial data.

There is limited knowledge across many aspects of some rare diseases, especially ultrarare ones, from their natural history to patho-mechanisms and correlations between genotype and phenotype. Rare diseases often face delays in diagnosis due to the time taken to first rule out more common diagnoses. Specialised tests may be needed for confirmation, but these are not available to all patients around the globe.

Patient heterogeneity arises from the underlying genetic cause in allelic diseases. In some, this results in a complete loss of function of a single disease gene; in others, the retention of partial function is due to genetic variation such as missense mutations. Heterogeneity may additionally reflect other genomic influences beyond the disease gene, and environmental factors such as diet and living conditions, which vary globally. Patient registries that include genetic variation data can be invaluable here.

> Unstructured data has a lot of richness due to its diverse, variable, and sometimes unpredictable nature, but it is not easy to code and analyse.

Benefits of patient registries

There are many benefits of setting up patient registries to provide RWD. They provide information on the natural history of a rare disease, the incidence, the expected numbers of patients eligible for a clinical trial, the choice of endpoints in clinical trial design, tracking treatment outcomes, quality of life assessments which lead to healthcare resource utilisation, and post-market surveillance once a new medicine is available in the clinic. For rare diseases, this has led to the recognition that no group needs to receive placebo treatment during clinical trials. Collection of RWD might reveal unmet care needs that can then be addressed. During expanded clinical trials, RWE provides information on treatment efficacy in patients who may differ in their genetic variation or support settings. RWE supplements the more restricted early clinical trials, typically involving only a small number of patients.

Best practices and examples of registry design

The holding of personal data, including health data, is regulated, creating challenges in sharing this data as regulations are different worldwide. Families with rare diseases are often very willing to allow the collection of their data, as they understand its importance in research and development. For registries to provide RWD, they require quality assurance processes for data organisation, data quality, consideration of potential biases, and for the database to be fit for purpose.

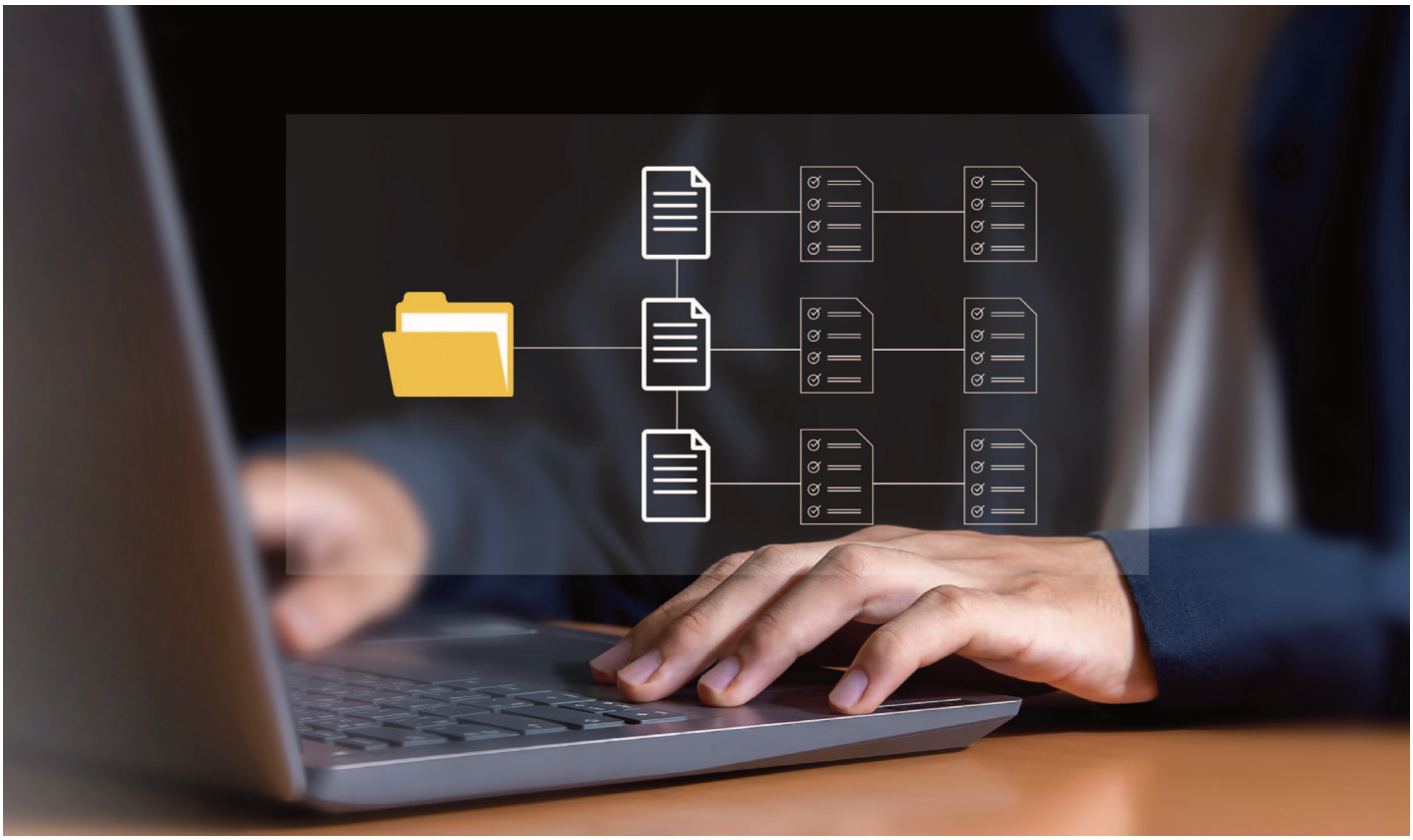

Photo: freepik

A recent survey identified many rare disease registries,[5] with most based in Europe, predominantly led from Germany, the UK, or the USA. Some hold more than 30,000 cases. Two-thirds cover a range of diseases, and a third only one disease. Most are national, with many others continental or global, which requires interoperability in terms of data elements ontologies, and common terminologies to allow data collected in different places to be combined. They aim to provide participants for clinical studies, to evaluate or improve clinical care, to describe epidemiology, or to improve understanding of the natural history. Data collected includes sociodemography, diagnosis, medical history, care pathway, and treatment history. Approximately one-fifth of registries use common or core data or ontological coding language, which considers what the data is about, defines variables, and translates the data to create standardised terms for global use. Nearly half have no clear governance. Many include patient-reported outcomes, but not all involve all potential users, such as patient organisations, in their design. Funding came mainly from federal or European Union bodies, with many funded by private pharmaceutical or technical companies.

> The age of onset is distinguishable from the age of diagnosis, given that the time between these may be considerable.

One example of a rare disease registry is Sanofi's Rare Disease Registries,[6] which was set up 30 years ago and expanded to collect data on rare lysosomal storage disorders (LSDs) (Fabry, Gaucher, Mucopolysaccharidosis type I, and Pompe disease). This registry contains data from over 18,000 patients who have one of these four LSDs and are enrolled at over 800 sites in 64 countries. There is now a Rare Disease Registries Patient Council, which is leading to further improvements. This RWD has led to more than 100 peer-reviewed articles published to advance learning on these diseases.[7]

Another example from our personal experience is DEM-CHILD, a patient registry for neuronal ceroid lipofuscinoses (NCL), also known as Batten disease. This registry was initiated by collaborating European clinicians and led by Dr Angela Schulz to improve early diagnosis and optimise standards of care.[8] DEM-CHILD registers patients with different forms of NCL to measure the prevalence of each type of NCL in participating countries. It collects retrospective and prospective patient data to precisely describe the clinical course and its variability in the different forms of NCL, correlating patients' genotypes with their phenotypes by linking clinical and genetic mutation data. DEM-CHILD also provides a tool for evaluating experimental therapy studies and palliative therapies. The registry currently has over 250 patients in the database.

DEM-CHILD follows best practices for registry design, with ethical approval, and it follows European data protection guidelines. There is an approved audit trail to ensure data safety, and the data is stored on different servers with emergency power supply and daily backup.

Since its founding, improvements have been made, and there are plans to allow parents to contribute data. The registry harmonises data collection and sharing and facilitates non-exclusive data sharing with third parties globally,

such as scientists and pharma. This supports the development of various therapies and the collection and sharing of patient samples with third parties. Established and novel clinical rating scales have been applied to assess disease progression for different NCL types, and quality-of-life questionnaires utilised. Clinical assessments are comprehensive for both the central nervous system (CNS) and extra CNS disease manifestations. A collection of serum and cerebrospinal fluid samples is available in the associated DEM-CHILD biobank.

A mark of its success is that the EMA and the FDA accepted the natural history data held in DEM-CHILD for late infantile CLN2 disease as valid natural-history controls for the efficacy evaluations in experimental therapies for CLN2 disease. This led to an expedited approval of intracerebroventricular enzyme replacement therapy with cerliponase alpha in May 2017.[9-11] There are other examples of similar successes utilising rare disease registries.[12]

There is a need to understand genetic variation and how this correlates with disease progression. In parallel with DEM-CHILD, the NCL-Resource contains the freely accessible NCL Mutation Database. This curated database collects published data on the genetics and phenotype of NCL patients and gathers this data in one place. The data inspires scientific design and can be used to predict disease severity and consider implications for therapeutic development.[13-15] More than 700 genetic variations in NCL genes are currently captured, together with details from more than 1,700 patients. The curation focuses on data quality and accuracy. For example, potentially duplicated patient records are highlighted and investigated further with relevant clinicians or researchers. The age of onset is distinguishable from the age of diagnosis, given that the time between these may be considerable. Each variant is checked for accurate Human Genome Variation Society nomenclature for the patient's genetic information. Errors in variant nomenclature are relatively common and, in some cases, have led to the publication of purported new variants when, in reality, they are misdescribed known variants. Thus, for the NCL database, consistent application of several checks by an expert curator increases the quality and accuracy of its data.

Emerging digital health technology allows the capture of digital biomarkers in a home-based disease assessment, which can be expected to provide more consistent RWD than a visit to an unfamiliar clinic. One example is the use of video capture to assess a key transition stage in the loss of independent walking but retention of weight bearing and transfer in the development of Duchenne muscular dystrophy. Such computer vision analysis can extract objective, quantitative measures, including time, movement trajectory patterns, and movement smoothness and symmetry, to identify voluntary or compensatory movements that can mark disease progression. Such RWD could inform clinical endpoints and be used in future clinical trials.[16]

The contribution of medical writing

Professional medical communication writers translate complex information into content that is more accessible in terms of clarity and appropriate for different platforms and target audiences. With respect to RWE and registry data, one important contribution is to enable those who are less familiar with the underpinning medical and scientific concepts to understand their importance and potential. This may allow patients to make an informed decision on whether to give permission for their medical data to be incorporated into a registry or to be analysed, or medical and allied professionals within the rare disease field to appreciate the potential of analysis of medical data and to contribute to this. Further, this requires working closely with those who run the registries and produce the RWE and who are ultimately responsible for driving the accessibility of this impactful research.

Conclusions and future perspectives

We have highlighted the contribution that RWD and registry data are already making towards effective treatments for rare diseases. As the industry seeks innovative solutions, RWE studies utilising RWD have grown in acceptance.[17] It has been argued by many that RWD provides valuable insight into how an investigational medicinal product performs in the real world. In contrast, a randomised controlled trial setting is heavily regulated, with robust patient inclusion and exclusion criteria defined in an approved protocol and trial settings. Therefore, RWD can provide insight that cannot be obtained through traditional means, and it brings in other patient populations that may have been overlooked, so it should not be ignored. This paradigm shift from traditional clinical data to real-world insights marks a new era for researchers, physicians, and patients alike. As the industry adapts, the implications of RWD are revealed, shaping the future of diagnosis, treatment, and patient care.

We suggest that every rare disease should be linked with a registry, and each should be standardised as necessary to offer the best practice for capturing global RWD. Access to this RWD should not be restricted unnecessarily. RWD provided by digital health technology could be improved by home-based regular longitudinal assessments appropriate to the disease. This will increase the potential of AI, including machine learning, to highlight key disease markers beyond clinical markers and open both contributions of data and clinical trials to patients around the world who do not have ready access to specialised centres of clinical excellence. Additionally, all registries should be fit for use by regulatory bodies.[17]

Finally, this perspective is written for medical writers. With their clear writing, these professionals can reach medical engineering professionals, young scientists, and future clinicians who may not yet be reading scientific publications and medical journals to inspire them to contribute to this area of work.

Acknowledgements

The authors thank all those who contribute to disease registries and databases, research publications, and case reports from which NCL mutation data is drawn. This work was supported by an award from the UK Medical Research Council (MR/V033956). Sara Mole (ORCID: 0000-0003-4385-4957) reports support by Biomarin for maintaining the NCL mutation database curated by Emily Gardner. All research at Great Ormond Street Hospital National Health Service (NHS) Foundation Trust and UCL Great Ormond Street Institute of Child Health is made possible by the National Institute for Health and Care Research (NIHR) Great Ormond Street Hospital Biomedical Research Centre.

Disclaimers

The opinions expressed in this article are the authors' own and not necessarily shared by their employer or EMWA. The views expressed are those of the authors and not necessarily those of the NHS, the NIHR, or the Department of Health.

Disclosures and conflicts of interest

The authors declare no conflicts of interest.

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. doi:10.1038/s41431-019-0508-0
2. Liu F, Panagiotakos D. Real-world data: a brief review of the methods, applications, challenges and opportunities. BMC Med Res Methodol. 2022;22(1):287. doi:10.1186/s12874-022-01768-6
3. Jefferies JL, Spencer AK, Lau HA, et al. A new approach to identifying patients with elevated risk for Fabry disease using a machine learning algorithm. Orphanet J Rare Dis. 2021;16(1):518. doi:10.1186/s13023-021-02150-3
4. Pisa F, Arias A, Bratton E, et al. Real world data for rare diseases research: the beginner's guide to registries. Expert Opin Orphan Drugs. 2023;11(1):9–15. doi:10.1080/21678707.2023.2241347
5. Hageman IC, van Rooij IA, de Blaauw I, et al. A systematic overview of rare disease patient registries: challenges in design, quality management, and maintenance. Orphanet J Rare Dis. 2023;18(1):106. doi:10.1186/s13023-023-02719-0
6. Mistry PK, Kishnani P, Wanner C, et al. Rare lysosomal disease registries: lessons learned over three decades of real-world evidence. Orphanet J Rare Dis. 2022;17(1):362. doi:10.1186/s13023-022-02517-0
7. Sanofi. Real-world evidence requires real-world patients. 2024 [cited 2025 Jan 03]. Available from: https://www.sanofi.com/en/magazine/your-health/real-world-evidence-requires-input-from-real-world-patients.
8. Schulz A, Simonati A, Laine M, et al. The DEM-CHILD NCL Patient Database: A tool for the evaluation of therapies in neuronal ceroid lipofuscinoses (NCL). Eur J Paediatr Neurol. 2015;19- Suppl 1: S16. doi:10.1016/S1090-3798(15)30049-0
9. Schulz A, Ajayi T, Specchio N, et al. Study of intraventricular cerliponase alfa for CLN2 disease. N Engl J Med. 2018;378(20):1898–907. doi:10.1056/nejmoa1712649
10. Nickel M, Simonati A, Jacoby D, et al. Disease characteristics and progression in patients with late-infantile neuronal ceroid lipofuscinosis type 2 (CLN2) disease: an observational cohort study. Lancet Child Adolesc Health. 2018;2(8):582–90. doi:10.1016/s2352-4642(18)30179-2
11. Nickel M, Schulz A. Natural history studies in NCL and their expanding role in drug development: Experiences from CLN2 disease and relevance for clinical trials. Front Neurol. 2022;13:785841. doi:10.3389/fneur.2022.785841
12. Liu J, Barrett JS, Leonardi ET, et al. Natural history and real-world data in rare diseases: applications, limitations, and future perspectives. J Clin Pharmacol. 2022;62(Suppl 2):S38–55. doi:10.1002/jcph.2134
13. Gardner E, Bailey M, Schulz A, et al. Mutation update: review of TPP1 gene variants associated with neuronal ceroid lipofuscinosis CLN2 disease. Hum Mutat. 2019;40(11):1924–38. doi:10.1002/humu.23860
14. Gardner E, Mole SE. The genetic basis of phenotypic heterogeneity in the neuronal ceroid lipofuscinoses. Front Neurol. 2021;12:754045. doi:10.3389/fneur.2021.754045
15. Lourenco CM, Pessoa A, Mendes CC, et al. Revealing the clinical phenotype of atypical neuronal ceroid lipofuscinosis type 2 disease: insights from the largest cohort in the world. J Paediatr Child Health. 2021;57(4):519–25. doi:10.1111/jpc.15250
16. Ferrer-Mallol E, Matthews C, Stoodley M, et al. Patient-led development of digital endpoints and the use of computer vision analysis in assessment of motor function in rare diseases. Front Pharmacol. 2022;13:916714. doi:10.3389/fphar.2022.916714
17. Giannuzzi V, Stoyanova-Beninska V, Hivert V. Editorial: The use of real world data for regulatory purposes in the rare diseases setting. Front Pharmacol. 2022; 13:1089033. doi:10.3389/fphar.2022.1089033

Author information

Sara E. Mole, PhD, is a professor of molecular cell biology and the UCL envoy for gender equality at University College London. She studies inherited neurodegenerative diseases that impact children, in particular, Batten disease.

Emily Gardner, PhD, is a molecular biologist with extensive experience in managing research funding and business operations. Emily is also the curator of the NCL database at UCL.

Heather L. Mason is a freelance medical writer with over 14 years of experience, specialising in rare diseases and specifically inborn errors of metabolism. She also has a passion for patient advocacy and inclusion.