www.nature.com/gim

Genetics in Medicine

Check for updates

BRIEF COMMUNICATION

Deep phenotyping unstructured data mining in an extensive pediatric database to unravel a common *KCNA2* variant in neurodevelopmental syndromes

Marie Hully[1], Tommaso Lo Barco[1], Anna Kaminska[1,2], Giulia Barcia[1,3], Claude Cances[4], Cyril Mignot[5], Isabelle Desguerre[1], Nicolas Garcelon[6,7], Edor Kabashi[8] and Rima Nabbout[1,8]✉

PURPOSE: Electronic health records are gaining popularity to detect and propose interdisciplinary treatments for patients with similar medical histories, diagnoses, and outcomes. These files are compiled by different nonexperts and expert clinicians. Data mining in these unstructured data is a transposable and sustainable methodology to search for patients presenting a high similitude of clinical features.

METHODS: Exome and targeted next-generation sequencing bioinformatics analyses were performed at the Imagine Institute. Similarity Index (SI), an algorithm based on a vector space model (VSM) that exploits concepts extracted from clinical narrative reports was used to identify patients with highly similar clinical features.

RESULTS: Here we describe a case of "automated diagnosis" indicated by Dr. Warehouse, a biomedical data warehouse oriented toward clinical narrative reports, developed at Necker Children's Hospital using around 500,000 patients' records. Through the use of this warehouse, we were able to match and identify two patients sharing very specific clinical neonatal and childhood features harboring the same de novo variant in *KCNA2*.

CONCLUSION: This innovative application of database clustering clinical features could advance identification of patients with rare and common genetic conditions and detect with high accuracy the natural history of patients harboring similar genetic pathogenic variants.

Genetics in Medicine (2021) 23:968 – 971; https://doi.org/10.1038/s41436-020-01039-z

INTRODUCTION

Over the last decade, bioinformatics technology has enabled large-scale clinical and epidemiological studies based on retrospective analysis of medical records stored in hospital data warehouses.[1–3] Software that exploits narrative reports in a clinical data warehouse can be especially helpful to define clinical features in a rare disease. These advances mirror the development of next-generation sequencing (NGS) technology that has empowered the discovery of novel genes in a wide range of rare diseases including epilepsy genetics, thus expanding the prevalence of a spectrum of genetic etiologies. However, phenotypic and genotypic heterogeneity in the majority of cases render genetic diagnosis rather complex.[4,5] Indeed, the clinical spectrum for the majority of the variants found in the voltage-gated potassium channel subfamily member *KCNA2* extend from isolated intellectual disability to developmental and epileptic encephalopathies.[6–8] Through multimodal deep phenotyping (clinical, imaging, and electroencephalogram [EEG] data) of two patients harboring the same p.T374A variant in *KCNA2*, patients with similar features were assessed through automated retrospective research using the local data warehouse query program.[9] Identification of the same *KCNA2* variant by exome sequencing for the patient sharing the majority of clinical features demonstrates a strong genotype–phenotype correlation for these three patients, as well as for the four patients harboring the same variant who were previously reported in the literature.[8,10]

MATERIALS AND METHODS

The *KCNA2* variant was identified on DNA extracted from blood samples according to routine procedure, by exome in family trios for patients 1 and 2, and targeted NGS for patient 3. Exome and targeted NGS bioinformatics analyses were performed in house according to routine procedures. The variant was verified by Sanger sequencing in all three patients. Exome sequencing was performed for the three patients and the de novo *KCNA2* variant, p.T374A, was validated by Sanger sequencing in each of these cases.

Data mining using the Necker Hospital database was implemented through the Dr. Warehouse program. This program has been developed through an ongoing collaboration between Imagine Institute and Necker Enfants Malades Hospital, using a vector space model (VSM) and including medical records of approximately 500,000 patients followed in this tertiary hospital.[2] Dr. Warehouse includes patients' records from 60 departments including pediatric neurology, cardiology, gastroenterology, nephrology, intensive care units, emergency, etc. The 6 million documents stored in this data warehouse are inpatient clinical reports, outpatient (clinic) reports, discharge letters, imaging reports, biological results, and pathology reports.

In this model, patients are represented as vectors of medical terms automatically extracted from their narrative reports and concatenated

[1]Reference Centre for Rare Epilepsies, Department of Pediatric Neurology, Necker Enfants Malades hospital, Universite de Paris, Paris, France. [2]Department of clinical neurophysiology, Necker Enfants Malades hospital, APHP, Paris, France. [3]Department of Genetics, Necker Enfants Malades hospital, APHP, Paris, France. [4]Competence Centre for Rare Epilepsies, Toulouse University Hospital, Toulouse, France. [5]Department of Genetics, Groupe Hospitalier Pitié Salpetriere-Trousseau, APHP, Sorbonne University, Paris, France. [6]INSERM, Imagine Institute, UMR 1163, Paris Descartes University, Paris, France. [7]INSERM, UMR 1138 Team 22, Paris Descartes University, Paris, France. [8]Universite de Paris, Imagine Institute, UMR 1163, Team Translational Research for Neurological Diseases, Paris Descartes University, Paris, France. ✉email: rima.nabbout@aphp.fr

based on the Unified Medical Language System (UMLS) Meta-thesaurus developed by the National Library of Medicine.[11]

Similarity Index (SI) is an algorithm based on a VSM that exploits UMLS concepts extracted from clinical narrative reports. The VSM uses three parameters: (1) the polarity of the concept (i.e., if it is negated/not negated), (2) the minimum number of concepts in common, and (3) the term frequency–inverse document frequency (TF-IDF) score. The TF-IDF score is a pertinence score that allows taking into account the high frequency of a phenotype in a patient record and the low frequency of patients with this phenotype in the entire data warehouse.[9] Term frequency (TF) is the number of occurrences of a term for the patient divided by the number of terms for this patient. Inverse document frequency (IDF) is the logarithm of the total number of patients in the data warehouse divided by the number of patients with this term. Therefore, the TF-IDF score allows to filter out noise in the extracted data, defining the phenotypic key concepts (K-concepts).

Clinical data were obtained for all patients with specific emphasis on neonatal history; description of movement or seizure disorders; neurological examinations; clinical evolution concerning developmental milestones, growth, feeding, and sleep disorders, and any relevant medical condition; video-EEG recordings; brain magnetic resonance image (MRI); and metabolic or genetic screenings performed. Video-EEG recordings were performed according to routine procedure, and data were retrospectively collected and analyzed (Fig. S1).

RESULTS

The first patient with a de novo heterozygous p.T374A variant in the *KCNA2* gene was initially detected through exome sequencing analysis in the reference center for rare epilepsies at Necker Hospital. During a polycentric staff meeting, the clinical case of a second patient with very similar clinical features carrying the same de novo *KCNA2* variant was discussed. Due to the strong genotype–phenotype correlation of these two initial patients, we sought to identify other patients harboring *KCNA2* variants ascertaining a specific phenotype that could be very specific for individuals heterozygous for *KCNA2*.

By using similarity queries integrated in the data mining system Dr. Warehouse, we searched 3 million clinical narrative reports for about 500,000 patient records that presented similar clinical features to patient 1, measured as the Similarity Index (SI) (Fig. 1).

We identified five individuals with the highest SI (Fig. 1) and reported their detailed phenotypic features in a clinical heat map (Fig. 2). Patient A showing the highest SI (SI = 66) shared early-onset (at H2) hyperexcitability, Myoclonic jerks (MJ), constant screaming, no eye contact, swallowing difficulties, and recurrent multifocal seizures. Multiple multifocal spikes at one year evolved toward severe encephalopathy at age 3, with permanent global dyskinesia with MJ (Supplementary Fig. 1), spastic quadriplegia, no head control and absent grasping, and scarce eye contact with frequent wandering eye movements. Patient A shared with patients 1 and 2 a number of clinical features, including acquired microcephaly (-4 DS), failure to thrive (-3 DS) with severe Gastro-oesophageal reflux (GOR) and severe sleep disorder with obstructive apnea (Fig. 2). Extensive metabolic screening was normal at 21 months; brain MRI showed cerebellar atrophy and delayed myelination. Patient A had not obtained a definitive genetic diagnosis at the time of the data mining. Through a targeted NGS panel including *KCNA2* for him and both parents, we identified the same de novo heterozygous *KCNA2* variant as in the two previous patients, validating our approach.

The other four patients (B–D) displayed fewer similarities, as described by lower SI scores than patient 1 (Fig. 2). Three patients were tested for *KCNA2* variants through the NGS panel for early onset epileptic encephalopathies, except for patient E due to lack of consent for genetic diagnostic test from the family. No pathogenic variants in *KCNA2* were identified in these individuals.

The data mining software was asked to display the K-concepts extracted from the clinical reports of patient 1 at different ages (0–6 months, 6–12 months, 12–24 months, 24–48 months, and 48–72 months) to analyze how they changed during the follow-up with available electronic reports. The most important concepts to emerge in the 0–6 month range were "myoclonia" (TF-IDF 17,92), "hyperexcitability" (TF-IDF 9,98), "stiffness" (TF-IDF 4,05), "startles" (TF-IDF 3,55), "irritability" (TF-IDF 3,31), "abnormal movements" (TF-IDF 2,94), and "opisthotonus" (TF-IDF 1,71). The main concepts sorted in the ranges 6–12 months and 12–24 months reflected how the clinical picture at this age was still strongly characterized by myoclonic events (TF-IDF 10,31; TF-IDF 14,97), hyperexcitability (TF-IDF 5,3 at 6–12 months), stiffness (TF-IDF 3,74 at 6–12 months), and, in addition, the term "encephalopathy" appeared from 6 months (TF-IDF 5,35; TF-IDF 6,61). From 2 years, the concepts "dystonia" (TF-IDF 3,56 in 24–48 months) and "epileptic encephalopathy" (TF-IDF 1,69 in 48–72 months) appeared, together with a series of new concepts like "respiratory distress," "malaise," "bradycardia," "polypnea," and "inhalation pneumopathy" (Fig. 2). These results provide an accurate description of the natural progression of the disease in this patient.

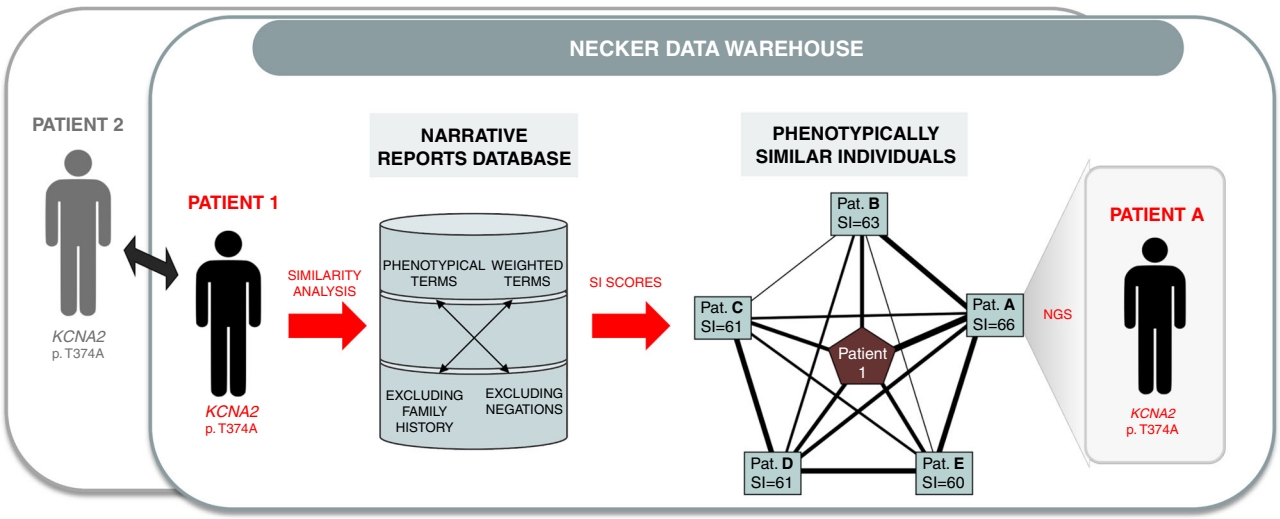

Fig. 1 Display of the two patients (patient 1 from our institution and patient 2 from another institution in our reference center network) sharing the same phenotype and the same *KCN2A* variant. Similarity analysis with all data warehouse narrative reports was performed, yielding a high similarity index (SI) in five patients (patients A–E). Exome sequencing validated that patient A, who had the highest SI, harbored the same *KCNA2* variant. NGS next-generation sequencing.

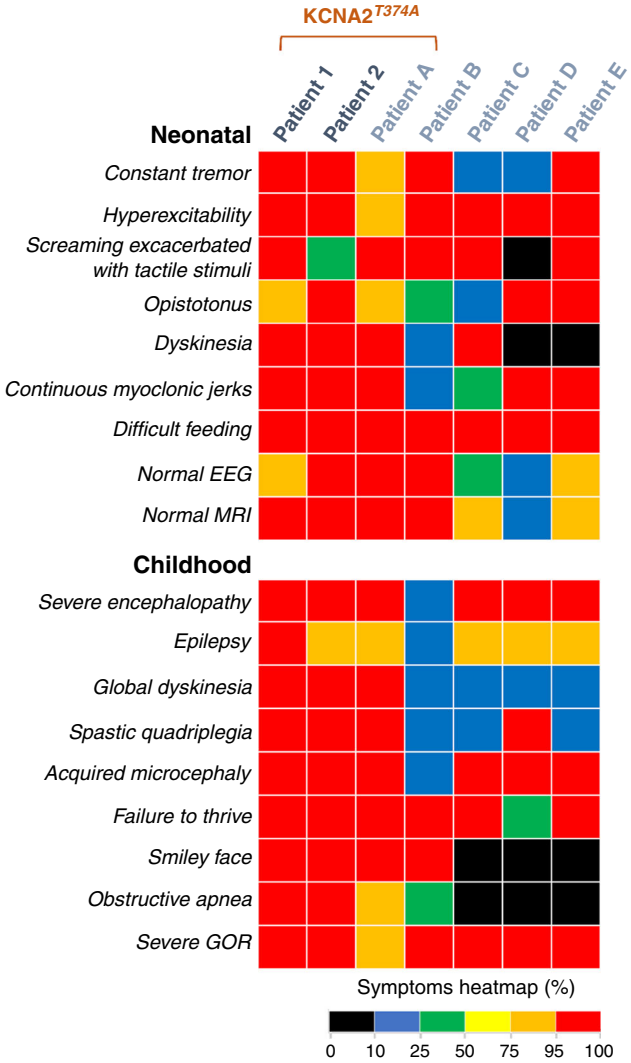

Fig. 2 Clinical heat map describing the detailed characteristics of the patients in this study. Heatmap for patient 1 and 2 with the p.T374A *KCNA2* variant as well as the other five patients (patients A–E) who were identified by the Dr. Warehouse database with the highest Similarity Index (SI). EEG electroencephalogram, MRI magnetic resonance image, GOR Gastro-oesophageal reflux.

DISCUSSION

Accurate assessment of symptoms is fundamental in the current context of both developing genetic screening technologies and bioinformatics data management, and should advance accurate diagnosis in tertiary medical centers taking in charge patients with rare and very rare diseases. The use of electronic systems able to process data exploiting narrative and documents is particularly suitable for the context of specialized centers for diagnosis and treatment of rare diseases. These systems allow the possibility to acquire information reported in a descriptive way, without structured a priori data acquisition that could induce missing rare and novel symptoms and signs. In parallel, increasing access to advanced genetic diagnostic techniques allows the continuous enrichment of known and novel genotype–phenotype spectra.

Dr. Warehouse, a data warehouse oriented toward narrative reports in use at our center, has already been shown to be reliable in extrapolating the phenotype of known conditions.[9,12] To validate this database usage for rare disorders, previous studies have been performed in our center. We evaluated this algorithm by identifying patients with similar clinical features for five different rare diseases: Lowe syndrome, dystrophic epidermolysis bullosa, activated PI3K delta syndrome, Rett syndrome, and Dowling–Meara. In this study, we showed how Dr. Warehouse integrated an efficient automated tool able to identify genetically similar patients using a VSM approach.[13] VSM was used to compute the similarity distance between an index patient and all the patients in the data warehouse. The dimensions of the VSM were built upon UMLS concepts extracted from clinical narratives stored in the clinical data warehouse. The VSM was enhanced using three parameters: a pertinence score (TF-IDF of the concepts), the polarity of the concept (negated/not negated), and the minimum number of concepts in common. It might help to detect patients with similar medical histories, diagnoses, and outcomes from a database including a large number of cases with automated methods.[9,13] Significantly, this database has been used for the first time to match two patients in our center carrying the same *KCNA2* variants with striking accuracy. In addition, this system is very efficacious to predict the longitudinal clinical outcome accurately assessing the major phenotypic concepts and their age of onset mimicking a natural history study among patients with genotype–phenotype correlation, as we demonstrate in patients with this specific *KCNA2* variant. More importantly, this report shows the possibility to accurately match patients sharing clinical features as well as to predict their natural history. This is of major interest in very rare diseases where every patient counts to understand the disease and to propose adapted trials for therapies. Indeed, in the absence of structured data, Dr. Warehouse showed the consistency of symptoms such as sleep disturbances and GOR as constant features of the disease as suggested by animal models carrying *KCNA2* variants.[14–16] This tool is adapted to correlate patients with new variant findings with specific phenotype description in the context of complex rare diseases where interpretation of exomes or NGS analysis is sometimes limited when facing new variants or variant of unknown significance.[17]

Finally, Dr. Warehouse is capable of integrating medical data to predict the genetic component for rare and particularly challenging clinical phenotypes and is able to match with striking accuracy the natural history and clinical progression, emphasizing the utility of these tools for public health practitioners. In the near future, these databases and electronic records could be further optimized to define genetic features coupled with clinical outcome and evolution of patients to elaborate targeted and efficient therapeutic strategies.

This work has some limitations. The first is related to natural language processing. The similarity calculation is largely dependent on the quality of the phenotypic extraction. The quality depends on the completeness of the thesaurus used (UMLS) and on the ability to detect the attributes of the extracted phenotypes (i.e., polarity, experimenter, hypothesis). In the work presented, the algorithm detects negation and family history but not the notion of hypothesis or nonreality. On the other hand, we use the VSM to calculate similarity between patients. This VSM method assumes that the concepts are statistically independent, which is rarely the case in a clinical context. Despite these methodological limitations, we describe interesting results for genotype–phenotype linkage and we are confident that this methodology can only be improved in future work.

Electronic medical files have a widespread use in many developed countries and are informed by different experts and nonexpert clinicians. Algorithms capable of data mining in these unstructured data are a “cheap,” transposable, and sustainable methodology to search for patient(s) presenting a high similitude of phenotypic characteristics. This algorithm, when a new causal variant for a disease is discovered, is able to find potential cases in a retrospective database by looking for patients similar to the patients already diagnosed and genotyped and could thus reduce their diagnostic journey and confirm the causal association. This

report supports the importance of deep phenotyping and paves the way for the use of mining unstructured data systems in the field of rare diseases for the purpose of diagnosis, extending cohorts, and looking at the natural history of rare and common diseases.

DATA AVAILABILITY

Clinical data sets are available upon request to the corresponding author (R.N.) and can be shared after consulting our local ethics committee.

Received: 20 May 2020; Revised: 29 October 2020; Accepted: 29 October 2020;
Published online: 26 January 2021

REFERENCES

1. Jannot, A. S. et al. The Georges Pompidou University Hospital Clinical Data Warehouse: a 8-years follow-up experience. *Int. J. Med. Inform.* **102**, 21–28 (2017).
2. Garcelon, N. et al. Improving a full-text search engine: the importance of negation detection and family history context to identify cases in a biomedical data warehouse. *J. Am. Med. Informatics Assoc.* **80**, 52–63 (2017).
3. Hardies, K., Weckhuysen, S., De Jonghe, P. & Suls, A. Lessons learned from gene identification studies in Mendelian epilepsy disorders. *Eur. J. Hum. Genet.* **24**, 961–967 (2016).
4. McTague, A. et al. The genetic landscape of the epileptic encephalopathies of infancy and childhood. *Lancet. Neurol.* **5**, 304–316 (2016).
5. Barcia, G. et al. Epilepsy with migrating focal seizures. *Neurol. Genet.* **5**, e363 (2019).
6. Pena, S. D. J. & Coimbra, R. L. M. Ataxia and myoclonic epilepsy due to a heterozygous new mutation in KCNA2: proposal for a new channelopathy. *Clin. Genet.* **87**, e1–e3 (2015).
7. Syrbe, S. et al. De novo loss-or gain-of-function mutations in KCNA2 cause epileptic encephalopathy. *Nat. Genet.* **47**, 393–399 (2015).
8. Masnada, S. et al. Clinical spectrum and genotype–phenotype associations of KCNA2-related encephalopathies. *Brain.* **140**, 2337–2354 (2017).
9. Garcelon, N. et al. A clinician friendly data warehouse oriented toward narrative reports: Dr Warehouse. *J. Biomed. Inform.* **80**, 52–63 (2018).
10. Hundallah, K., Alenizi, A., Alhashem, A. & Tabarki, B. Severe early-onset epileptic encephalopathy due to mutations in the KCNA2 gene: expansion of the genotypic and phenotypic spectrum. *Eur. J. Paediatr. Neurol.* **20**, 657–660 (2016).
11. Lindberg, D. A., Humphreys, B. L. & McCray, A. T. The Unified Medical Language System. *Methods Inf. Med.* **32**, 281–291 (1993).
12. Garcelon, N. et al. Next generation phenotyping using narrative reports in a rare disease clinical data warehouse. *Orphanet. J. Rare Dis.* **13**, 85 (2018).
13. Garcelon, N. et al. Finding patients using similarity measures in a rare diseases-oriented clinical data warehouse: Dr. Warehouse and the needle in the needle stack. *J. Biomed. Inform.* **73**, 51–61 (2017).
14. Brew, H. M. et al. Seizures and reduced life span in mice lacking the potassium channel subunit Kv1.2, but hypoexcitability and enlarged Kv1 currents in auditory neurons. *J. Neurophysiol.* **98**, 1501–1525 (2007).
15. Xie, G. et al. A new Kv1.2 channelopathy underlying cerebellar ataxia. *J. Biol. Chem.* **285**, 32160–32173 (2010).
16. Srdanović, S. et al. Transient knock-down of kcna2 reduces sleep in larval zebrafish. *Behav. Brain Res.* **326**, 13–21 (2017).
17. Richards, S. et al. Standards and guidelines for the interpretation of sequence variants: a joint consensus recommendation of the American College of Medical Genetics and Genomics and the Association for Molecular Pathology. *Genet. Med.* **17**, 405–424 (2015).

ACKNOWLEDGEMENTS

We thank the patients and their families; Diane Doummar and Bastien Estublier for providing some of the patients' clinical data; and Boris Keren and Claude Besmond, who analyzed the exome results of patients 1 and 2. The authors acknowledge the Bettencourt Foundation (R.N.) and the ERC Consolidator Grant (E.K.) for funding. This work was supported by funding from The French National Research Agency (ANR) under "Inverstissements d'Avenir" programs (reference ANR-10-IAHU-01).

AUTHOR CONTRIBUTIONS

R.N. and M.H. conceptualized the study and provided the methodology; M.H., T.L.B., C.C., I.D., and R.N. analyzed the clinical data; A.K., C.C., and R.N. analyzed the EEG data; M.H., T.L.B., R.N., and N.G. analyzed Dr. Warehouse concepts and data (software); G.B. and C.M. analyzed and interpreted the exome and Sanger sequencing; R.N. supervised the study; M.H., T.L.B., E.K., and R.N. drafted the first version of the manuscript; all authors participated to the writing, reviewing, and editing of the manuscript and agreed on the final version for submission; A.K. and R.N. provided Fig. 1; R.N., T.L.B., and E.K. conceptualized Fig. 2; R.N., E.K., and N.G. worked on validation. All authors participated to the first revision and agreed on the revised manuscript submission (based on the CRediT Contributor Roles Taxonomy categories).

ETHICS DECLARATION

This research was approved by the ethics committee of our institution (Necker Enfants Malades Hospital) and informed consent was obtained from the parents of patients accordingly.

COMPETING INTERESTS

The authors declare no competing interests.

ADDITIONAL INFORMATION

Supplementary information The online version of this article (https://doi.org/10.1038/s41436-020-01039-z) contains supplementary material, which is available to authorized users.

Correspondence and requests for materials should be addressed to R.N.

Reprints and permission information is available at http://www.nature.com/reprints

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.