Automated detection of genetic etiology prior to diagnosis using electronic medical records

Peter D. Galer [1,2,3,4], Shiva Ganesan [1,2,3], Michael Kaufman [1,2,3], Shridhar Parthasarathy [1,2,3], Sarah Ruggiero [1,3], Stacey Cohen [1,3], Olivia Wilmarth [1,3], and Ingo Helbig [1,2,3,5]

[1] Division of Neurology, Children's Hospital of Philadelphia. [2] Department of Biomedical and Health Informatics (DBHi). [3] Epilepsy Genetics Initiative (ENGIN), Children's Hospital of Philadelphia. [4] Department of Bioengineering, University of Pennsylvania. [5] Department of Neurology, University of Pennsylvania, Perelman School of Medicine.

Introduction

- Developmental epileptic encephalopathies (DEE) are a rare, heterogenous set of disorders thought to be largely genetic in origin.
- A genetic etiology can be invaluable to treatment, but many patients go years before a diagnosis is made. Key to this diagnosis is phenotypic data.
- We propose a novel pipeline to automate extraction of phenome data from electronic medical records (EMR) and consequently detect key features prior to genetic diagnosis.

Methods

- We used the NLP pipeline cTAKES to extract human phenotype ontology (HPO) terms from free-text of all notes available for each individual in the EMR.
- Terms are placed in 3-month time bins by age of individual at each EMR note.
- Terms of individuals with a particular genetic etiology prior to diagnoses are compared to the remainder of the DEE cohort.

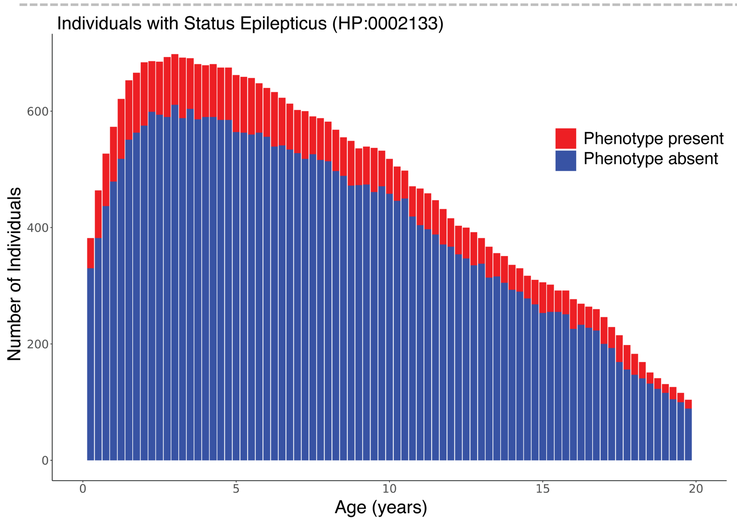

Figure 1. Status epilepticus (HP:0002133) prevalence in the cohort across 3-month age bins.

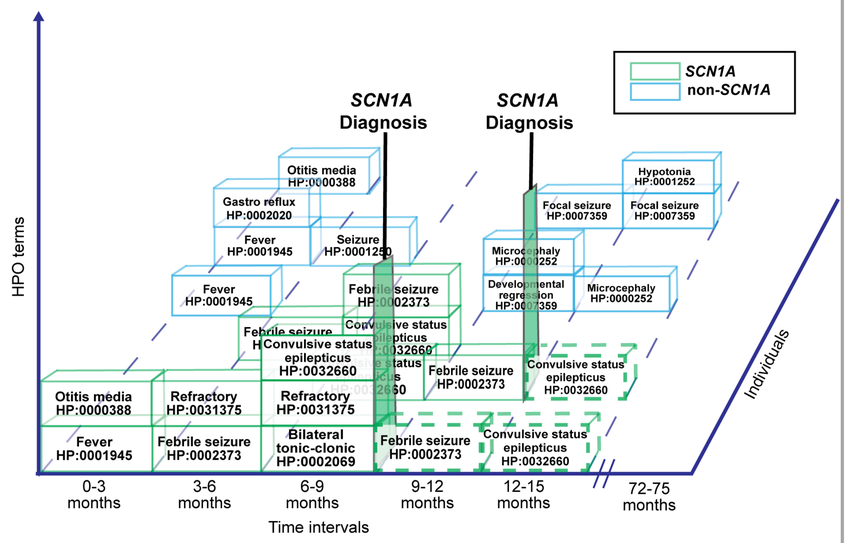

Figure 2. Parsing individuals phenotype terms into 3-month bins, separated by genetic etiology and age of genetic diagnosis.

Results

- We extracted 2,988,549 HPO terms from the EMR of 1,574 individuals with DEE with a confirmed or presumed genetic etiology placed, binned by age.
- 54,407 terms occur significantly more frequently in individuals with a particular genetic etiology prior to diagnosis in a particular age bin compared to the other individuals with DEE, including:
  - *SCN1A* – Focal clonic seizures (HP:0002266), 1.50-1.75 years (CI: 2.30-Inf; PPV=1)
  - *SCN1A* – Focal motor status epilepticus (HP:0032663), 0.50-0.75 years (CI: 4.44-5531.4; PPV=0.67)
  - *KCNB1* – Central sleep apnea (HP:0010536), 3.25-3.50 years (CI: 11.15-13790.3; PPV=0.29)
  - *IQSEC2* – Microcephaly (HP:0000252), 3.00-3.25 years (CI: 7.62-Inf; PPV=0.081)

Discussion

- Phenotypes can be automatically extracted from the EMR and grouped by genetic diagnosis on a longitudinal scale. From this, a phenotypic landscape of disease and genetic disorders can be formed.
- Significant phenotype signatures of genetic disorders can be isolated prior to diagnosis.
- Using these isolated signatures, in a completely automated fashion prediction scores of genetic etiologies for an individual can be generated from phenotypes extracted from free-text from an individual's EMR.
- This paves the way for improved genetic testing and automated diagnostics in individuals with DEE.

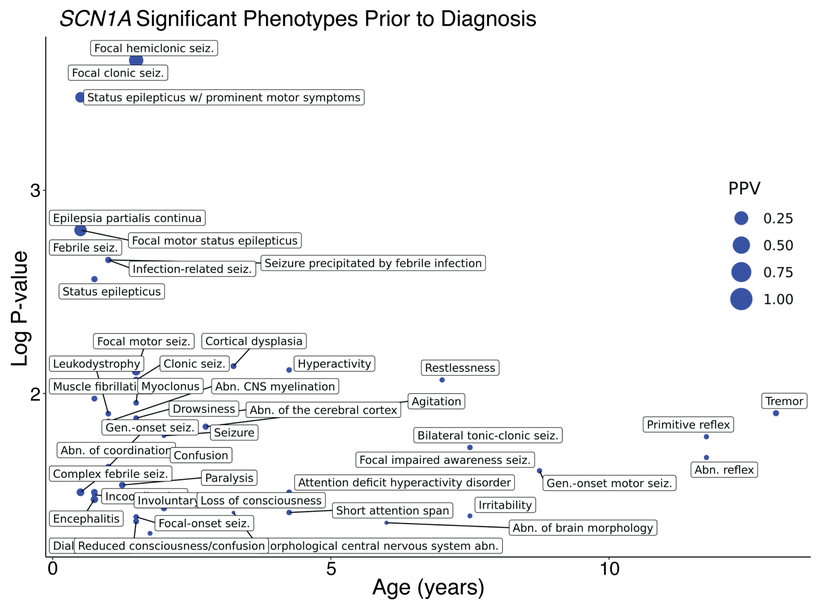

Figure 2. Significant phenotypes prior to genetic diagnosis in DEE patients with a *de novo* mutation in *SCN1A*. Size of dots indicate positive predictive value (PPV) of the phenotype.