Visualizing Clinical Profiles of Rare Metabolic Diseases

Project Team: Zhong Huang, Nishant Iyengar
Project Manager: Zach White
Project Lead: Rachel Richesson, PhD

- Project Summary
  - Two undergraduate students spent ten-weeks adopting Latent Dirichlet Allocation, a natural language processing technique, as a clustering mechanism for the comorbidities of rare metabolic diseases.

- Data
  - 1.2 million DUHS patients over the past 5 years; all have been diagnosed with a “rare disease.”
  - Includes ICD 9 codes, Event Date, Age, Sex, and Medications (however only ICD 9 codes and Medications were used to cluster patients).

Duke University School of Nursing
Duke
SOCIAL SCIENCE RESEARCH INSTITUTE
Duke University School of Medicine

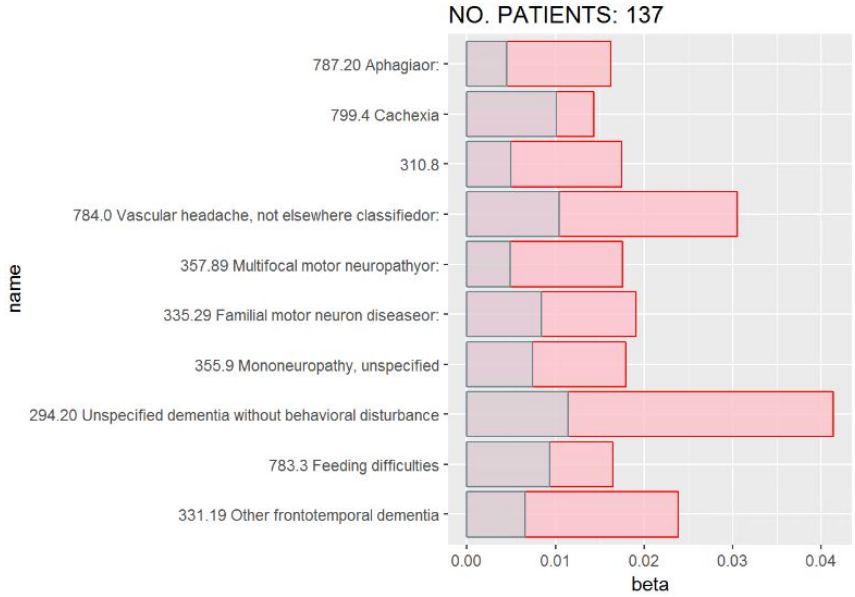
Methodology & Model Validation

- Methodology:
  - After parsing out diseases populations from the larger data set, a long-format table was constructed that tallied the number of times a patient ID was associated with an ICD9 code.
  - Patients with <5 unique ICD9 codes and ICD9 codes appearing in >50% of all patients were discarded. The table also underwent TF-IDF down weighing, analogous to down weighing “stop words” in natural language processing.
  - The table was then fed into the Latent Dirichlet Model, implemented in the topicmodels package.

- Model Validation:
  - The K value (number of topics/clusters) was selected using the ldatuning package, which integrates metrics from Griffiths (2004), CaoJuan (2009), Arun (2010), and Deveaud (2014).
  - The K value was limited between 2 and 6 for practicality.
  - Only disease populations that could be modeled with a p-value of (approximately) 0 were selected for continued analysis.

Results / Next Steps

The resultant model produces an optimal K clusters and the top ten most prevalent symptoms/medications in each cluster. Using Principal Component Analysis, it also visualizes the difference between different clusters.



- Next Steps
  - Invite more clinicians at DUHS to use our custom R Shiny interface to manually down weigh ICD9 codes that are irrelevant in a cluster (i.e. downweigh clinically known comorbidities).
  - Determine novel correlations between the comorbidities of rare metabolic diseases.
  - Conduct a more in-depth statistical analysis within clusters.