Towards Patient-Driven Phenotyping and Similarity for Precision Medicine

Tiffany J. Callahan
Computational Biosciences Program
University of Colorado Anschutz Medical Campus

Olivier Bodenreider
Lister Hill National Center for Biomedical Communications
National Library of Medicine
Electronic Medical Records

- Digital version of a patient’s medical history:
  - Inpatient notes
  - Labs and physical exams
  - Prescribed medications
  - Diagnoses and procedures
  - Treatment plans
  - Discharge instructions

The big promise of lies in large-scale use, automatically feeding clinical research, quality improvement, and clinical phenotyping.

Hripcsak & Albers, JAMIA. 2013
Clinical Phenotyping

**Goal:** identify cohorts of patients with specific clinical features characteristic of a disease of interest

**Typical approaches:**

- **Rule-based**  
  + interpretable, fast to implement, good results on limited datasets  
  - requires expert knowledge and multiple iterations, not easily generalizable

- **Natural language processing and text mining**  
  + rich data not found in other sources  
  - sensitive to misspelling/bad grammar, redundancy, ambiguity

- **Machine learning**  
  + many standardized approaches, easy implementation, robust  
  - curse of dimensionality; difficult with rare disease/small patient cohorts
Motivation

• Traditional approaches are good at providing information on the "average patient"

• What evidence can physicians use when trying to treat a patient whose symptoms deviate from average?

• **Patient Similarity**: derive insights from patients that are similar to an index patient to provide personalized predictions

• Diagnostic cohort identification
  • Drug repurposing
  • Identify and tailor treatment recommendations

1Sharafoddini et al. JMIR Med Inform. 2017
1. Similarity function
   - Data-driven; automatic
   - Pediatric data - OMOP CDM v5

2. Clustering
   - Similarity function-driven

3. Cluster identification/labeling
   - Clinical terminologies/value sets
   - Biomedical Knowledgebase
   - Literature

4. Evaluation
   - Compare to PheKB clusters
   - Verify algorithm reproducibility across data warehouses
Concept Normalization

Patient_similarity = \[0.0 + 1.0 + 0.2 + 0.389 + 0.456 + 0.027\]

\[\frac{1}{5} = 0.2\]

\[\text{sim}(t_1, t_2) = -\log_2 \frac{|T(t_1) \cap T(t_2)|}{|T(t_1) \cup T(t_2)|}\]

Evaluation

- Children’s Hospital of Colorado EHR data
  - De-identified (COMIRB # 15-0445)
- PEDSnet OMOP version 5
  - Concepts normalized to standardized terminologies
- Test Case – 2 groups (N = 20)
  - Huntington’s Chorea (ICD-9-CM 333.4)
  - Cystic Fibrosis (ICD-9-CM 722.0)

```sql
SELECT person_id, condition_source_value, COUNT(condition_source_value) AS count
FROM omop5deid.condition_occurrence
WHERE condition_source_value LIKE '333.4 %'
GROUP BY person_id, condition_source_value
ORDER BY count DESC
LIMIT 10;
```
### Huntington's Chorea
- Fatal disorder caused by breakdown of nerve cells in the brain
- 30,000 Americans have been diagnosed
- Symptoms include:
  - Personality, mood changes
  - Unsteadiness, poor coordination
- Diagnoses ($\overline{x} = 320.2; \text{unique} = 334$)
- Laboratory tests ($\overline{x} = 114.1, \text{unique} = 119$)
- Medications ($\overline{x} = 1528.7, \text{unique} = 177$)

### Cystic Fibrosis
- Genetic disease that causes mucus buildup resulting in persistent lung infection and difficulty breathing
- >30,000 people diagnosed worldwide
- Symptoms include:
  - Coughing, wheezing, frequent lung infections
  - Poor growth, male infertility
- Diagnoses ($\overline{x} = 982.5, \text{unique} = 447$)
- Laboratory tests ($\overline{x} = 3104.4, \text{unique} = 124$)
- Medications ($\overline{x} = 3120.3, \text{unique} = 392$)
Clustering

- Convert pairwise patient similarity to distance matrix
- Agglomerative hierarchical clustering with complete linkage
- 3 Clusters
  - Cystic Fibrosis
  - Huntington’s Chorea – red
  - Huntington’s Chorea – yellow
    - Pulmonary fibrosis (3; top 4 frequent dx)
    - Asthma (2; top 10 frequent dx)
Conclusions

- Developed a patient similarity algorithm
  - Data-driven
  - Composite semantic similarity for heterogeneous data types
  - Adjust weights to customize by use case
  - Promising initial proof of concept with pediatric EHR is promising

- Limitations
  - Small test group, need to scale to larger groups
  - Limited evaluation
  - Several unmapped Generic Product Identifiers

- Future Work
  - Explore alternative semantic similarity algorithms
  - Optimize algorithm
  - Develop machine learning approach to determine patient similarity attribute weights
Acknowledgments

- Dr. Olivier Bodenreider
- Drs. Paul Fontelo and Clement McDonald
- Summer Follows: Ann Cirincione and Raja Cholan
- Dr. Michael Kahn and Health Data Compass Team