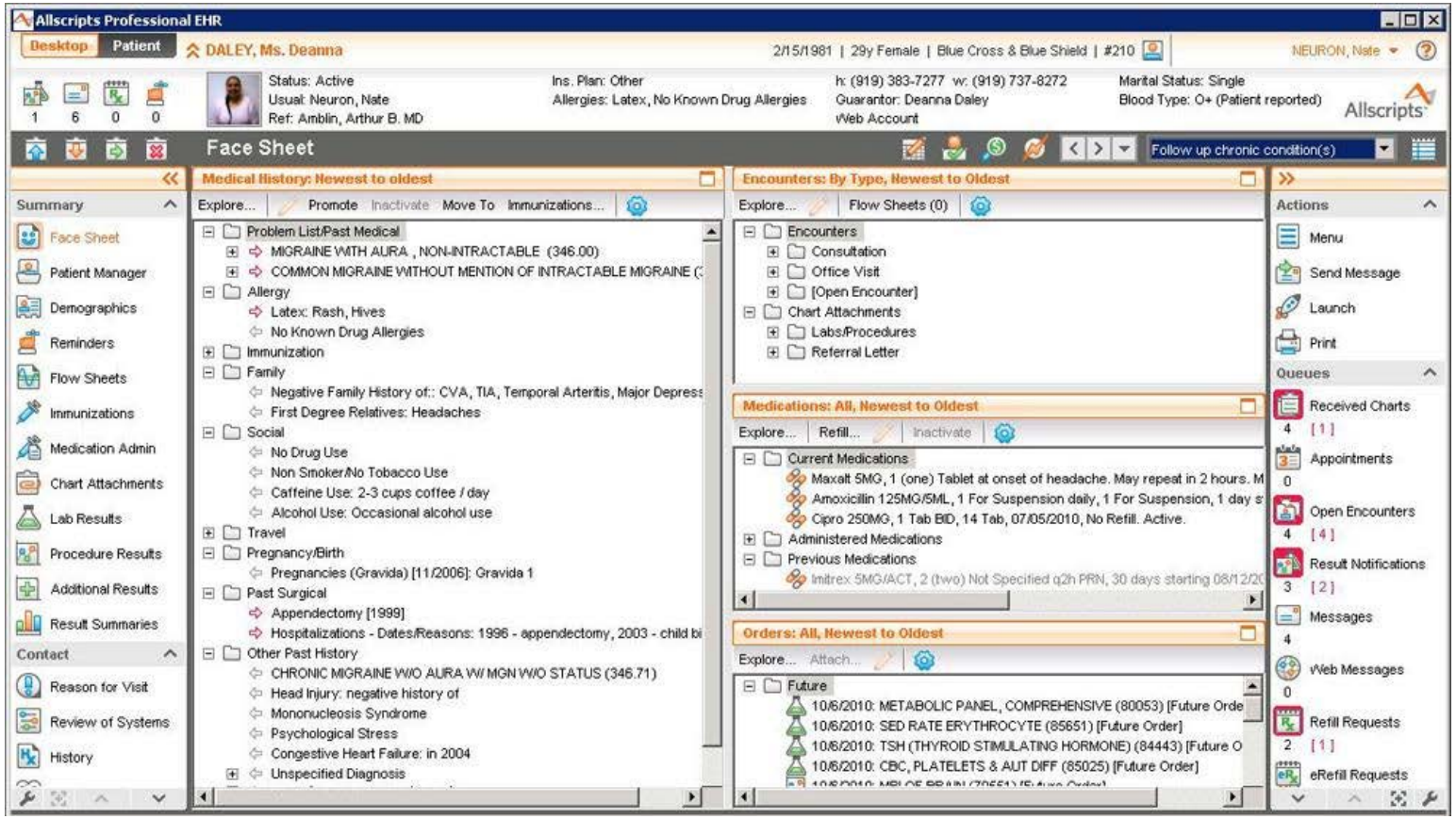
Towards Patient-Driven Phenotyping and Similarity for Precision Medicine

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.

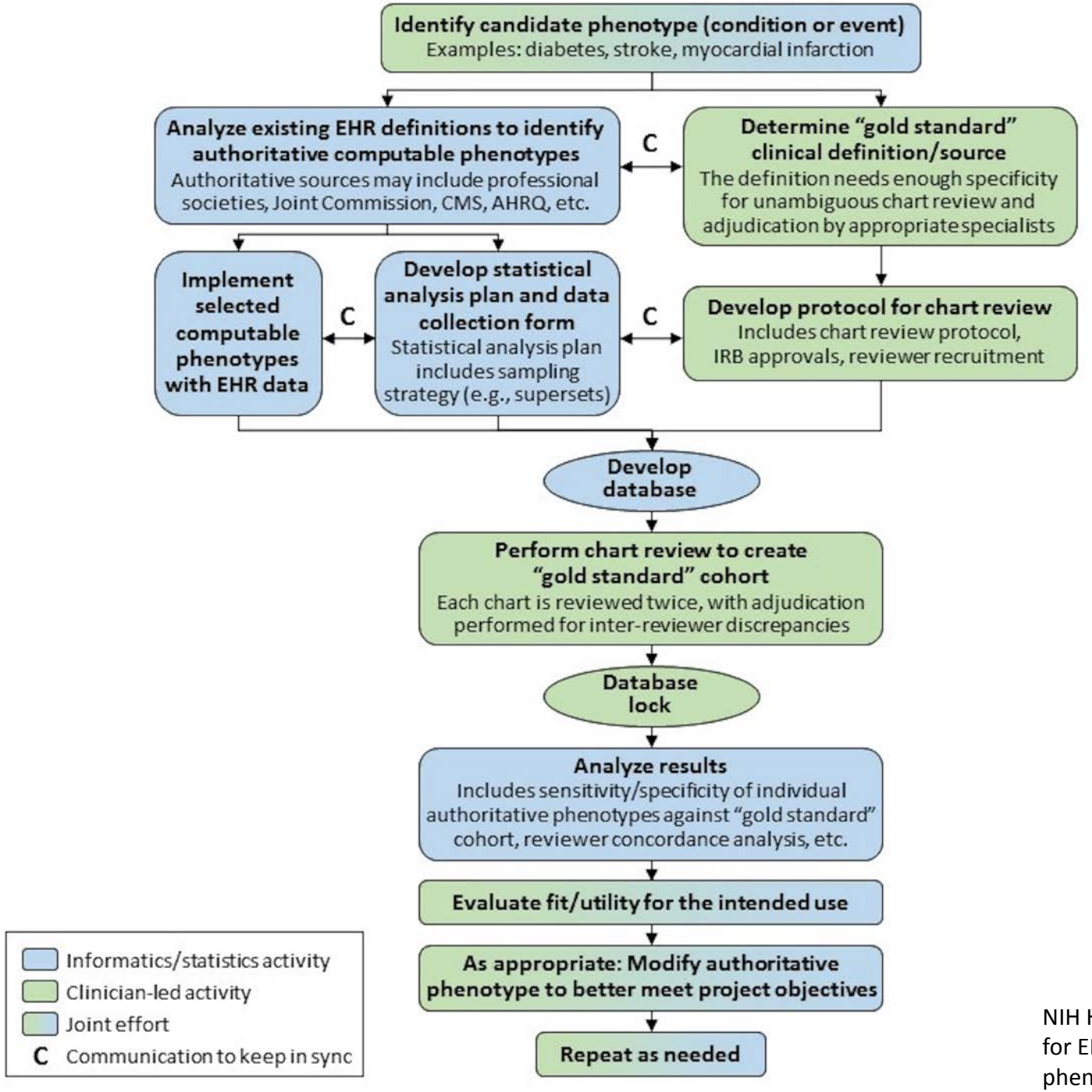
Hripcsak & Albers, *JAMIA*. 2013

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

PheKB

eMERGE Network
electronic medical records & genomics

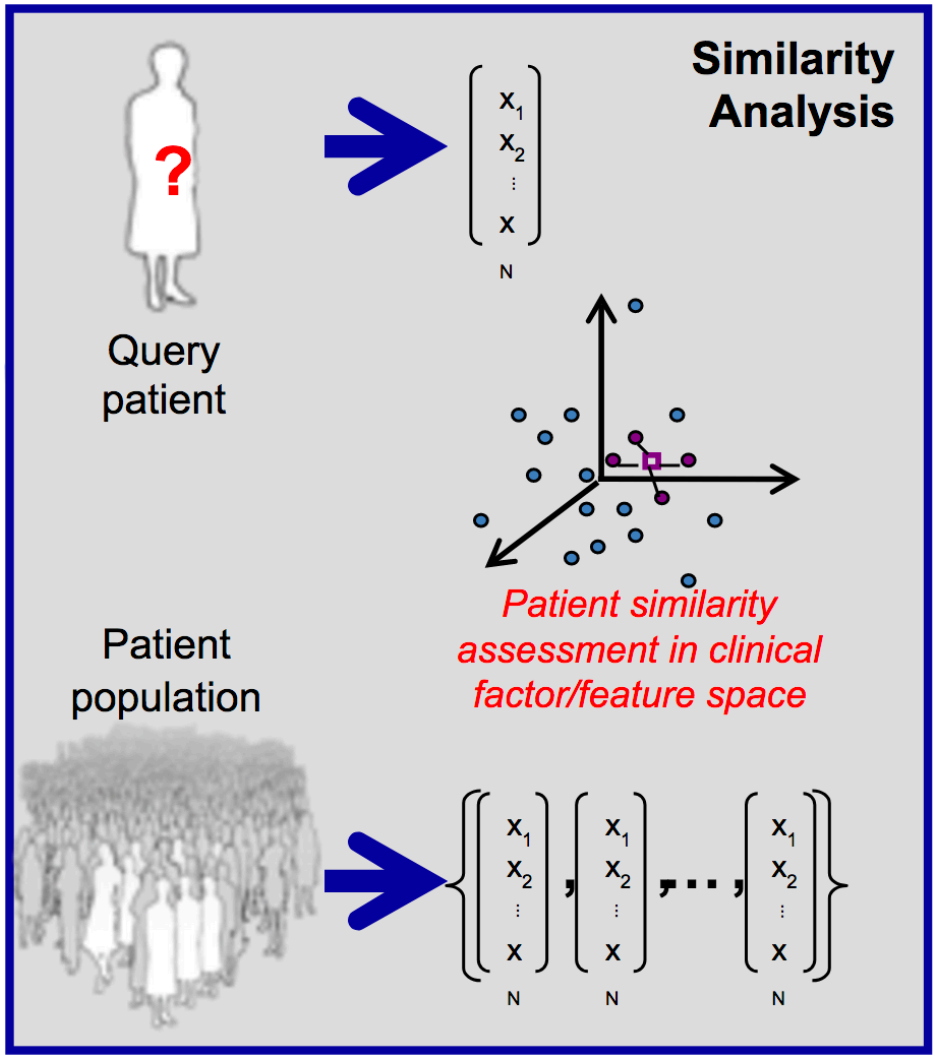


NIH Health Care Systems Research Collaboratory. Tools for EHR-based phenotyping. Type 2 diabetes mellitus phenotype definitions (Version 1.0). 2014.

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[1]
- Diagnostic cohort identification
    - Drug repurposing
    - Identify and tailor treatment recommendations



[1]*Sharafoddini et al. JMIR Med Inform. 2017*

Approach

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

Labs
Medications
Diagnoses
Demographics

MRN	Labs	Diagnosis	Medication
12345	Abs Cd4+ T Lymp: 1093 cells/mm^3	Age-related macular degeneration	0.05 ML Ranibizumab 10 Mg/ML injection

MRN	Age	Gender	Race
12345	14.0	Male	White

Patients | **Clinical Concepts**

Male != Female | White == White | $1 - \frac{(|14 - 30|)}{20}$

0.0 | 1.0 | 0.20

Concept Normalization

LOINC: | SNOM

gender, race, age | conditions, labs, medications

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

MxN matrix

ce conditions labs meds

ce conditions labs meds

stors

$\frac{1}{5} = 0.2$

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

81423003

86555001

277869007

190909002

57676002

0.02 0.01 0.08 0.13 0.13 0.13 0.01 0.04 0.07 0.04 0.02 0.01 0.02 0.02

1.0 + 0.13 + 0.07 = 1.193

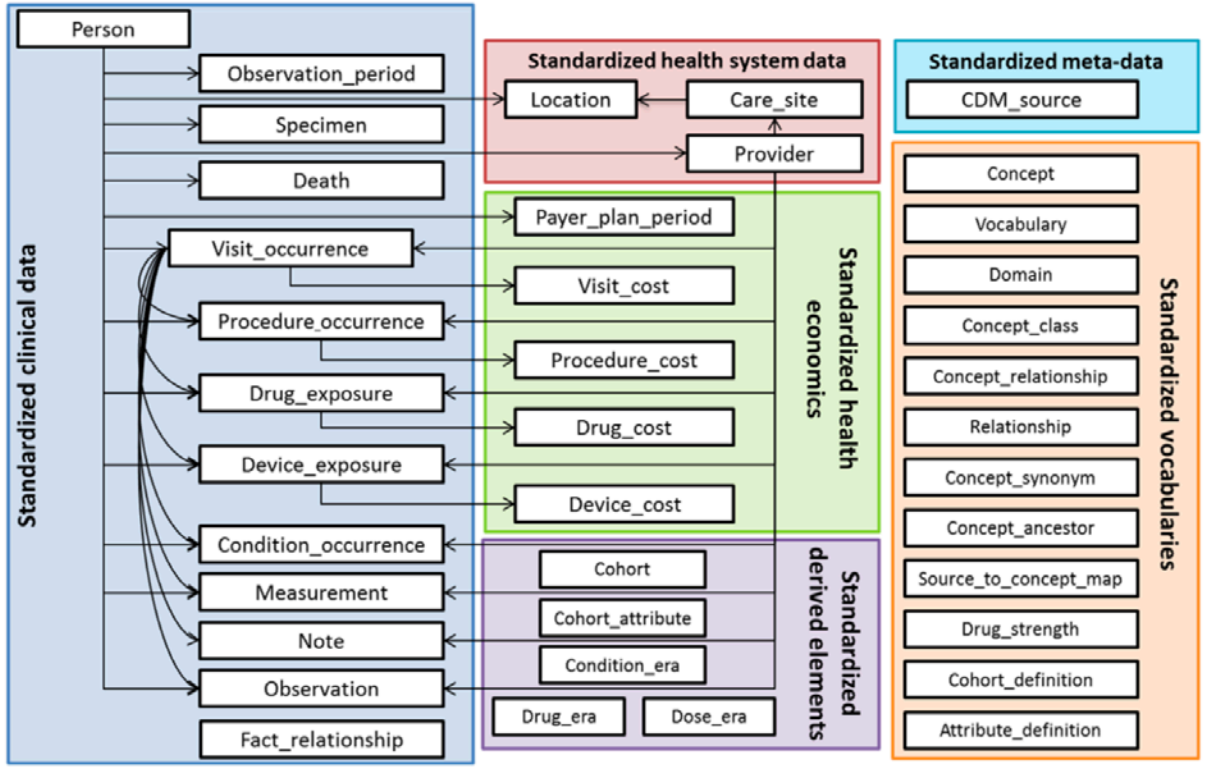
1.0 + 0.13 + 0.02 = 1.144

$(1.193 + 1.144)x \frac{1}{3+3} = 0.389$

[1]Batet et al. *J Biomed Inform.* 2011; [2]Azuaje et al. *Proc ISMB'2005 SIG. 2005*

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)



Google BigQuery

```
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
;
```

Row	person_id	condition_source_value	count
1	151130	333.4 | Huntington's chorea	125
2	322136	333.4 | Huntington's chorea	47
3	456778	333.4 | Huntington's chorea	21
4	65447	333.4 | Huntington's chorea	19
5	53666	333.4 | Huntington's chorea	16
6	434388	333.4 | Huntington's chorea	15

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 ($\bar{x}$ = 320.2; unique = 334)
- Laboratory tests ($\bar{x}$ = 114.1, unique = 119)
- Medications ($\bar{x}$ = 1528.7, unique = 177)

Hdsa.org/what-is-hd/

Row	co_person_id	OMOP_concept	SNOMED_code	source_code	count
1	151130	374341	58756001	333.4 | Huntington's chorea	125
2	151130	435642	29164008	784.5 | Other speech disturbance	59
3	151130	441277	25766007	315.32 | Mixed receptive-expressive language disorder	57
4	151130	441417	281016006	781.3 | Lack of coordination	40
5	151130	440086	192127007	314.00 | Attention deficit disorder without mention of hyperactivity	38
6	151130	440374	191736004	300.3 | Obsessive-compulsive disorders	22
7	151130	442077	197480006	300.00 | Anxiety state, unspecified	17
8	151130	436077	248290002	315.9 | Unspecified delay in development	17
9	151130	438409	406506008	314.9 | Unspecified hyperkinetic syndrome	16
10	151130	436073	69322001	298.9 | Unspecified psychosis	16
11	151130	379782	5158005	307.23 | Tourette's disorder	14

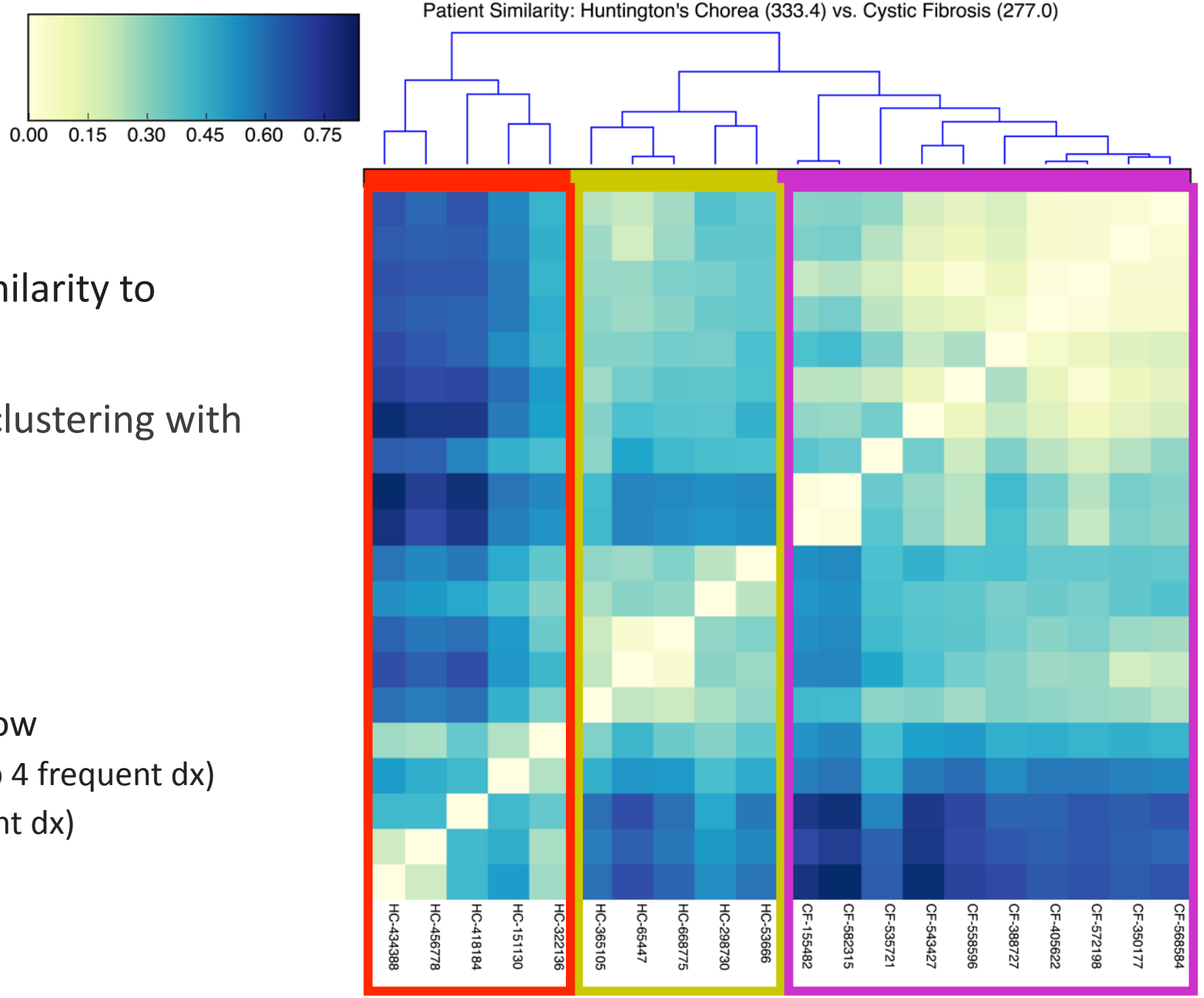
Cystic Fibrosis

- 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 ($\bar{x}$ = 982.5, unique = 447)
- Laboratory tests ($\bar{x}$ = 3104.4, unique = 124)
- Medications ($\bar{x}$ = 3120.3, unique = 392)

https://www.cff.org/What-is-CF/About-Cystic-Fibrosis/

Row	co_person_id	OMOP_concept	SNOMED_code	source_code	count
1	388727	434615	81423003	277.00 | Cystic fibrosis without mention of meconium ileus	538
2	388727	254320	86555001	277.02 | Cystic fibrosis with pulmonary manifestations	406
3	388727	378253	25064002	784.0 | Headache	352
4	388727	442077	197480006	300.00 | Anxiety state, unspecified	343
5	388727	194325	190909002	277.03 | Cystic fibrosis with gastrointestinal manifestations	272
6	388727	436096	82423001	338.29 | Other chronic pain	215
7	388727	4174281	277869007	031.0 | Pulmonary diseases due to other mycobacteria	194
8	388727	377889	15188001	389.9 | Unspecified hearing loss	182
9	388727	77074	57676002	719.40 | Pain in joint, site unspecified	150
10	388727	317009	195967001	493.90 | Asthma, unspecified type, unspecified	111
11	388727	4032376	108305003	V57.1 | Care involving other physical therapy	95

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