Office of the National Coordinator for Health Information Technology (ONC)

• **Mission**
  • Improve the health and well-being of individuals and communities through the use of technology and health information that is accessible when and where it matters most

• **Strategic Goals**
  • Advance Person-Centered and Self-Managed Health
  • Transform Health Care Delivery and Community Health
  • **Foster Research, Scientific Knowledge, and Innovation**
  • Enhance Nation's Health IT Infrastructure

https://healthit.gov/
Patient-Centered Outcomes Research (PCOR)

- Produce evidence to inform health care decisions made by patients, families, and their health care providers
- Patient-Centered Outcomes Research Trust Fund (PCORTF) managed by the Assistant Secretary for Planning and Evaluation (ASPE)

ONC Synthetic Health Data Project

Accelerate ability to conduct PCOR by:

• Enhancing an open-source synthetic data generator Synthea™, developed by The MITRE Corporation, to increase the number and variety of synthetic data
  • Opioid use
  • Pediatric populations
  • Patients with complex care needs

• Engaging broader community to validate the realism and demonstrate the potential uses of newly available synthetic data

Prize competition invited a wide array of innovators, researchers, and technology developers to create and test innovative solutions that enhance Synthea and the synthetic data it generates

• Advance novel uses of synthetic data for patient-centered outcomes research

• Validate the realism of Synthea-generated synthetic data
Challenge Structure

Phase I: Proposals for Innovative Models

- Written proposal describing proposed solution
- Proposals were invited from teams or individuals
- Proposals had to include methodology and intended outcomes

Phase II: Prototype/Solution Development

- Approved proposals moved on to Phase II
- Solutions designed and tested
- Final paper describing the solution
- Video demonstration
- Evidence of validation
- Non-proprietary source code

Winning Solutions

- Total cash prize pool: $100,000
- Solutions were judged by a panel based on criteria
Winning Solutions

**FIRST PLACE**

**CodeRx**

Medication Diversification Tool

$40,000

**SECOND PLACE**

**The Generalistas**

Virtual Generalist: Modeling Co-morbidities in Synthea

**Team LMI**

On Improving Realism of Disease Modules in Synthea: Social Determinant-Based Enhancements to Conditional Transition Logic

$15,000

**THIRD PLACE**

**Particle Health**

The Necessity of Realistic Synthetic Health Data Development Environments

**Team TeMa**

Empirical Inference of Underlying Condition Probabilities Using Synthea-Generated Synthetic Health Data

**UI Health**

Spatiotemporal Big Data Analysis of Opioid Epidemic in Illinois

$10,000
Medication Diversification Tool (MDT)

Team CodeRx
### Challenge Category and Use Case

<table>
<thead>
<tr>
<th>Challenge Category 1</th>
<th>Enhancement to Synthea</th>
</tr>
</thead>
<tbody>
<tr>
<td>Use Case</td>
<td>Pediatrics (pediatric asthma)</td>
</tr>
</tbody>
</table>

Source: https://www.istockphoto.com/vector/asthmatic-girl-gm185960170-27680137
Who we are

CodeRx is a collective of pharmacists and other healthcare professionals who have a skill set in tech and apply it towards building useful things

Website: coderx.io

Founded: early 2020

Membership: 150+ (30-40 active weekly). mostly pharmacists and pharmacy students

What we do

- **Slack channel** - engage in discussions about coding / data / tech as it relates to pharmacy and healthcare
- **GitHub organization** - collaborate on open source pharm tech / health tech projects
- **Newsletter and website** - share guides, resources, and pertinent topics
# Team CodeRx

For this challenge, Team CodeRx consists of six PharmDs from across the U.S.

<table>
<thead>
<tr>
<th>Name</th>
<th>Title/Role</th>
<th>Organization/Salary Unit</th>
</tr>
</thead>
<tbody>
<tr>
<td>Joseph LeGrand, PharmD, MS (team lead)</td>
<td>Lead Application Analyst</td>
<td>Vanderbilt University Medical Center</td>
</tr>
<tr>
<td>Kent Bridgeman, PharmD, MHI</td>
<td>Informatics Pharmacist</td>
<td>Allina Health</td>
</tr>
<tr>
<td>Kristen Tokunaga, PharmD, BCGP</td>
<td>Analytics Consulting Manager</td>
<td>Komodo Health</td>
</tr>
<tr>
<td>Yevgeny Bulochnik, PharmD, ACE, CACP</td>
<td>Formulary Administration Consultant</td>
<td>HealthPartners</td>
</tr>
<tr>
<td>Robert Hodges, PharmD, MSDS, MBA</td>
<td>Sr. Data Scientist, McKesson RelayHealth (CoverMyMeds)</td>
<td></td>
</tr>
<tr>
<td>Dalton Fabian, PharmD</td>
<td>Data Scientist</td>
<td>UnityPoint Health</td>
</tr>
</tbody>
</table>
Objective

To programmatically generate Synthea medication orders through the use of RxClass and Medical Expenditure Panel Survey (MEPS) data sources
In the current Synthea asthma module, 100% of Asthma patients are prescribed the same asthma maintenance inhaler (Flovent HFA 44 mcg) regardless of age.

This is not consistent with real-world clinical practice.

Source: https://synthetichealth.github.io/module-builder/#asthma
Asthma Maintenance Inhalers on the Market

Pediatric patients are prescribed one of many inhaled corticosteroids as a first line therapy.

Synthea vs MEPS

Medical Expenditure Panel Survey (MEPS) is a nationwide set of surveys of households and their medical providers.

The AHRQ conducts this survey annually to collect information on the use and cost of health care.
Tools

Open-source tools & code

Packages
- Python
- Pandas

Sources
- RxNav API
- RxClass API
- MEPS

Publicly-available, government-maintained data sources
Available Online GUIs

Already exist - maintained and hosted by National Library of Medicine
Developer Inputs

**Required**

1. RxClass class ID(s) OR RxNorm ingredient ID(s)

**Optional**

1. Dose form filters
2. Patient demographic info breakpoints
   a. Age range(s)
   b. Gender M/F
   c. State of residence
3. Single vs multi ingredient drugs
4. Other Synthea settings
Methods

**RxClass**
Medication class -> ingredient(s)

**RxNav**
Medication ingredient -> product(s) -> NDC(s)  
+ dose form filters  
+ single vs multi ingredient filters

(RxNorm)

**MEPS**
NDCs and demographic info and counts of patients who report taking them
Use RxClass API to return list of medication ingredients

- "Corticosteroids" = ATC
- Class R01AD

Medication ingredients

- beclomethasone
- betamethasone
- budesonide
- ciclesonide
- flunisolide
- fluticasone
- mometasone
- prednisolone
- tixocortol
- triamcinolone
Filter on RxNorm dose form in:
- Metered Dose Inhaler
- Dry Powder Inhaler
- Inhalation Suspension
- Single ingredient products only
- RxNorm term type (TTY) = IN

<table>
<thead>
<tr>
<th>Medication ingredients</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>beclomethasone</td>
<td>YES</td>
</tr>
<tr>
<td>betamethasone</td>
<td>NO</td>
</tr>
<tr>
<td>budesonide</td>
<td>YES</td>
</tr>
<tr>
<td>ciclesonide</td>
<td>YES</td>
</tr>
<tr>
<td>flunisolide</td>
<td>YES</td>
</tr>
<tr>
<td>fluticasone</td>
<td>YES</td>
</tr>
<tr>
<td>mometasone</td>
<td>YES</td>
</tr>
<tr>
<td>prednisolone</td>
<td>NO</td>
</tr>
<tr>
<td>tixocortol</td>
<td>NO</td>
</tr>
<tr>
<td>triamcinolone</td>
<td>NO</td>
</tr>
</tbody>
</table>
Cross-reference population prescription utilization data

- 0-5 year old age range

NOTE: for the actual submodule, we would also include the 6-103 year old age range

<table>
<thead>
<tr>
<th>Medication Ingredients</th>
<th>Percentage</th>
</tr>
</thead>
<tbody>
<tr>
<td>beclomethasone</td>
<td>3%</td>
</tr>
<tr>
<td>budesonide</td>
<td>48%</td>
</tr>
<tr>
<td>ciclesonide</td>
<td>0%</td>
</tr>
<tr>
<td>flunisolide</td>
<td>0%</td>
</tr>
<tr>
<td>fluticasone</td>
<td>49%</td>
</tr>
<tr>
<td>mometasone</td>
<td>0%</td>
</tr>
</tbody>
</table>

N/A - Not applicable
Asthma module JSON:

```
... "Maintenance_Medication_End": {
  "type": "MedicationEnd",
  "referenced_by_attribute": "maintenance_inhaler",
  "direct_transition": "Emergency_Medication_End"
}, ...
```

1. Place MDT module JSON file in the medications folder
2. Place MDT lookup table CSV files in the lookup_tables folder
3. Replace existing `MedicationOrder` state in asthma module JSON file with a `CallSubmodule` state referencing the MDT module
4. Ensure asthma module `MedicationEnd` states end medications by attribute, not by name
Results
Asthma Medications using the MDT - Ingredient

```
"Ingredients in lookup table:",
"# [ % pop ] Name",
"------------------",
"1. [ 3.1% ] Beclomethasone",
"2. [ 47.7% ] Budesonide",
"3. [ 49.2% ] Fluticasone"
```

```json
"type": "simple",
"lookup_table_transition": [
  {
    "transition": "Prescribe_Beclomethasone",
    "default_probability": "1",
    "lookup_table_name": "maintenance_inhaler_ingredient_distribution.csv"
  },
  {
    "transition": "Prescribe_Budesonide",
```
## Asthma Medications using the MDT - Product

<table>
<thead>
<tr>
<th>#</th>
<th>% pop</th>
<th>Name</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>29.5%</td>
<td>Budesonide_0.125_Mg_Ml_Inhalation_Suspension</td>
</tr>
<tr>
<td>2</td>
<td>15.1%</td>
<td>Budesonide_0.125_Mg_Ml_Inhalation_Suspension_Fulmicort</td>
</tr>
<tr>
<td>3</td>
<td>55.3%</td>
<td>Budesonide_0.25_Mg_Ml_Inhalation_Suspension</td>
</tr>
</tbody>
</table>

```json
{

type: "Simple",
lookup_table_transition: [
{
  transition: "Prescribe_Budesonide_0.125_Mg_Ml_Inhalation_Suspension",
  default_probability: "1",
  lookup_table_name: "maintenance_inhaler_Budesonide_product_distribution.csv"
},
{
  transition: "Prescribe_Budesonide_0.25_Mg_Ml_Inhalation_Suspension",
  default_probability: "0",
  lookup_table_name: "maintenance_inhaler_Budesonide_product_distribution.csv"
},
{
  transition: "Prescribe_Budesonide_0.125_Mg_Ml_Inhalation_Suspension_Fulmicort",
  default_probability: "0",
  lookup_table_name: "maintenance_inhaler_Budesonide_product_distribution.csv"
}]
```
Asthma Medications In Synthea Module Builder
Asthma Medications In Synthea Module Builder
## Validation

### Chi-Square Goodness-of-Fit Test

#### Current State

- 0%  100%  0%  0%  0%
  - $X^2 = 7.168, 52$
  - $df = 5$
  - $N = 14410$
  - $P = 0.00$
  - Significantly different

#### Gold Standard

- 16%  34%  7%  14%  26%
  - $X^2 = 2.73$
  - $df = 6$
  - $N = 13906$
  - $P = 0.184$
  - Not significantly different

#### Future State

- 15%  34%  8%  14%  26%
  - 3%
Benefits for Researchers

- Creates micro-validated medication distributions in Synthea modules
- Loosely test hypothesis prior to obtaining access to PHI
- Identify drug trends to validate on real data
Benefits for developers

- Encourages incorporation of complex drug treatment options into Synthea modules
- Offers complexity that developers will need to account for in software
- Allows drug related development before having PHI
Challenges & Successes

Challenges

- Complexity of data
- Difficulty in defining disease-specific drug lists

Successes

- Can be used with other Synthea modules & disease states
- Can generate medication distributions by patient age, gender, geographic location, and more
- Can be re-run with updated data
- Has flexibility for user to select drugs
- Designed by 6 Pharmacists who are experts on medication use
Future for MDT
Future Enhancements

The groundbreaking methods used in MDT to integrate public data sources allows for future enhancements:

1. Enhance Synthea with other MEPS data elements.
   a. insurance type, social determinants of health, medical conditions.
2. Model prescription fill and medication adherence through meps MEPS data.
3. Add allergy or drug-drug interactions logics through NLM\(^1\) data.
4. Capture dose ranges (low/medium/high) and progression.
5. Improve medication costs modeling through NADAC\(^2\) dataset.

\(^1\)NLM [National Library of Medicine]
\(^2\)NADAC [National Average Drug Acquisition Cost]
Thank You!!!!
Links

GitHub Repo: github.com/coderxio/medication-diversification

Joseph LeGrand, PharmD, MS linkedin.com/in/jrlegrand/

Kent Bridgeman, PharmD, MHI linkedin.com/in/kentvbridgeman/

Kristen Tokunaga, PharmD, BCGP linkedin.com/in/kristen-tokunaga-pharmd/

Robert Hodges, PharmD, MSDS, MBA linkedin.com/in/robhodgespharmd/

Dalton Fabian, PharmD linkedin.com/in/daltonfabian/

Yevgeny (Eugene) Bulochnik, PharmD ACE CACP linkedin.com/in/yevgeny-eugene-bulochnik-b429a6155/

CodeRx Website: coderx.io/
Outline

Bayes nets & CPTs
- comorbidities matter!
- use in Synthea’s new lookup_table_transition

Mapping concepts
- ICD10
- SNOMED
- Synthea attributes

Feature engineering
- SQL/Pyspark/R on Databricks
- Now demonstrated with Synthea data!

Validation
- co-occurrence matrix
- COPM

Suggested design patterns
- No delay: disease is NOT your destiny
- Mechanistic progression: allow interventions
- More modular modules
  - Separate treatment from disease incidence and progression
  - Use attributes to communicate between modules.
Bayes Nets decompose joint probabilities into a network of conditional probabilities

We trained a Bayes Net model on real hospital data ...

... and incorporated the results into a Synthea simulation module.
Each node in the Bayes Net contains a Conditional Probability Table.
### Mapping concepts

#### ICD10 & SNOMED attributes

<table>
<thead>
<tr>
<th>attribute</th>
<th>test</th>
<th>attribute_type</th>
<th>icd_pattern</th>
<th>icd_name</th>
<th>snomed_concept_name</th>
</tr>
</thead>
<tbody>
<tr>
<td>ckd_1</td>
<td>==1</td>
<td>integer</td>
<td>N18.1</td>
<td>Chronic_kidney_disease_stage_1</td>
<td>Chronic kidney disease stage 1 (disorder)</td>
</tr>
<tr>
<td>ckd_2</td>
<td>==2</td>
<td>integer</td>
<td>N18.2</td>
<td>Chronic_kidney_disease_stage_2</td>
<td>Chronic kidney disease stage 2 (disorder)</td>
</tr>
<tr>
<td>ckd_3</td>
<td>==3</td>
<td>integer</td>
<td>N18.3</td>
<td>Chronic_kidney_disease_stage_3</td>
<td>Chronic kidney disease stage 3 (disorder)</td>
</tr>
<tr>
<td>ckd_4</td>
<td>==4</td>
<td>integer</td>
<td>N18.4</td>
<td>Chronic_kidney_disease_stage_4</td>
<td>Chronic kidney disease stage 4 (disorder)</td>
</tr>
<tr>
<td>ckd_5</td>
<td>==5</td>
<td>integer</td>
<td>N18.[56]</td>
<td>Chronic_kidney_disease_stage_5</td>
<td>End stage renal disease (disorder)</td>
</tr>
<tr>
<td>smoker</td>
<td>is true</td>
<td>boolean</td>
<td>F17</td>
<td>Nicotine_dependence</td>
<td>Smokes tobacco daily</td>
</tr>
<tr>
<td>diabetes</td>
<td>is true</td>
<td>boolean</td>
<td>E11</td>
<td>Type_2_diabetes_mellitus</td>
<td>Diabetes</td>
</tr>
<tr>
<td>coronary_heart_disease</td>
<td>is true</td>
<td>boolean</td>
<td>I25</td>
<td>Chronic_ischemic_heart_disease</td>
<td>Coronary Heart Disease</td>
</tr>
<tr>
<td>copd_variant</td>
<td>is not nil</td>
<td>ConditionOnset</td>
<td>J44</td>
<td>Chronic_obstructive_pulmonary_disease</td>
<td>Chronic obstructive bronchitis (disorder)</td>
</tr>
</tbody>
</table>
Synthea-Specific ICD10 to SNOMED Mapping

<table>
<thead>
<tr>
<th>icd_set_id</th>
<th>snomed_concept_id</th>
<th>snomed_concept_name</th>
<th>icd_code</th>
<th>icd_description</th>
</tr>
</thead>
<tbody>
<tr>
<td>43</td>
<td>44054006</td>
<td>Diabetes</td>
<td>E11</td>
<td>Type 2 diabetes mellitus</td>
</tr>
<tr>
<td>446</td>
<td>46177005</td>
<td>End stage renal disease (disorder)</td>
<td>N18.5</td>
<td>Chronic kidney disease, stage 5</td>
</tr>
<tr>
<td>447</td>
<td>46177005</td>
<td>End stage renal disease (disorder)</td>
<td>N18.6</td>
<td>End stage renal disease</td>
</tr>
<tr>
<td>899</td>
<td>75498004</td>
<td>Acute bacterial sinusitis (disorder)</td>
<td>B96.89</td>
<td>Other specified bacterial agents as the cause of diseases classified elsewhere</td>
</tr>
<tr>
<td>899</td>
<td>75498004</td>
<td>Acute bacterial sinusitis (disorder)</td>
<td>J01.90</td>
<td>Acute sinusitis, unspecified</td>
</tr>
<tr>
<td>1323</td>
<td>301011002</td>
<td>Escherichia coli urinary tract infection</td>
<td>B96.20</td>
<td>Unspecified Escherichia coli [E. coli] as the cause of diseases classified elsewhere</td>
</tr>
<tr>
<td>1323</td>
<td>301011002</td>
<td>Escherichia coli urinary tract infection</td>
<td>N39.0</td>
<td>Urinary tract infection, site not specified</td>
</tr>
</tbody>
</table>

**Patterns**

E11 matches E11.0, E11.1, etc

**Sets**

All ICD elements of the set must be present for the SNOMED concept to apply
### Feature engineering

Now with Synthea data!

<table>
<thead>
<tr>
<th>patient</th>
<th>month_number</th>
<th>age_group</th>
<th>gender</th>
<th>race</th>
<th>ethnicity</th>
<th>ckd_stage</th>
<th>smoker</th>
</tr>
</thead>
<tbody>
<tr>
<td>726</td>
<td>003370e2-d54b-9d6b-c7e7-800c7f1144bf</td>
<td>515</td>
<td>age_45_64</td>
<td>F</td>
<td>white</td>
<td>ckd_3</td>
<td>F</td>
</tr>
<tr>
<td>727</td>
<td>003370e2-d54b-9d6b-c7e7-800c7f1144bf</td>
<td>516</td>
<td>age_45_64</td>
<td>F</td>
<td>white</td>
<td>ckd_3</td>
<td>F</td>
</tr>
<tr>
<td>728</td>
<td>003370e2-d54b-9d6b-c7e7-800c7f1144bf</td>
<td>517</td>
<td>age_45_64</td>
<td>F</td>
<td>white</td>
<td>ckd_3</td>
<td>F</td>
</tr>
<tr>
<td>729</td>
<td>003370e2-d54b-9d6b-c7e7-800c7f1144bf</td>
<td>518</td>
<td>age_45_64</td>
<td>F</td>
<td>white</td>
<td>ckd_3</td>
<td>F</td>
</tr>
<tr>
<td>730</td>
<td>003370e2-d54b-9d6b-c7e7-800c7f1144bf</td>
<td>519</td>
<td>age_45_64</td>
<td>F</td>
<td>white</td>
<td>ckd_3</td>
<td>F</td>
</tr>
<tr>
<td>731</td>
<td>003370e2-d54b-9d6b-c7e7-800c7f1144bf</td>
<td>520</td>
<td>age_45_64</td>
<td>F</td>
<td>white</td>
<td>ckd_3</td>
<td>F</td>
</tr>
<tr>
<td>732</td>
<td>003370e2-d54b-9d6b-c7e7-800c7f1144bf</td>
<td>521</td>
<td>age_45_64</td>
<td>F</td>
<td>white</td>
<td>ckd_3</td>
<td>F</td>
</tr>
<tr>
<td>733</td>
<td>003370e2-d54b-9d6b-c7e7-800c7f1144bf</td>
<td>522</td>
<td>age_45_64</td>
<td>F</td>
<td>white</td>
<td>ckd_3</td>
<td>F</td>
</tr>
<tr>
<td>734</td>
<td>003370e2-d54b-9d6b-c7e7-800c7f1144bf</td>
<td>523</td>
<td>age_45_64</td>
<td>F</td>
<td>white</td>
<td>ckd_3</td>
<td>F</td>
</tr>
<tr>
<td>735</td>
<td>003370e2-d54b-9d6b-c7e7-800c7f1144bf</td>
<td>524</td>
<td>age_45_64</td>
<td>F</td>
<td>white</td>
<td>ckd_3</td>
<td>F</td>
</tr>
<tr>
<td>736</td>
<td>003370e2-d54b-9d6b-c7e7-800c7f1144bf</td>
<td>525</td>
<td>age_45_64</td>
<td>F</td>
<td>white</td>
<td>ckd_3</td>
<td>F</td>
</tr>
<tr>
<td>737</td>
<td>003370e2-d54b-9d6b-c7e7-800c7f1144bf</td>
<td>526</td>
<td>age_45_64</td>
<td>F</td>
<td>white</td>
<td>ckd_3</td>
<td>F</td>
</tr>
<tr>
<td>738</td>
<td>003370e2-d54b-9d6b-c7e7-800c7f1144bf</td>
<td>527</td>
<td>age_45_64</td>
<td>F</td>
<td>white</td>
<td>ckd_3</td>
<td>F</td>
</tr>
<tr>
<td>739</td>
<td>003370e2-d54b-9d6b-c7e7-800c7f1144bf</td>
<td>528</td>
<td>age_45_64</td>
<td>F</td>
<td>white</td>
<td>ckd_3</td>
<td>F</td>
</tr>
<tr>
<td>740</td>
<td>003370e2-d54b-9d6b-c7e7-800c7f1144bf</td>
<td>529</td>
<td>age_45_64</td>
<td>F</td>
<td>white</td>
<td>ckd_3</td>
<td>F</td>
</tr>
<tr>
<td>741</td>
<td>003370e2-d54b-9d6b-c7e7-800c7f1144bf</td>
<td>530</td>
<td>age_45_64</td>
<td>F</td>
<td>white</td>
<td>ckd_3</td>
<td>F</td>
</tr>
<tr>
<td>742</td>
<td>003370e2-d54b-9d6b-c7e7-800c7f1144bf</td>
<td>531</td>
<td>age_45_64</td>
<td>F</td>
<td>white</td>
<td>ckd_3</td>
<td>F</td>
</tr>
<tr>
<td>743</td>
<td>003370e2-d54b-9d6b-c7e7-800c7f1144bf</td>
<td>532</td>
<td>age_45_64</td>
<td>F</td>
<td>white</td>
<td>ckd_3</td>
<td>F</td>
</tr>
<tr>
<td>744</td>
<td>003370e2-d54b-9d6b-c7e7-800c7f1144bf</td>
<td>533</td>
<td>age_45_64</td>
<td>F</td>
<td>white</td>
<td>ckd_3</td>
<td>F</td>
</tr>
<tr>
<td>745</td>
<td>003370e2-d54b-9d6b-c7e7-800c7f1144bf</td>
<td>534</td>
<td>age_45_64</td>
<td>F</td>
<td>white</td>
<td>ckd_3</td>
<td>F</td>
</tr>
<tr>
<td>746</td>
<td>003370e2-d54b-9d6b-c7e7-800c7f1144bf</td>
<td>535</td>
<td>age_45_64</td>
<td>F</td>
<td>white</td>
<td>ckd_3</td>
<td>F</td>
</tr>
<tr>
<td>747</td>
<td>003370e2-d54b-9d6b-c7e7-800c7f1144bf</td>
<td>536</td>
<td>age_45_64</td>
<td>F</td>
<td>white</td>
<td>ckd_3</td>
<td>F</td>
</tr>
<tr>
<td>748</td>
<td>003370e2-d54b-9d6b-c7e7-800c7f1144bf</td>
<td>537</td>
<td>age_45_64</td>
<td>F</td>
<td>white</td>
<td>ckd_3</td>
<td>F</td>
</tr>
<tr>
<td>749</td>
<td>003370e2-d54b-9d6b-c7e7-800c7f1144bf</td>
<td>538</td>
<td>age_45_64</td>
<td>F</td>
<td>white</td>
<td>ckd_3</td>
<td>F</td>
</tr>
<tr>
<td>750</td>
<td>003370e2-d54b-9d6b-c7e7-800c7f1144bf</td>
<td>539</td>
<td>age_45_64</td>
<td>F</td>
<td>white</td>
<td>ckd_3</td>
<td>F</td>
</tr>
<tr>
<td>751</td>
<td>003370e2-d54b-9d6b-c7e7-800c7f1144bf</td>
<td>540</td>
<td>age_45_64</td>
<td>F</td>
<td>white</td>
<td>ckd_4</td>
<td>F</td>
</tr>
</tbody>
</table>
Validation

```sql
select description, count(*) tally from conditions
    where description rlike('(Chronic kidney disease|End stage renal)')
group by description order by description
```

<table>
<thead>
<tr>
<th>description</th>
<th>tally</th>
</tr>
</thead>
<tbody>
<tr>
<td>Chronic kidney disease stage 1 (disorder)</td>
<td>3024</td>
</tr>
<tr>
<td>Chronic kidney disease stage 2 (disorder)</td>
<td>436</td>
</tr>
<tr>
<td>Chronic kidney disease stage 3 (disorder)</td>
<td>25</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>description</th>
<th>tally</th>
</tr>
</thead>
<tbody>
<tr>
<td>Chronic kidney disease stage 1 (disorder)</td>
<td>3837</td>
</tr>
<tr>
<td>Chronic kidney disease stage 2 (disorder)</td>
<td>6557</td>
</tr>
<tr>
<td>Chronic kidney disease stage 3 (disorder)</td>
<td>18607</td>
</tr>
<tr>
<td>Chronic kidney disease stage 4 (disorder)</td>
<td>10585</td>
</tr>
<tr>
<td>End stage renal disease (disorder)</td>
<td>17272</td>
</tr>
</tbody>
</table>
### Individual concepts

#### Including Covid

<table>
<thead>
<tr>
<th>Rank</th>
<th>Synthea Concept</th>
<th>Mercy EHR Concept</th>
<th>Pre-Covid Synthea Concept</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Suspected COVID-19</td>
<td>Essential hypertension (disorder)</td>
<td>Viral sinusitis (disorder)</td>
</tr>
<tr>
<td>2</td>
<td>COVID-19</td>
<td>Diabetes mellitus type 2 (disorder)</td>
<td>Acute viral pharyngitis (disorder)</td>
</tr>
<tr>
<td>3</td>
<td>Fever (finding)</td>
<td>Hyperlipidemia (disorder)</td>
<td>Acute bronchitis (disorder)</td>
</tr>
<tr>
<td>4</td>
<td>Cough (finding)</td>
<td>Cough (finding)</td>
<td>Normal pregnancy</td>
</tr>
<tr>
<td>5</td>
<td>Loss of taste (finding)</td>
<td>Asthma (disorder)</td>
<td>Streptococcal sore throat (disorder)</td>
</tr>
<tr>
<td>6</td>
<td>Viral sinusitis (disorder)</td>
<td>Gastroesophageal reflux disease (disorder)</td>
<td>Otitis media</td>
</tr>
<tr>
<td>7</td>
<td>Fatigue (finding)</td>
<td>Coronary arteriosclerosis (disorder)</td>
<td>Unhealthy alcohol drinking behavior (finding)</td>
</tr>
<tr>
<td>8</td>
<td>Sputum finding (finding)</td>
<td>Hypertriglyceridemia (disorder)</td>
<td>Severe anxiety (panic) (finding)</td>
</tr>
<tr>
<td>9</td>
<td>Hypoxemia (disorder)</td>
<td>Joint pain (finding)</td>
<td>Sprain of ankle</td>
</tr>
<tr>
<td>10</td>
<td>Respiratory distress (finding)</td>
<td>Body mass index 30+ - obesity (finding)</td>
<td>Prediabetes</td>
</tr>
</tbody>
</table>

**dominated by Covid-19**

**common chronic conditions**

**largely acute conditions**
Co-occurrence heatmap for Synthea data
EMR data
Co-Occurrence Probability Measure (COPM)

a) across all 107 conditions

<table>
<thead>
<tr>
<th>From</th>
<th>To</th>
<th>actual</th>
<th>simulation_pre</th>
<th>simulation_post</th>
</tr>
</thead>
<tbody>
<tr>
<td>actual</td>
<td>actual</td>
<td>0</td>
<td>0.2123</td>
<td>0.2008</td>
</tr>
<tr>
<td>simulation_pre</td>
<td>simulation_pre</td>
<td>0</td>
<td>0</td>
<td>0.0110</td>
</tr>
<tr>
<td>simulation_post</td>
<td>simulation_post</td>
<td></td>
<td></td>
<td>0</td>
</tr>
</tbody>
</table>

b) focused on just 6 CKD-related conditions

<table>
<thead>
<tr>
<th>From</th>
<th>To</th>
<th>actual</th>
<th>simulation_pre</th>
<th>simulation_post</th>
</tr>
</thead>
<tbody>
<tr>
<td>actual</td>
<td>actual</td>
<td>0</td>
<td>0.8680</td>
<td>0.2430</td>
</tr>
<tr>
<td>simulation_pre</td>
<td>simulation_pre</td>
<td>0</td>
<td>0</td>
<td>0.1207</td>
</tr>
<tr>
<td>simulation_post</td>
<td>simulation_post</td>
<td></td>
<td></td>
<td>0</td>
</tr>
</tbody>
</table>
**Design pattern recommendations**

- **No delay**
  - disease is not your destiny
- **Mechanistic progression**
  - allow interventions
- **More modular modules**
  - separate treatment from disease incidence and progression
  - use attributes to communicate between modules
Bonus Slides!

data visualization helps find bugs
Debugging Synthea with co-occurrence

EMR data:
mutually exclusive stages

Synthea data:
stages can co-occur
‘End stage renal disease’ condition toggles on and off

```sql
select patient, description, start, stop, encounter, code from conditions
where description rlike('\(Chronic kidney disease|End stage renal)\)')
order by patient, start
```

<table>
<thead>
<tr>
<th>patient</th>
<th>description</th>
<th>start</th>
<th>stop</th>
<th>encounter</th>
</tr>
</thead>
<tbody>
<tr>
<td>00118915-7610-1be1-fd03-21811ac23b71</td>
<td>Chronic kidney disease stage 1 (disorder)</td>
<td>2014-09-17</td>
<td>null</td>
<td>cdfd0045-026b-a2f9</td>
</tr>
<tr>
<td>00154e43-88d2-f074-2810-23c2ed04235f</td>
<td>Chronic kidney disease stage 3 (disorder)</td>
<td>2010-08-26</td>
<td>null</td>
<td>ac3401ab-e626-73e3</td>
</tr>
<tr>
<td>001792aa-daec-beab-969b-1a7a98c0dc67</td>
<td>Chronic kidney disease stage 3 (disorder)</td>
<td>2005-04-17</td>
<td>null</td>
<td>6b2c1374-3280-f9dd</td>
</tr>
<tr>
<td>001792aa-daec-beab-969b-1a7a98c0dc67</td>
<td>End stage renal disease (disorder)</td>
<td>2008-07-13</td>
<td>2009-07-19</td>
<td>f6787445-a37a-a0bf</td>
</tr>
<tr>
<td>001792aa-daec-beab-969b-1a7a98c0dc67</td>
<td>End stage renal disease (disorder)</td>
<td>2010-07-25</td>
<td>2011-07-31</td>
<td>4f70d0ec-3965-1a60</td>
</tr>
<tr>
<td>001792aa-daec-beab-969b-1a7a98c0dc67</td>
<td>End stage renal disease (disorder)</td>
<td>2012-08-05</td>
<td>2013-08-11</td>
<td>93ce8bb5-5949-1831</td>
</tr>
</tbody>
</table>
End Stage Renal Disease

Nothing writes this attribute; it is always nil
ON IMPROVING REALISM OF DISEASE MODULES IN SYNTHEA™

Social Determinant-Based Enhancements to Conditional Transition Logic

Response to the 2021 HHS Synthetic Health Data Challenge
Category I: Enhancements to Synthea™
Opioids Use Case

Team LMI
Brant Horio
Greg Pekar
Simon Whittle
Maureen Merkl
Linna Qiao

October 19, 2021
• Opioid Use Disorder (OUD) is a crisis
  – **Deadly outcomes**: 69,700+ Americans dying from opioid overdoses in 2020 (36% increase from prior year)
  – **Hard to fix**: OUD is a highly individual and complex care condition and highly influenced by SDOH
  – **Ongoing problem**: U.S. CBP confiscated more fentanyl in first half 2021 than last 3 years (Kaminsky 2021)
• Our research focused on one pathway to OUD
  – **Prescribed opioids may result in OUD**: often due to long term need to help with chronic pain
  – **SDOH (occupation) can drive onset of chronic pain**: particularly for heavy manual labor in rural areas
• The key technical idea was to demonstrate
  – a **software workflow** that mines open-source secondary data sources for SDOH (occupation)
  – modifications to the Synthea codebase to **operationalize the SDOH** (following published findings that could bridge occupation to relevant transition logic in Synthea’s OUD-related state machines, and
  – **increased realism** in Synthea’s generated populations with respect to opioid prescriptions
OUR APPROACH MINED SECONDARY DATA FOR OUD-RELEVANT SDOH TO AUGMENT SIMULATED PERSON JOURNEYS

Our technical approach seeks to automate data mining of publicly available data sources for SDOH that better characterize communities at the local level.

Lat/Lon \rightarrow Census Tract \rightarrow Localized SDOH (i.e., occupation)

User specifications inform Synthea runs
- e.g., location, population, diseases

Synthea draws from publicly available information and knowledge
- Publicly-available demographic data, e.g., US Census
- Health incidence statistics, e.g., CDC prevalence and incidence rates, NIH reports
- Clinical practice guidelines and protocols
- Medical coding dictionaries

Data is transformed
- PATIENT PROFILE
  - Demographics
  - Residence location
- HEALTH PROFILE
  - Health condition and disease
- CAREMAPS
- CODED NARRATIVES
- SOCIAL DETERMINANTS OF HEALTH
  - Household financial status
  - Insurance
  - Occupation (w.r.t. chronic pain)

An agent-based approach is used to generate a fully-synthetic population, simulating individuals from birth-to-death while tracking their healthcare journey.

Repeatable and accessible enhancement to Synthea codebase that introduces SDOH (or other Census tract-related Person attributes) for more realistic modeling.
We operationalize SDOH data as agent attributes to enhance transition logic in Synthea’s module builder.

**Prepare SDOH Data**
Mine data and create census tract-indexed file to occupation information for Bangor, ME.

**Assign SDOH Attributes to Person Agents**
Assign an occupation to each person based on Census statistics for our targeted occupation classes in their Census tract communities.

**Modify Relevant Disease Modules**
Adjust transition probabilities for a chronic low back pain condition, based on occupation and gender.

**Run Synthea and Validate**
Run ten trials of 32,000 patients with legacy_Synthea and LMI_Synthea.

Compare simulated outcomes for number of opioid prescriptions per capita.
We scoped research to Bangor, ME due to higher than national average prevalence of OUD and predominant SDOH that strongly influence OUD

Based on collaboration with University of Maine OUD researchers and published literature, we focused on:
- Maine’s prevalence of forestry and fishing occupations
- Correlation of these occupations to chronic musculoskeletal pain conditions
- Chronic pain as a pathway to prescribed opioids, potential abuse, and OUD

We processed American Community Survey data to derive likelihood of individuals having forestry and fishing occupations by Census tract in Penobscot County, ME (where Bangor is located) to build a tract-indexed file to assign Person occupations.

Drawing from literature (Yang, Halderman, Lu, and Baker 2016) assessing chronic low back pain risk associated with occupations and gender, to inform state machine transition probabilities.

Validation patterns were collaborated on with University of Maine researchers and focused on opioid prescription counts for State and County.
• Built a data ingestion method to access the tract-indexed file we created for occupation data
• Assigned Persons to a Census tract
  – For greater localized level of detail than zip code and to align with our data sources of interest
  – Based on shortest distance between Synthea’s provided latitude-longitude residence coordinates to the centroid of the nearest Census block
  – Given Block assignment, then we assigned the Person to the appropriate Census tract
• Assigned occupation to Persons
  – Using assigned tract number, probabilistically assign occupation to Person by referencing occupation file
• Person owns one new attribute for occupation which can now be used for transitional conditional logic in the disease module state machines
MODIFICATION OF THE DISEASE MODULES
To validate our modifications to Synthea:
  – We ran 10 trials, each with different random seeds
  – Each run generated 32,000 patients to approximately represent the entire City of Bangor, ME
  – We used the same 10 seeds for legacy_Synthea and LMI_Synthea to allow comparison between software versions

We used the Prescription Monitoring Program Annual Report 2020 (Maine HSS 2021) and data from the CDC (CDC 2020) as our ground truth data for the number of opioid prescriptions in Maine
VALIDATION RESULTS SHOW BETTER OUTCOMES FOR OUR SUBPOPULATION OF INTEREST

Legacy Synthea results for Bangor, ME fall short of both County and State benchmarks.

LMI Synthea shows slight improvement, but as the subpopulation we augmented was very small compared to the overall population, it did little to shift metrics at the Bangor scale.

Evaluating only the subpopulation we changed in the fishing and forestry occupation, both State and County benchmarks fall within the interquartile range of our simulated results.

The results are promising in that outcomes are demonstrative of how population level representation might be improved by adjusting many relevant subpopulations.
BENEFITS TO HEALTHCARE COMMUNITY

• Benefits to Synthea
  – Minimal codebase modifications to allow an easy pull request
  – Method of ingesting Census tract level data and assigning to patients can be generalized to other similar data sets at a local scale
  – Focus on enhancing Person attributes to integrate SDOH allows Synthea ecosystem to be fully leveraged (e.g., ModuleBuilder)

• Benefits to researchers
  – Provides additional experiment factors to incorporate into Synthea’s disease modules for greater detail for the conditional logic and state transition path possibilities
  – Repeatable and accessible enhancement to Synthea codebase that introduces SDOH (or other Census tract-related Person attributes) for more realistic modeling

• Benefits to health IT developers
  – Provides a working framework to ingest secondary data and assign them as patient attributes

• Broader healthcare community
  – Individualized SDOH drives complex care and critically needs to be better understood, analyzed, and modeled to further advance patient-centered outcomes research
FUTURE WORK

- Additional SDOH factors based on Census tracts and blocks may be examined (e.g., homelessness, access to care, access to food)
- Work is needed to find better ways to account for relevant correlations between SDOH (e.g., our assignment of occupation neglected potentially relevant relationships to Synthea’s assigned socioeconomic status or age)
- The team will be continuing to collaborate with University of Maine researchers with the objective of integrating Synthea for OUD research
There is a lot of code

- Big learning curve to understand the codebase and how the modules interact with one another (e.g., geography folder has a fully implemented quad tree we could have used for centroid distance calculations—we implemented a less efficient in a sorted map-based approach to finding the closest centroid)

- Successes in working with existing Synthea code base
  - Minimal modifications to preserve functioning software with least risk of “breaking it”
    - Integration of external analysis (e.g., tract and occupation) into the Synthea process
    - Use of input files and existing data loader methods
    - Primarily modified the pickDemographics() step of the Person generation logic
  - Preserved software practices and workflows to make sure our enhancements looked roughly like the rest of the project (would not be a stumbling block to existing Synthea developers and users)
  - Focused on capitalizing on the very polished Module Builder application to operationalize our enhancements in other modules
The Necessity of Realistic Synthetic Health Data Development Environments

Category II Entry: Novel Uses of Synthea Generated Synthetic Data

Team: Particle Health, Submitter: Parker Bannister
Background and Objectives

- APIs will be the tool driving national access to health data
  - 21st Century Cures Act and TEFCA
- Synthetic health data vs. Real health data
  - Synthea
    - Single CCDA documents, separately generated free-text notes
  - Access to Data
    - Download CCDA to Local Machine
    - APIs for FHIR, not CCDA
- Why is a realistic development environment needed?
  - Allows researchers and developers to seamlessly transition to real patient information
    - Time, Cost, and Liability
- The Particle Health Sandbox Objective
  - Point-In-Time documents
  - In-Document Synthetic provider notes
  - Focus on synthetic populations with specific conditions
  - Validation
Document Generation pt.1

- Synthea_Runner.py
  - Generates base Synthea documents
  - Finds patients with conditions specified to create population of interest
    - Regex - Synthea symptoms.csv
  - Stores files for population of interest and processes them with the point in time document generator
  - Finally Validates Results
```python
list_of_conditions = ['covid19', 'diabetes', 'lung cancer', 'opiod addiction']
p = arg_parser.ArgumentParser()
p.add_argument('--num-patients', type=int, required=True)
p.add_argument('--condition', type=str, required=True, choices=list_of_conditions)
args = p.parse_args(sys.argv[1:])

num_patients = args.num_patients
condition = args.condition

list_of_modules = ['covid19', 'metabolic', 'lung cancer', 'opiod addiction']

if condition == 'covid19':
    module = list_of_modules[0]
if condition == 'diabetes':
    module = list_of_modules[1]
if int(num_patients) < 10:
    num_patients = 25
if condition == 'lung cancer':
    module = list_of_modules[2]
    num_patients = 2500
if condition == 'opiod addiction':
    module = list_of_modules[3]
if int(num_patients) < 10:
    num_patients = 25

# run synthea with module and number of patients specified
print(f'
 Running Synthea to Generate Base Synthetic Patient Data: {num_patients}')
n = os.system(f'../run_synthea.py --num_patients {num_patients} --module {module} --a 0-99')

# load csv produced from synthea with conditions of population generated
print(f'
Finding Patients with Conditions of Interest: {num_patients}')
symptoms_csv_load = []
with open(f'../output/symptoms/csv/symptoms.csv') as symptoms_csv:
    for i in symptoms_csv:
        symptoms_csv_load.append(i)

# define function to find condition of interest
def check_if_match(words):
    p = re.split(r'\(', words)
    for i in range(len(p)):
        pattern = p[i].lower()
        results = get_close_matches(condition, patterns)
        return results, patterns

# change opioid addiction to terms found in symptom csv
if condition == 'opiod addiction':
    condition = 'drug overdose'

# get unique patient ids for patients with condition of interest
```
os.system('python point_in_time_document_generator.py')

# copy output files and directory to output folder:
for i in os.listdir('./pitt_gen_output/'):  
    shutil.copytree('./pitt_gen_output/' + i, '..', '+condition + '+output_ + str(today) + '+generator_output/' + i)

copyfile('./output_directory.csv', '..', '+condition + '+output_ + str(today) + '+output_directory.csv')

# clear notes, synthea ccds, pitt gen output_file, directory
for i in os.listdir('Notes/'):  
    os.remove('Notes/' + i)
for i in os.listdir('./Synthea_CCDs/'):  
    os.remove('./Synthea_CCDs/' + i)
for i in os.listdir('./Output/'):  
    if i.startswith('output_'):  
        os.remove('output_ + i')
    else:  
        shutil.rmtree('output_ + i', ignore_errors=True)
for i in os.listdir('./pitt_gen_output/'):  
    shutil.rmtree('./pitt_gen_output/' + i, ignore_errors=True)
    os.remove('./output_directory.csv')

print('
' + 'Generation Complete' + '
')

# validate point in time coda documents
print('
' + 'Validating Output Point In Time CDAD Documents' + '
')

# loop thru output files to validate
patients = []
val_output = []

urllib3.disable_warnings()
for i in os.listdir('../' + condition + '+output_ + str(today) + '+generator_output/'):  
    if not i.startswith('output_'):  
        print(i)
        for j in os.listdir('../' + condition + '+output_ + str(today) + '+generator_output/ + i):
            if not j.startswith('.'):  
                print('%\FOLDER: ' + j)  
                # loop thru folders for patient
                counter = 0
                for k in os.listdir('../' + condition + '+output_ + str(today) + '+generator_output/ + i + '/' + j):
                    if not k.startswith('.'):  
                        print('%\FILE: ' + k)
                        patients.append(i)
                        folder_name = i

url = 'https://cdad.healthit.gov/scorercard/ccdascorecardservice2'
myfile = {'ccdasFile': (k, open(data_file, 'rb'))}

r = requests.post(url, data=myfile, verify=False).json()
Document Generation pt.2

- Point_in_Time_Document_Generator.py
  - Parses base Synthea CCDA and stores sections relevant to point in time document types
    - LXML
  - Templates sections into new document
    - Jinja2
    - XML Templates for new point in time documents
    - Addition of in-document notes
Validation

- HealthIT.gov’s CCDA Scorecard 2.0
  - API Implementation
- CSV of Results for synthetic population generated
  - File score validation per file generated for entire population of synthetic patients
# Validation Portion of Synthea_Runner.py and Results

```python
# loop thru output files to validate
patients = []
val_output = []
urllib3.disable_warnings()
for i in os.listdir('../' + condition + '_output_' + str(today) + '/generator_output/'):  
    if not i.startswith('.'):  
        print(i)  
        # loop thru folders for patient
        counter = 0
        for j in os.listdir('../' + condition + '_output_' + str(today) + '/generator_output/' + i):  
            if not j.startswith('.'):  
                print('tFOLDER: ' + j)  
                # loop thru files for patient
                for k in os.listdir('../' + condition + '_output_' + str(today) + '/generator_output/' + i + '/' + j):  
                    if not k.startswith('.'):  
                        print('tFile: ' + k)
                        patients.append(i)
                        folder_name = i
                        data_file = '../' + condition + '_output_' + str(today) + '/generator_output/' + i + '/' + j + '/' + k
                        url = 'https://ccda.healthit.gov/scorecard/ccdascorecardservice2'
                        myfile = ('ccdaFile': (k, open(data_file, 'rb')))
                        r = requests.post(url, files=myfile, verify=False).json()
                        val_output.append(list(r.items()))
                        data = pd.DataFrame(data=val_output, columns=['ErrorMessage', 'FileName', 'CCDADocumentType', 'Results', 'ReferenceResults', 'ErrorList', 'SchemaErrors'])
                        val_table = data
                        val_table.insert(0, 'Patient', patients, True)
                        val_table.to_csv('../' + condition + '_output_' + str(today) + '/validation_results.csv', index=False)
                        print('
' + 'Validation Complete' + '
')
```
<table>
<thead>
<tr>
<th>Patient ID</th>
<th>Error Message</th>
<th>File Name</th>
<th>CDDA Document Type</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>Error Message</td>
<td>Continuity of Care Document</td>
<td>CCDaDocumentType</td>
</tr>
<tr>
<td>1</td>
<td>Error Message</td>
<td>Encounter Summary</td>
<td>CCDaDocumentType</td>
</tr>
<tr>
<td>2</td>
<td>Error Message</td>
<td>Lab Summary</td>
<td>CCDaDocumentType</td>
</tr>
<tr>
<td>3</td>
<td>Error Message</td>
<td>Immunizations Summary</td>
<td>CCDaDocumentType</td>
</tr>
<tr>
<td>4</td>
<td>Error Message</td>
<td>Encounter Summary</td>
<td>CCDaDocumentType</td>
</tr>
<tr>
<td>5</td>
<td>Error Message</td>
<td>Continuity of Care Document</td>
<td>CCDaDocumentType</td>
</tr>
<tr>
<td>6</td>
<td>Error Message</td>
<td>Refill Summary</td>
<td>CCDaDocumentType</td>
</tr>
<tr>
<td>7</td>
<td>Error Message</td>
<td>Encounter Summary</td>
<td>CCDaDocumentType</td>
</tr>
<tr>
<td>8</td>
<td>Error Message</td>
<td>Encounter Summary</td>
<td>CCDaDocumentType</td>
</tr>
<tr>
<td>9</td>
<td>Error Message</td>
<td>Immunizations Summary</td>
<td>CCDaDocumentType</td>
</tr>
<tr>
<td>10</td>
<td>Error Message</td>
<td>Immunizations Summary</td>
<td>CCDaDocumentType</td>
</tr>
<tr>
<td>11</td>
<td>Error Message</td>
<td>Lab Summary</td>
<td>CCDaDocumentType</td>
</tr>
<tr>
<td>12</td>
<td>Error Message</td>
<td>Lab Summary</td>
<td>CCDaDocumentType</td>
</tr>
<tr>
<td>13</td>
<td>Error Message</td>
<td>Continuity of Care Document</td>
<td>CCDaDocumentType</td>
</tr>
<tr>
<td>14</td>
<td>Error Message</td>
<td>Continuity of Care Document</td>
<td>CCDaDocumentType</td>
</tr>
<tr>
<td>15</td>
<td>Error Message</td>
<td>Lab Summary</td>
<td>CCDaDocumentType</td>
</tr>
<tr>
<td>16</td>
<td>Error Message</td>
<td>Continuity of Care Document</td>
<td>CCDaDocumentType</td>
</tr>
<tr>
<td>17</td>
<td>Error Message</td>
<td>Encounter Summary</td>
<td>CCDaDocumentType</td>
</tr>
<tr>
<td>18</td>
<td>Error Message</td>
<td>Lab Summary</td>
<td>CCDaDocumentType</td>
</tr>
<tr>
<td>19</td>
<td>Error Message</td>
<td>Encounter Summary</td>
<td>CCDaDocumentType</td>
</tr>
<tr>
<td>20</td>
<td>Error Message</td>
<td>Encounter Summary</td>
<td>CCDaDocumentType</td>
</tr>
<tr>
<td>21</td>
<td>Error Message</td>
<td>Encounter Summary</td>
<td>CCDaDocumentType</td>
</tr>
<tr>
<td>22</td>
<td>Error Message</td>
<td>Encounter Summary</td>
<td>CCDaDocumentType</td>
</tr>
<tr>
<td>23</td>
<td>Error Message</td>
<td>Encounter Summary</td>
<td>CCDaDocumentType</td>
</tr>
<tr>
<td>24</td>
<td>Error Message</td>
<td>Encounter Summary</td>
<td>CCDaDocumentType</td>
</tr>
<tr>
<td>25</td>
<td>Error Message</td>
<td>Encounter Summary</td>
<td>CCDaDocumentType</td>
</tr>
<tr>
<td>26</td>
<td>Error Message</td>
<td>Immunizations Summary</td>
<td>CCDaDocumentType</td>
</tr>
<tr>
<td>27</td>
<td>Error Message</td>
<td>Lab Summary</td>
<td>CCDaDocumentType</td>
</tr>
<tr>
<td>28</td>
<td>Error Message</td>
<td>Encounter Summary</td>
<td>CCDaDocumentType</td>
</tr>
<tr>
<td>29</td>
<td>Error Message</td>
<td>Immunizations Summary</td>
<td>CCDaDocumentType</td>
</tr>
<tr>
<td>30</td>
<td>Error Message</td>
<td>Continuity of Care Document</td>
<td>CCDaDocumentType</td>
</tr>
<tr>
<td>31</td>
<td>Error Message</td>
<td>Continuity of Care Document</td>
<td>CCDaDocumentType</td>
</tr>
<tr>
<td>32</td>
<td>Error Message</td>
<td>Encounter Summary</td>
<td>CCDaDocumentType</td>
</tr>
<tr>
<td>33</td>
<td>Error Message</td>
<td>Lab Summary</td>
<td>CCDaDocumentType</td>
</tr>
<tr>
<td>34</td>
<td>Error Message</td>
<td>Continuity of Care Document</td>
<td>CCDaDocumentType</td>
</tr>
<tr>
<td>35</td>
<td>Error Message</td>
<td>Continuity of Care Document</td>
<td>CCDaDocumentType</td>
</tr>
<tr>
<td>36</td>
<td>Error Message</td>
<td>Encounter Summary</td>
<td>CCDaDocumentType</td>
</tr>
</tbody>
</table>
Final Output and Usage

- Multiple Point in Time documents generated from Synthea CCDAs
  - Encounter Summary
  - Immunizations Summary
  - Lab Summary
  - Refill Summary
- In-Document Synthetic Provider Notes
- API Interface to Access Information that mirrors the process of National Network APIs
import requests

url = 'https://sandbox.acr8949-apv8948-904-90a4-7659f113062e/api/v1/profiles'
headers = {'Content-Type': 'application/json', 'Authorization': 'JWT'}
data = {
    "address_city": "West Bridgewater",
    "address_lines": [
        "126 McLaughlin Ferry"
    ],
    "address_state": "Massachusetts",
    "date_of_birth": "1976-03-11",
    "email": "Arlie@doc.com",
    "family_name": "Rolfson",
    "gender": "Male",
    "given_name": "Arlie",
    "mpi": "1234",
    "postal_code": "02324",
    "purpose_of_use": "TREATMENT",
    "asv": "123-45-6789",
    "telephone": "1-234-567-8910"
}

r = requests.post(url, headers=headers, json=data)

print(r.json())

query_id = r.json()['id']

url = 'https://sandbox.acr8949-apv8948-904-90a4-7659f113062e/api/v1/profiles/' + query_id
headers = {'Content-Type': 'application/json', 'Authorization': 'JWT'}
r = requests.get(url, headers=headers)

print(r.json())
## Final Output and Usage

### Continuity of Care Document

#### Encounter Summary

<table>
<thead>
<tr>
<th>Description</th>
<th>Code System</th>
<th>Code Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>History of encounters</td>
<td>LOINC</td>
<td>2.16.840.1.113883.6.1</td>
</tr>
</tbody>
</table>

#### Drug Rehabilitation and Detoxification

<table>
<thead>
<tr>
<th>Description</th>
<th>Code System</th>
<th>Code Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Drug rehabilitation and detoxification</td>
<td>SNOMED-CT</td>
<td>56876005</td>
</tr>
</tbody>
</table>

#### Allergy List

<table>
<thead>
<tr>
<th>Description</th>
<th>Code System</th>
<th>Code Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Allergy List</td>
<td>LOINC</td>
<td>2.16.840.1.113883.6.1</td>
</tr>
</tbody>
</table>
Final Output and Usage

Chief Complaint
No complaints.

History of Present Illness
Austin578 is a 46 year-old non-hispanic white male.

Social History
Patient is married. Patient has a documented history of opioid addiction. Patient is an active smoker and is an alcoholic. Patient identifies as heterosexual.

Patient comes from a low socioeconomic background. Patient has a high school education. Patient currently has UnitedHealthcare.

Allergies
No Known Allergies.

Medications
No Active Medications.

Assessment and Plan

Plan
Discussion

- Progression from original Synthea (Single CCDA)
  - Why realistic data helps developers innovate
- Validation Component
- Delivering a specific sub-population of patients
  - Impact on innovation and research (Opioid or Complex-Care)
- Synthea solution publically available on our GitHub Repository:
  - https://github.com/ParticleHealth/Particle-Health-Sandbox-Environment
- Leveraged Open-Source tools and libraries
- Visit our website to use our sandbox environment
  - Pre-loaded with Synthea Data modified with our solution
  - https://wwwparticlehealth.com
Lessons Learned and Future Work

- Lessons Learned
  - Synthea is a powerful tool for generating synthetic CCDA and equivalent FHIR data
  - There are many potential opportunities to develop on top of Synthea to improve upon it and the use cases it can deliver
  - Real Clinical Information is highly variable and comes in different shapes and sizes
  - Quality development environments are important for enabling innovation
  - Policy greatly impacts the future direction of health information technology

- Future Work
  - Expanding Point In Time Document Types
    - ie. Discharge summaries
  - Generating other types of data seen in clinical practice
    - ie. Allergies
Thank you!

Questions?

go@particlehealth.com
Empirical inference of Underlying Condition Probabilities Using Synthea-Generated Synthetic Health Data

Team TeMa

Dr. Michael D. Teter  
miketeter@yahoo.com

Dr. Christopher E. Marks  
cemarks@alum.mit.edu

Challenge Category: II (Novel Uses of Synthea Generated Data)
Problem Motivation

• Simulation is often used to investigate complicated phenomena for which analytic determination of outcome probabilities is intractable.
• Synthea is built in a way that makes it well-suited for this purpose.
  • It inputs conditional probabilities that can be validated.
  • Its outputs are the result of tailorable combinations of these input probabilities.

Our Task

• Use Synthea-generated data to investigate relationships between a patient's pathology and a given set of symptoms and severities.
Our Methods (1 of 2)

- Canonical Bayesian analysis
- Using empirical distributions in Synthea data, no need to enumerate a complicated tree.
Our Methods (2 of 2)

- Graph-based machine learning method.
- We decide which variables are related.
- More versatile than the strictly empirical model.

Bayesian Network
Example Outputs

Patient
- 5 years old
- Female
- Fever

Empirical Bayes

Bayesian Network
Validation

• Internal Validation through code testing

• External Validation through comparison with existing tools.

• A useful method for validating Synthea!

External validation: comparison with WebMD

<table>
<thead>
<tr>
<th>WebMD</th>
<th>Synthea Bayes</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bacterial Pneumonia</td>
<td>Viral sinusitis (disorder)</td>
</tr>
<tr>
<td>Middle Ear Infection</td>
<td>Acute bronchitis (disorder)</td>
</tr>
<tr>
<td>Viral Pneumonia</td>
<td>Acute viral pharyngitis (disorder)</td>
</tr>
<tr>
<td>Influenza (Flu) Child</td>
<td>Streptococcal sore throat (disorder)</td>
</tr>
<tr>
<td>Strep Throat</td>
<td>Sinusitis (disorder)</td>
</tr>
</tbody>
</table>
Summary & Next Steps

• Our approach can be extended to account for demographics, encounter types, patient location, patient history, etc.

• Compare results to known distributions as a way of validating Synthea.

• Identify areas where Synthea can be improved.
  • Standardization!

• Look for real-world applications.
  • Rare pathologies?
SPATIOTEMPORAL BIG DATA ANALYSIS OF THE OPIOID EPIDEMIC IN ILLINOIS

• Office of the National Coordinator for Health Information Technology (ONC) Synthetic Health Data Challenge
• Category II: Novel Uses of Synthea™ Generated Synthetic Data
• Arash Jalali, MPH, MSHI
• Sean Huang, MD
• Karl Kochendorfer, MD, FAAFP
UI HEALTH

- Comprehensive care, education, and research to train health care leaders and foster healthy communities in Illinois and beyond.
- 465 bed tertiary care hospital, 21 outpatient clinics, 11 federally qualified Mile Square Health Center locations
- Campuses in Chicago, Peoria, Quad Cities, Rockford, Springfield, and Urbana
- 7 Health Science colleges: Applied Health Sciences, Dentistry, Medicine, Nursing, Pharmacy, School of Public Health, Jane Addams College of Social Work

1 in 3
IL Physicians trained at UIC

Only
Public, Research 1 University in Chicago

7
Health Sciences Colleges at UIC

$243M
in total annual health sciences research funding

1 in 4
IL Social Workers Trained at UIC

Over 8,000
IL Nurses trained at UIC

Smarter Public Health Prevention Systems
**OPIOID CRISIS: CHICAGO STATISTICS**

- 2019: 855 people died from opioid overdoses (from 793 previous year)
- 2018 -> 2019: Opioid-related overdose death rate increased by 10.1%
- CFD EMS team responded to average 29 responses per day (increase in 25.4%)
OPIOID CRISIS: CHICAGO STATISTICS

- Men
- Aged 45-64 years old
- Non-Hispanic African-Americans
- Use of combination of other opioids and illicit drugs
  - Cocaine
- High economic hardship
  - Education
  - Income levels, Unemployment
  - Crowded Housing
In 2019, CFD EMS responded to a total of 10,490 opioid-related overdoses occurring throughout all 77 community areas.

Most impacted community areas for EMS responses in 2019:
- Austin (n=1,343)
- West Garfield Park (n=1,094)
- Humboldt Park (n=1,090)

Least impacted area for EMS responses 2019:
- Mount Greenwood (n=8)
- Burnside (n=7)
- Hegewisch (n=7)
- Edison Park (n=6)
WORKPLACE INJURIES

- Especially with prescription pain relievers
- NIOSH, John Howard:
  - Potential for addiction may be preceded by injuries that happen in the workplace, with the consequences affecting both an individual’s working life as well as their home life
- Exposure to opiate powders -> hazardous environment for healthcare workers
“develop and implement systems to collect timely, adequate and standardized data to identify at-risk populations, and implement public health interventions that directly address removing structural and racial inequities.”

While data is critical to improving outcomes, current data is:

...incomplete
...not standardized for comparison
...not timely
...widely variable from location to location

Difficulties remain in accessing high quality, timely, comprehensive and standardized data. While metrics are generally available for drug-related overdoses, data for non-fatal overdoses and other key indicators are not widely collected or standardized across states and communities. These data gaps greatly hinder understanding of local situations and advancing prevention, treatment and harm reduction efforts.

Inadequate data collection prevents effective public health interventions to reduce overdose and death.

Data categories
- Fatal overdoses
- Non-fatal overdoses
- No data

Prescriptions
PDMP
HITAC RECOMMENDATIONS ON PUBLIC HEALTH DATA SYSTEMS

- Improving interoperability
  - NEMSIS
  - Cloud computing

- Synthetic syndromic surveillance to assist “traditionally under-resourced areas to support creation of a public health system able to support health equity and health disparities”
  - Help facilitate interoperability, geolocation
  - Merging with census and other SDOH data

- Explore traditional and non-traditional data sources to assist with early identification of early clusters/outbreaks of disease incidence
SPATIOTEMPORAL BIG DATA ANALYSIS OF OPIOID EPIDEMIC IN ILLINOIS

- Spatiotemporal distribution of EMS 911 calls and ambulance dispatches related to drug overdoses
- Obtain Chicago EMS data – store in Azure Data Lake. Scripts execute over Azure Cloud
- Opiate cases identified and geospatial information extracted
- Analysis on ArcGIS enterprise
- Enrich opiate cases with Esri demographic and census data of surrounding neighborhoods
- Machine learning to understand features to predict opiate use
SMARTER PUBLIC HEALTH PREVENTION SYSTEM (SPHPS) SYNTHETIC DATA INTEGRATION OF PRE-HOSPITAL TO HOSPITAL DATA
AZURE MODERN ANALYTICS ARCHITECTURE FOR SYNTHETIC SYNDROMIC SURVEILLANCE
CHICAGO EMS DATA

• National Emergency Medical Services Information System (NEMSIS)
• Identify opiate cases on SQL Server
• Provider’s Primary Impression
• Primary Symptom

https://nemsis.org/
CHICAGO EMS DATA

• Provider’s Primary Impression and/or Primary Symptom = ‘Opioid related disorders’, ‘Opioid use, unspecified’. Or ICD-10 codes in F11, T40 categories or Z79.891
ARCGIS ENTERPRISE

- Geospatial data management, data visualization, analytics, geospatial forecasting of opiates and overdoses
- Space time cubes for all overdoses and opioid cases
- Create predictive models using time series forecasting tools
- Curve fit forecast models
  - Forecast future future values using curve fitting
- Exponential smoothing forecast model – predicts values by decomposing time series at each location into seasonal and trend components
<table>
<thead>
<tr>
<th>Method</th>
<th>Forecast Equation</th>
</tr>
</thead>
<tbody>
<tr>
<td>xt+1</td>
<td>a*t - b; a=0.009804, b=0.137255</td>
</tr>
<tr>
<td>xt</td>
<td>k + a*exp(b*t); k=0.084258, a=-0.000747, b=0.316302</td>
</tr>
<tr>
<td>xt</td>
<td>a*t^2+b*t+c; a=0.001548, b=-0.039474, c=0.238390</td>
</tr>
<tr>
<td>xt</td>
<td>a*t^2+b*t+c; a=-0.004902, b=0.083333, c=-0.117647</td>
</tr>
<tr>
<td>xt</td>
<td>a*t^2+b*t+c; a=-0.002580, b=0.046182, c=-0.083591</td>
</tr>
<tr>
<td>xt</td>
<td>k+a*exp(-b<em>exp(-c</em>t)); k=0.000000, a=0.224779, b=52.858909, c=0.6076</td>
</tr>
<tr>
<td>xt</td>
<td>a*t^2+b*t+c; a=-0.002580, b=0.046182, c=-0.083591</td>
</tr>
<tr>
<td>xt</td>
<td>a*t^2+b*t+c; a=-0.001935, b=0.038313, c=-0.077399</td>
</tr>
<tr>
<td>xt</td>
<td>a*t+b; a=-0.007353, b=0.117647</td>
</tr>
<tr>
<td>xt</td>
<td>a*t+b; a=0.017157, b=-0.078431</td>
</tr>
<tr>
<td>xt</td>
<td>a*t^2+b*t+c; a=0.008256, b=-0.073271, c=0.094943</td>
</tr>
<tr>
<td>xt</td>
<td>k+a*exp(-b<em>exp(-c</em>t)); k=0.000010, a=0.557301, b=91.371438, c=0.3877</td>
</tr>
<tr>
<td>xt</td>
<td>a*t+b; a=-0.007353, b=0.117647</td>
</tr>
<tr>
<td>xt</td>
<td>a*t+b; a=0.014706, b=-0.058824</td>
</tr>
<tr>
<td>xt</td>
<td>a*t^2+b*t+c; a=-0.006579, b=0.107714, c=-0.106295</td>
</tr>
<tr>
<td>xt</td>
<td>a*t^2+b*t+c; a=-0.005934, b=0.099845, c=-0.158927</td>
</tr>
<tr>
<td>xt</td>
<td>k+a*exp(-b<em>exp(-c</em>t)); k=0.000000, a=0.194351, b=108.978494, c=0.859</td>
</tr>
<tr>
<td>xt</td>
<td>a*t^2+b*t+c; a=0.004902, b=-0.063725, c=0.254902</td>
</tr>
<tr>
<td>xt</td>
<td>k+a*exp(b*t); k=0.084258, a=-0.000747, b=0.316302</td>
</tr>
<tr>
<td>xt</td>
<td>a*t+b; a=0.009804, b=-0.019608</td>
</tr>
</tbody>
</table>
EXPONENTIAL SMOOTHING FORECAST MODEL
• Geospatial information -> upload into ESRI
• Demographic data store that allows for geocoding and automated enrichment

• USA 2020 demographic data
• USA 2010 Census Demographic Data
• USA 2014/2018 American Community Survey (ACS) Demographic Data,
• USA 2020 Consumer Expenditure data
• USA 2020 Tapestry Segmentation Data.
• Its geography information is updated to 2020/2021.
5-MINUTE WALK TIME
### 2010 Census Profile

#### Prepared by Esri

**2010 Census Profile**

1740 W Taylor St, Chicago, Illinois, 60612-5

**2000-2010 Annual Rate**

<table>
<thead>
<tr>
<th>Population</th>
<th>2000</th>
<th>2010</th>
<th>Annual Rate</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total</td>
<td>654</td>
<td>658</td>
<td>0.61%</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Households</th>
<th>2006</th>
<th>2010</th>
<th>Annual Rate</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total</td>
<td>554</td>
<td>558</td>
<td>0.72%</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Population by Race</th>
<th>Number</th>
<th>Percent</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total</td>
<td>654</td>
<td>100.0%</td>
</tr>
<tr>
<td>Population Reporting One Race</td>
<td>622</td>
<td>95.1%</td>
</tr>
<tr>
<td>White</td>
<td>276</td>
<td>42.4%</td>
</tr>
<tr>
<td>Black</td>
<td>145</td>
<td>22.2%</td>
</tr>
<tr>
<td>American Indian</td>
<td>7</td>
<td>0.1%</td>
</tr>
<tr>
<td>Asian</td>
<td>195</td>
<td>29.8%</td>
</tr>
<tr>
<td>Pacific Islander</td>
<td>1</td>
<td>0.2%</td>
</tr>
<tr>
<td>Some Other Race</td>
<td>20</td>
<td>4.3%</td>
</tr>
<tr>
<td>Population Reporting Two or More Races</td>
<td>32</td>
<td>4.9%</td>
</tr>
</tbody>
</table>

#### Population by Sex

<table>
<thead>
<tr>
<th>Male</th>
<th>Female</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total</td>
<td>352</td>
</tr>
<tr>
<td>20</td>
<td>46.2%</td>
</tr>
<tr>
<td>276</td>
<td>43.8%</td>
</tr>
</tbody>
</table>

#### Population by Age

<table>
<thead>
<tr>
<th>Age</th>
<th>2000</th>
<th>2010</th>
<th>Annual Rate</th>
</tr>
</thead>
<tbody>
<tr>
<td>0-4</td>
<td>60</td>
<td>59</td>
<td>0.17%</td>
</tr>
<tr>
<td>5-9</td>
<td>52</td>
<td>51</td>
<td>0.2%</td>
</tr>
<tr>
<td>10-14</td>
<td>56</td>
<td>0.2%</td>
<td></td>
</tr>
<tr>
<td>15-19</td>
<td>61</td>
<td>62</td>
<td>0.72%</td>
</tr>
<tr>
<td>20-24</td>
<td>62</td>
<td>63</td>
<td>0.72%</td>
</tr>
<tr>
<td>25-29</td>
<td>58</td>
<td>60</td>
<td>0.72%</td>
</tr>
<tr>
<td>30-34</td>
<td>54</td>
<td>56</td>
<td>0.72%</td>
</tr>
<tr>
<td>35-39</td>
<td>51</td>
<td>53</td>
<td>0.72%</td>
</tr>
<tr>
<td>40-44</td>
<td>53</td>
<td>54</td>
<td>0.72%</td>
</tr>
<tr>
<td>45-49</td>
<td>50</td>
<td>51</td>
<td>0.72%</td>
</tr>
<tr>
<td>50-54</td>
<td>48</td>
<td>50</td>
<td>0.72%</td>
</tr>
<tr>
<td>55-59</td>
<td>40</td>
<td>42</td>
<td>0.72%</td>
</tr>
<tr>
<td>60-64</td>
<td>35</td>
<td>36</td>
<td>0.72%</td>
</tr>
<tr>
<td>65-69</td>
<td>30</td>
<td>31</td>
<td>0.72%</td>
</tr>
<tr>
<td>70-74</td>
<td>20</td>
<td>21</td>
<td>0.72%</td>
</tr>
<tr>
<td>75-79</td>
<td>8</td>
<td>9</td>
<td>0.72%</td>
</tr>
<tr>
<td>80-84</td>
<td>5</td>
<td>6</td>
<td>1.2%</td>
</tr>
<tr>
<td>85+</td>
<td>2</td>
<td>3</td>
<td>1.2%</td>
</tr>
</tbody>
</table>

#### Households by Type

<table>
<thead>
<tr>
<th>Type</th>
<th>2010</th>
<th>Annual Rate</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total</td>
<td>558</td>
<td>100.0%</td>
</tr>
<tr>
<td>Households by Person</td>
<td>558</td>
<td>100.0%</td>
</tr>
<tr>
<td>Households with 1 Person</td>
<td>20</td>
<td>3.6%</td>
</tr>
<tr>
<td>Households with 2 People</td>
<td>45</td>
<td>8.1%</td>
</tr>
<tr>
<td>Family Households</td>
<td>59</td>
<td>10.6%</td>
</tr>
<tr>
<td>Husband Wife Families</td>
<td>27</td>
<td>4.9%</td>
</tr>
<tr>
<td>With Minor Children</td>
<td>26</td>
<td>4.7%</td>
</tr>
<tr>
<td>Other Family Households</td>
<td>18</td>
<td>3.3%</td>
</tr>
<tr>
<td>Male, Female Children</td>
<td>10</td>
<td>1.8%</td>
</tr>
<tr>
<td>Nonfamily Households</td>
<td>35</td>
<td>6.3%</td>
</tr>
<tr>
<td>Households with Children</td>
<td>10</td>
<td>1.8%</td>
</tr>
<tr>
<td>Unrelated Partner Households</td>
<td>2</td>
<td>0.4%</td>
</tr>
<tr>
<td>Male-female</td>
<td>4</td>
<td>0.7%</td>
</tr>
<tr>
<td>Same sex</td>
<td>3</td>
<td>0.5%</td>
</tr>
<tr>
<td>Average Household Size</td>
<td>2.0</td>
<td>4.4%</td>
</tr>
</tbody>
</table>

#### Family Households by Size

<table>
<thead>
<tr>
<th>Size</th>
<th>2010</th>
<th>Annual Rate</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total</td>
<td>558</td>
<td>100.0%</td>
</tr>
<tr>
<td>2 People</td>
<td>27</td>
<td>4.9%</td>
</tr>
<tr>
<td>3 People</td>
<td>46</td>
<td>8.3%</td>
</tr>
<tr>
<td>4 People</td>
<td>26</td>
<td>4.7%</td>
</tr>
<tr>
<td>5 People</td>
<td>21</td>
<td>3.8%</td>
</tr>
<tr>
<td>6 People</td>
<td>15</td>
<td>2.7%</td>
</tr>
<tr>
<td>7+ People</td>
<td>10</td>
<td>1.8%</td>
</tr>
<tr>
<td>Average Family Size</td>
<td>3.0</td>
<td>4.4%</td>
</tr>
</tbody>
</table>

#### Nonfamily Households by Size

<table>
<thead>
<tr>
<th>Size</th>
<th>2010</th>
<th>Annual Rate</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total</td>
<td>280</td>
<td>100.0%</td>
</tr>
<tr>
<td>1 Person</td>
<td>250</td>
<td>56.8%</td>
</tr>
<tr>
<td>2 People</td>
<td>26</td>
<td>5.8%</td>
</tr>
<tr>
<td>3 People</td>
<td>14</td>
<td>3.1%</td>
</tr>
<tr>
<td>4 People</td>
<td>11</td>
<td>2.5%</td>
</tr>
<tr>
<td>5 People</td>
<td>8</td>
<td>1.8%</td>
</tr>
<tr>
<td>6 People</td>
<td>5</td>
<td>1.1%</td>
</tr>
<tr>
<td>7+ People</td>
<td>5</td>
<td>1.1%</td>
</tr>
<tr>
<td>Average Nonfamily Size</td>
<td>2.9</td>
<td>4.4%</td>
</tr>
</tbody>
</table>

#### Population by Relationship and Household Type

<table>
<thead>
<tr>
<th>Relationship</th>
<th>2010</th>
<th>Annual Rate</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total</td>
<td>658</td>
<td>100.0%</td>
</tr>
<tr>
<td>In Households</td>
<td>558</td>
<td>85.0%</td>
</tr>
<tr>
<td>In Family Households</td>
<td>120</td>
<td>18.3%</td>
</tr>
<tr>
<td>Householder</td>
<td>120</td>
<td>18.3%</td>
</tr>
<tr>
<td>Spouse</td>
<td>120</td>
<td>18.3%</td>
</tr>
<tr>
<td>Child</td>
<td>120</td>
<td>18.3%</td>
</tr>
<tr>
<td>Other relative</td>
<td>120</td>
<td>18.3%</td>
</tr>
<tr>
<td>Nonrelative</td>
<td>120</td>
<td>18.3%</td>
</tr>
<tr>
<td>In Group Quarters</td>
<td>300</td>
<td>45.5%</td>
</tr>
<tr>
<td>Institutionalized Population</td>
<td>0</td>
<td>0.0%</td>
</tr>
<tr>
<td>Noninstitutionalized Population</td>
<td>658</td>
<td>99.9%</td>
</tr>
</tbody>
</table>

---

**Smarter Public Health Prevention Systems**
• After data cleaning, variables entered into Azure Machine Learning to run predictions
• Many AutoML experiments created
• Target: opioid activations by EMS. Total overdose case activations by EMS
• Classification machine learning models
- Variables cleaned and filtered. Removed based on overfitting and imbalanced data

![Azure Machine Learning](image.png)

<table>
<thead>
<tr>
<th>Details</th>
<th>Data cleaning</th>
<th>Model</th>
<th>Outputs</th>
<th>Logs</th>
<th>Metrics</th>
<th>Snapshots</th>
</tr>
</thead>
<tbody>
<tr>
<td>Type</td>
<td>Status</td>
<td>Description</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Validation split harvesting</td>
<td>Done</td>
<td>The input data has been split into a training dataset and a validation dataset for validation of the models. The validation dataset is generated to improve model performance.</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Class balancing detection</td>
<td>Passed</td>
<td>Your inputs were analyzed, and all classes are balanced in your training data.</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Missing feature values imputation</td>
<td>Done</td>
<td>Missing feature values were detected in your training data and imputed. If the missing values are expected, let the run complete. Otherwise cancel the current run and use a script to customize the handling of missing feature values that may be more appropriate based on the data type and business requirement.</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>High cardinality feature detection</td>
<td>Passed</td>
<td>No high cardinality features were detected.</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
RESULTS

• Second to last runs
SPARSENORMALIZER, XGBOOST CLASSIFIER
## RESULTS

- Final runs

<table>
<thead>
<tr>
<th>Algorithm name</th>
<th>Explained</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>VotingEnsemble</td>
<td>View explanation</td>
<td>0.84012</td>
</tr>
<tr>
<td>StackEnsemble</td>
<td></td>
<td>0.83637</td>
</tr>
<tr>
<td>SparseNormalizer, XGBoostClassifier</td>
<td></td>
<td>0.83388</td>
</tr>
<tr>
<td>SparseNormalizer, XGBoostClassifier</td>
<td></td>
<td>0.83388</td>
</tr>
<tr>
<td>SparseNormalizer, XGBoostClassifier</td>
<td></td>
<td>0.83388</td>
</tr>
<tr>
<td>SparseNormalizer, XGBoostClassifier</td>
<td></td>
<td>0.83341</td>
</tr>
<tr>
<td>SparseNormalizer, XGBoostClassifier</td>
<td></td>
<td>0.83326</td>
</tr>
<tr>
<td>SparseNormalizer, XGBoostClassifier</td>
<td></td>
<td>0.83326</td>
</tr>
<tr>
<td>SparseNormalizer, XGBoostClassifier</td>
<td></td>
<td>0.83326</td>
</tr>
<tr>
<td>SparseNormalizer, XGBoostClassifier</td>
<td></td>
<td>0.83232</td>
</tr>
<tr>
<td>MaxAbsScaler, LightGBM</td>
<td></td>
<td>0.83216</td>
</tr>
<tr>
<td>SparseNormalizer, XGBoostClassifier</td>
<td></td>
<td>0.83216</td>
</tr>
</tbody>
</table>
CONSIDERATIONS WITH SYNTHEA

---

C:\\Synthea\synthea>\run_synthea.bat -m "onc_opioids" -p 255000 Illinois Chicago

54084 -- Elva122 Langworth352 (25 y/o F) Chicago, Illinois
54082 -- Drucci1a444 Paucek755 (29 y/o F) Chicago, Illinois
54091 -- Coleen678 Sauer652 (17 y/o F) Chicago, Illinois
54088 -- Tracey100 Gottlieb798 (30 y/o M) Chicago, Illinois
54085 -- Luigi346 Schmeler639 (39 y/o M) Chicago, Illinois
54083 -- Kevin729 Hahn503 (54 y/o M) Chicago, Illinois
54089 -- Long300 Hammes673 (41 y/o M) Chicago, Illinois DECEASED

---

<table>
<thead>
<tr>
<th>CITY</th>
<th>STATE</th>
<th>COUNTY</th>
<th>ZIP</th>
<th>LAT</th>
<th>LON</th>
<th>HE</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>CI</td>
<td><em>DuPage County</em></td>
<td>60016</td>
<td>41.880</td>
<td>-87.916</td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>CI</td>
<td><em>DuPage County</em></td>
<td>60061</td>
<td>41.784</td>
<td>-87.806</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>CI</td>
<td><em>DuPage County</em></td>
<td>60616</td>
<td>41.904</td>
<td>-87.779</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>CI</td>
<td><em>DuPage County</em></td>
<td>60060</td>
<td>41.950</td>
<td>-87.617</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>CI</td>
<td><em>DuPage County</em></td>
<td>60667</td>
<td>41.866</td>
<td>-87.642</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>CI</td>
<td><em>DuPage County</em></td>
<td>60640</td>
<td>41.634</td>
<td>-87.670</td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>CI</td>
<td><em>DuPage County</em></td>
<td>60640</td>
<td>41.634</td>
<td>-87.670</td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>CI</td>
<td><em>DuPage County</em></td>
<td>60011</td>
<td>41.734</td>
<td>-87.695</td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>CI</td>
<td><em>DuPage County</em></td>
<td>50510</td>
<td>41.607</td>
<td>-87.528</td>
<td></td>
</tr>
<tr>
<td>9</td>
<td>CI</td>
<td><em>DuPage County</em></td>
<td>60011</td>
<td>41.772</td>
<td>-87.582</td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>CI</td>
<td><em>DuPage County</em></td>
<td>60062</td>
<td>41.786</td>
<td>-87.687</td>
<td></td>
</tr>
<tr>
<td>11</td>
<td>CI</td>
<td><em>DuPage County</em></td>
<td>60610</td>
<td>41.972</td>
<td>-87.725</td>
<td></td>
</tr>
<tr>
<td>12</td>
<td>CI</td>
<td><em>DuPage County</em></td>
<td>60017</td>
<td>41.715</td>
<td>-87.815</td>
<td></td>
</tr>
<tr>
<td>13</td>
<td>CI</td>
<td><em>DuPage County</em></td>
<td>60014</td>
<td>41.809</td>
<td>-87.609</td>
<td></td>
</tr>
<tr>
<td>14</td>
<td>CI</td>
<td><em>DuPage County</em></td>
<td>60001</td>
<td>41.701</td>
<td>-87.686</td>
<td></td>
</tr>
<tr>
<td>15</td>
<td>CI</td>
<td><em>DuPage County</em></td>
<td>60004</td>
<td>41.519</td>
<td>-87.672</td>
<td></td>
</tr>
<tr>
<td>16</td>
<td>CI</td>
<td><em>DuPage County</em></td>
<td>60610</td>
<td>41.816</td>
<td>-87.776</td>
<td></td>
</tr>
<tr>
<td>17</td>
<td>CI</td>
<td><em>DuPage County</em></td>
<td>60017</td>
<td>41.862</td>
<td>-87.762</td>
<td></td>
</tr>
<tr>
<td>18</td>
<td>CI</td>
<td><em>DuPage County</em></td>
<td>60054</td>
<td>42.026</td>
<td>-87.621</td>
<td></td>
</tr>
</tbody>
</table>

---
PROPOSED METHOD WITH SYNTHEA

• Similar 5 minute walk times
• Enrich synthetic data using USA 2020 demographic data, USA 2010 Census Demographic Data, USA 2014/2018 American Community Survey (ACS) Demographic Data, USA 2020 Consumer Expenditure data, and USA 2020 Tapestry Segmentation Data.
• Azure Machine Learning for similar classification techniques.
EXPERIENCE WITH SYNTHEA

• Machine learning results not as robust as with real 911 data (NEMSIS).
• Location data not reflective of true demographics of Chicago
• Consider integration of other U.S. Synthetic Household population data (Ex. RTI) into Synthea workflow
REFERENCES

Thank You

Phone: 202-690-7151

Health IT Feedback Form: https://www.healthit.gov/form/healthit-feedback-form

Twitter: @onc_healthIT

LinkedIn: Search “Office of the National Coordinator for Health Information Technology”

Subscribe to our weekly eblast at healthit.gov for the latest updates!