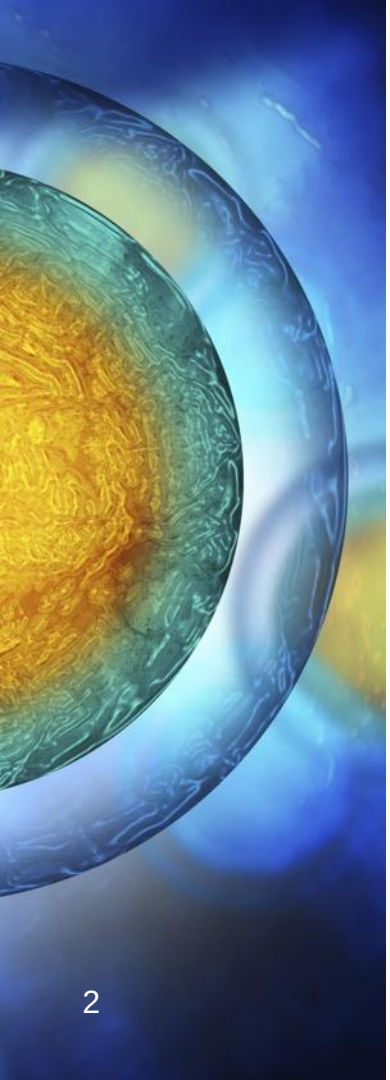
OHSU

Standardizing Electronic Health Records to Improve Medical Research and Analysis

The OMOP CDM

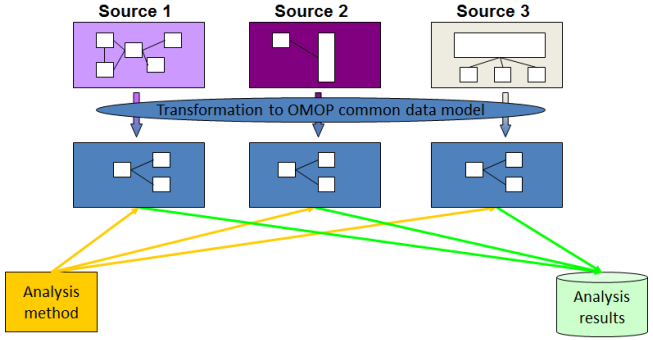
DATE: AUGUST 22, 2019 PRESENTED BY: ALYSSA HUQUE AND AARON BERGER, INTERNS



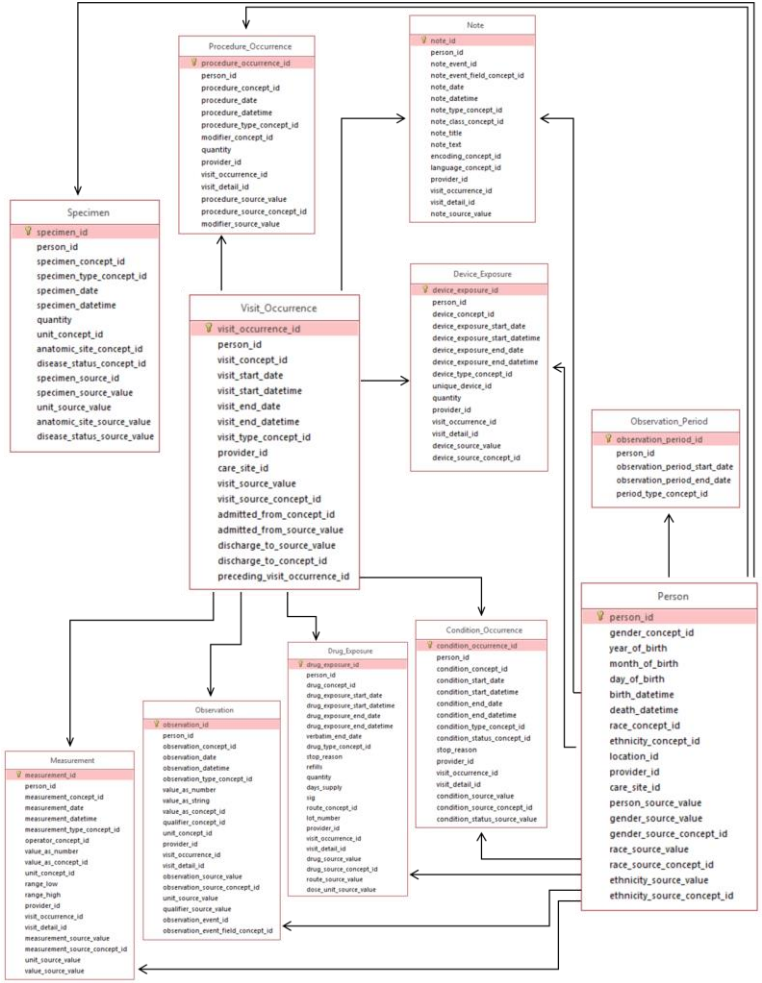
Background

- Passage of the Health and Information Technology for Economic and Clinical Health Act (HITECH Act) of American Recovery and Reinvestment Act of 2009 (ARRA – Obama's economic stimulus bill) incentivized medical institutions to digitize health records
- Inconsistencies of data captured
- Desire to standardize EHRs
  - Secondary research
  - Collaboration and compatibility with other data sets
  - Increased accuracy, reproducibility

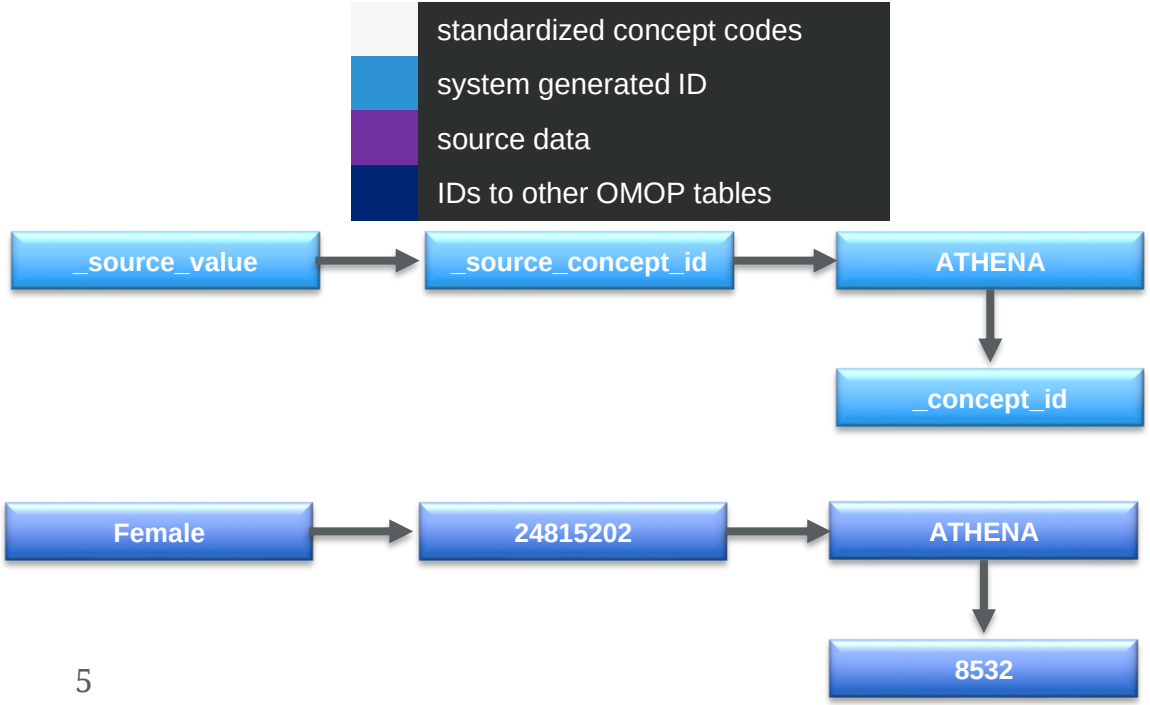
OHDSI's OMOP CDM



- Standardizes data
  - Common format
  - Common representation
- Perform systematic and standard analytics
- Support research to identify and evaluate associations
- Allows data analysis from multiple, disparate data sources
  - Handles administrative claims and EHR


Procedure_Occurrence
procedure_occurrence_id
person_id
procedure_concept_id
procedure_date
procedure_datetime
procedure_type_concept_id
modifier_concept_id
quantity
provider_id
visit_occurrence_id
visit_detail_id
procedure_source_value
procedure_source_concept_id
modifier_source_value
Note
note_id
person_id
note_event_id
note_event_field_concept_id
note_date
note_datetime
note_type_concept_id
note_class_concept_id
note_title
note_text
encoding_concept_id
language_concept_id
provider_id
visit_occurrence_id
visit_detail_id
note_source_value
Specimen
specimen_id
person_id
specimen_concept_id
specimen_type_concept_id
specimen_date
specimen_datetime
quantity
unit_concept_id
anatomic_site_concept_id
disease_status_concept_id
specimen_source_id
specimen_source_value
unit_source_value
anatomic_site_source_value
disease_status_source_value
Device_Exposure
device_exposure_id
person_id
device_concept_id
device_exposure_start_date
device_exposure_start_datetime
device_exposure_end_date
device_exposure_end_datetime
device_type_concept_id
unique_device_id
quantity
provider_id
visit_occurrence_id
visit_detail_id
device_source_value
device_source_concept_id
Visit_Occurrence
visit_occurrence_id
person_id
visit_concept_id
visit_start_date
visit_start_datetime
visit_end_date
visit_end_datetime
visit_type_concept_id
provider_id
care_site_id
visit_source_value
visit_source_concept_id
admitted_from_concept_id
admitted_from_source_value
discharge_to_source_value
discharge_to_concept_id
preceding_visit_occurrence_id
Observation_Period
observation_period_id
person_id
observation_period_start_date
observation_period_end_date
period_type_concept_id
Person
person_id
gender_concept_id
year_of_birth
month_of_birth
day_of_birth
birth_datetime
death_datetime
race_concept_id
ethnicity_concept_id
location_id
provider_id
care_site_id
person_source_value
gender_source_value
gender_source_concept_id
race_source_value
race_source_concept_id
ethnicity_source_value
ethnicity_source_concept_id
Condition_Occurrence
condition_occurrence_id
person_id
condition_concept_id
condition_start_date
condition_start_datetime
condition_end_date
condition_end_datetime
condition_type_concept_id
condition_status_concept_id
stop_reason
provider_id
visit_occurrence_id
visit_detail_id
condition_source_value
condition_source_concept_id
condition_status_source_value
Drug_Exposure
drug_exposure_id
person_id
drug_concept_id
drug_exposure_start_date
drug_exposure_start_datetime
drug_exposure_end_date
drug_exposure_end_datetime
verbatim_end_date
drug_type_concept_id
stop_reason
refills
quantity
days_supply
sig
route_concept_id
lot_number
provider_id
visit_occurrence_id
visit_detail_id
drug_source_value
drug_source_concept_id
route_source_value
dose_unit_source_value
Observation
observation_id
person_id
observation_concept_id
observation_date
observation_datetime
observation_type_concept_id
value_as_number
value_as_string
value_as_concept_id
qualifier_concept_id
unit_concept_id
provider_id
visit_occurrence_id
visit_detail_id
observation_source_value
observation_source_concept_id
unit_source_value
qualifier_source_value
observation_event_id
observation_event_field_concept_id
Measurement
measurement_id
person_id
measurement_concept_id
measurement_date
measurement_datetime
measurement_type_concept_id
operator_concept_id
value_as_number
value_as_concept_id
unit_concept_id
range_low
range_high
provider_id
visit_occurrence_id
visit_detail_id
measurement_source_value
measurement_source_concept_id
unit_source_value
value_source_value

Example OMOP Person Table



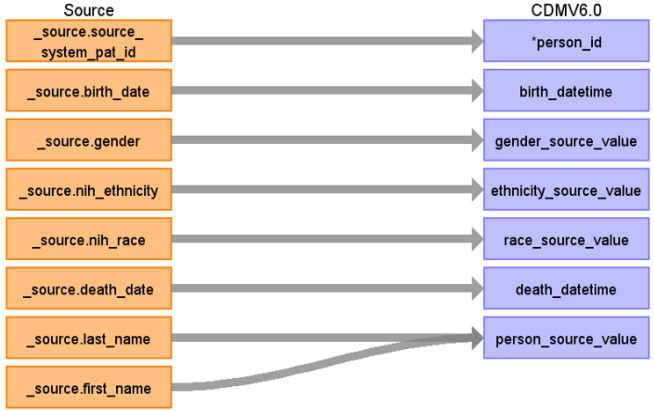
Field	Type
person_id	integer
gender_concept_id	integer
year_of_birth	integer
month_of_birth	integer
day_of_birth	integer
birth_datetime	datetime
death_datetime	datetime
race_concept_id	integer
ethnicity_concept_id	integer
location_id	integer
provider_id	integer
care_site_id	integer
person_source_value	varchar(50)
gender_source_value	varchar(50)
gender_source_concept_id	integer
race_source_value	varchar(50)
race_source_concept_id	integer
ethnicity_source_value	varchar(50)
ethnicity_source_concept_id	integer

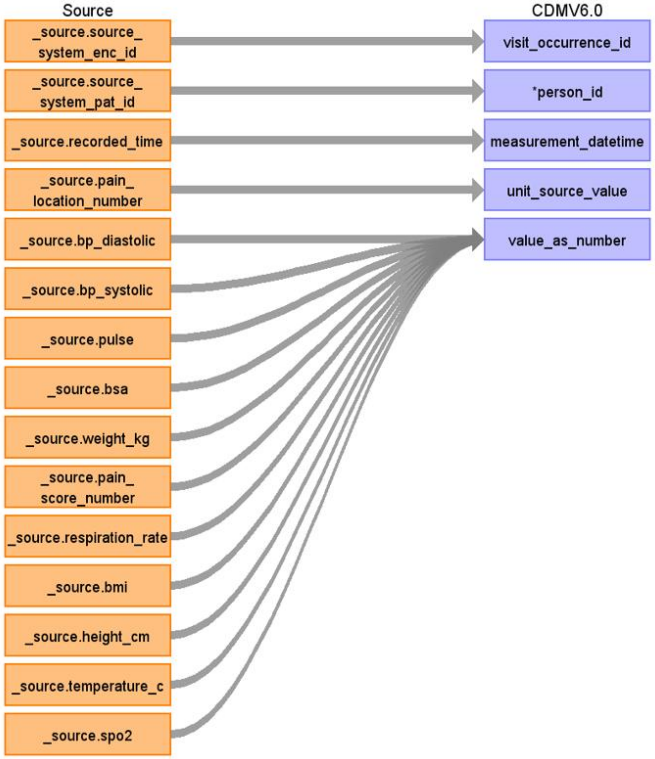
Example OHSU Demographics Table

Field	Type
OHSU_MRN	integer
SOURCE_SYSTEM_PAT_ID	integer
CURRENT_AGE_YRS	integer
BIRTH_DATE	datetime
GENDER	varchar(50)
NIH_ETHNICITY	varchar(50)
NIH_RACE	varchar(50)
PATIENT_ALIVE	varchar(50)
DEATH_DATE	datetime
LAST_NAME	varchar(50)
FIRST_NAME	varchar(50)
ADDRESS_CITY	varchar(50)
ADDRESS_LINE1	varchar(50)
ADDRESS_LINE2	varchar(50)
ADDRESS_STATE	varchar(50)
ADDRESS_ZIP	integer
ADDRESS_PHONE	integer
EMAIL	varchar(50)
ADDRESS_COUNTY	varchar(50)
CURRENT_PCP	varchar(50)
BIO_SAMPL_OPT_OUT_FLG	varchar(50)
GENETC_OPT_OUT_FLG	varchar(50)

Mapping Data

Demographics vs. Vitals





Issues with Mapping Data

- Loss of uniqueness of certain OHSU tables
- Fields can map but terminology won't

encounter_diagnosis			
BILLING_DX_FLAG	Y/N		
DX_DATE	YYYY-MM-DD	(CONDITION_OCCURRENCE) condition_start_datetime	datetime
DX_ICD	code		
DX_ICD10	code	(CONDITION_OCCURRENCE) condition_source_concept_id	varchar(50)
DX_ICD10_NAME	varchar, blank	(CONDITION_OCCURRENCE) condition_source_value	varchar(50)
DX_NAME	varchar		
ENC_DX_FLAG	Y/N		
FOLLOWUP_DX_FLAG	Y/N		
HOSP_ADMIT_FLAG	Y/N		
MED_HX_DX_FLAG	Y/N		
OHSU_MRN	integer		
ORDER_MED_DX_FLAG	Y/N		
ORDER_PROC_DX_FLAG	Y/N		
REFERRAL_DX_FLAG	Y/N		
SOURCE_SYSTEM_ENC_ID	integer	(CONDITION_OCCURRENCE) visit_occurrence_id	varchar(50)
SOURCE_SYSTEM_PAT_ID	Z + integer	(CONDITION_OCCURRENCE) person_id	integer

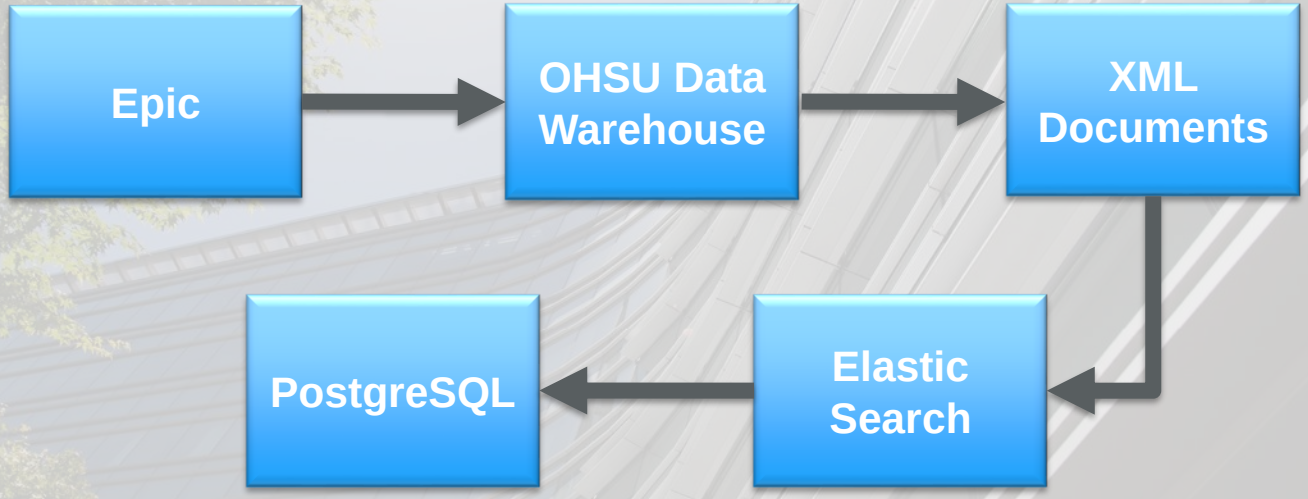
Success in Mapping Data

- 74% (198/268) source fields mapped successfully
- 62% (96/154) OMOP fields of pure data mapped successfully
  - Adding _concept_ids will increase percentages of mapped OMOP fields

Aaron - PostgreSQL

- Use PostgreSQL to create CDM tables and source tables
- Extract data from Elasticsearch and import into created source tables
- Transform source data format into CDM data format
- Load newly formatted data into CDM tables
- Implement primary, foreign keys for relationality

The Path of Aaron's Data



PostgreSQL Tables and Fields

```
-- Fields that map from 'Demographics' to 'person'
INSERT INTO public.person(person_id, person_source_value, birth_datetime, death_datetime, ger
SELECT "SOURCE_SYSTEM_PAT_ID", "FULL_NAME", "BIRTH_DATE", "DEATH_DATE", "GENDER", "NIH_RACE",
FROM public."Demographics";

--Fields that map from 'Demographics' to 'location' (Health System Data Table)
INSERT INTO public.location(address_1, address_2, city, state, zip, county)
SELECT "ADDRESS_LINE1", "ADDRESS_LINE2", "ADDRESS_CITY", "ADDRESS_STATE", "ADDRESS_ZIP", "ADD
FROM public."Demographics";

--Fields that map from 'Notes' to 'note' (Clinical Data Table)
INSERT INTO public.note(note_datetime, note_text, note_id, person_id)
SELECT "NOTE_DATE", "NOTE_TEXT", "SOURCE_SYSTEM_NOTE_CSN_ID", "SOURCE_SYSTEM_PAT_ID"
FROM source."Notes";

--Fields that map from 'Problem_List' to 'condition_occurrence' (Clinical Data Table)
INSERT INTO public.condition_occurrence(person_id, condition_start_datetime, condition_end_da
SELECT "SOURCE_SYSTEM_PAT_ID", "DX_START_DATE", "DX_END_DATE", "DX_ICD10_NAME", "DX_ICD10"
FROM source."Problem_List";
```

	birth_datetime timestamp without time zone	death_datetime timestamp without time zone	gender_source_value character varying (50)	gender_concept_id integer	race_source_value character varying (50)	race_concept_id integer
1	1955-12-25 00:00:00	9999-12-31 00:00:00	FEMALE	[null]	WHITE	[null]
2	1956-12-25 00:00:00	9999-12-31 00:00:00	FEMALE	[null]	WHITE	[null]
3	1957-12-25 00:00:00	9999-12-31 00:00:00	FEMALE	[null]	WHITE	[null]
4	2002-12-25 00:00:00	9999-12-31 00:00:00	MALE	[null]	WHITE	[null]
5	1960-12-25 00:00:00	9999-12-31 00:00:00	FEMALE	[null]	WHITE	[null]
6	1943-12-25 00:00:00	9999-12-31 00:00:00	FEMALE	[null]	WHITE	[null]
7	1963-12-25 00:00:00	9999-12-31 00:00:00	FEMALE	[null]	WHITE	[null]
8	1989-12-25 00:00:00	9999-12-31 00:00:00	FEMALE	[null]	WHITE	[null]

The Path of Alyssa's Data

Alyssa – Python

- Goal: to develop and implement a software program that maps OHSU's EHR data stored in XML files to new XML files that are formatted and organized by the standards set by the OMOP CDM.

Example OHSU EHR XML Document

Demographics


```xml
<?xml version="1.0"?>
- <main>
  - <DATA_RECORD>
      <MEDICAL_RECORD_NUMBER>00000000</MEDICAL_RECORD_NUMBER>
      <PATIENT_ID>Z0000000</PATIENT_ID>
      <CURRENT_AGE_IN_YEARS>20</CURRENT_AGE_IN_YEARS>
      <BIRTH_DATE>04/21/1999</BIRTH_DATE>
      <GENDER>FEMALE</GENDER>
      <ETHNICITY>NOT HISPANIC OR LATINO</ETHNICITY>
      <RACE>ASIAN</RACE>
      <ALIVE>Y</ALIVE>
      <DEATH_DATE>12/31/1999</DEATH_DATE>
      <LAST_NAME>HUQUE</LAST_NAME>
      <FIRST_NAME>ALYSSA</FIRST_NAME>
      <CITY>WASHINGTON</CITY>
      <ADDRESS_LINE1>1600 PENNSYLVANIA AVENUE</ADDRESS_LINE1>
      <ADDRESS_LINE2>N/A</ADDRESS_LINE2>
      <STATE>DC</STATE>
      <ZIP>20500</ZIP>
      <PHONE>2024561414</PHONE>
      <EMAIL>ALYHUQUE&#64GMAIL.COM</EMAIL>
      <COUNTY>WASHINGTON DC</COUNTY>
      <PRIMARY_CARE_PROVIDER>HERSH, WILLIAM R.</PRIMARY_CARE_PROVIDER>
      <BIOLOGY_SAMPLE_OPT_OUT>N</BIOLOGY_SAMPLE_OPT_OUT>
      <GENETIC_OPT_OUT>N</GENETIC_OPT_OUT>
    </DATA_RECORD>
  - <DATA_RECORD>
      <MEDICAL_RECORD_NUMBER>11111111</MEDICAL_RECORD_NUMBER>
      <PATIENT_ID>Z1111111</PATIENT_ID>
      <CURRENT_AGE_IN_YEARS>15</CURRENT_AGE_IN_YEARS>
      <BIRTH_DATE>09/05/2003</BIRTH_DATE>
      <GENDER>FEMALE</GENDER>
      <ETHNICITY>NOT HISPANIC OR LATINO</ETHNICITY>
      <RACE>ASIAN</RACE>
      <ALIVE>Y</ALIVE>
      <DEATH_DATE>12/31/1999</DEATH_DATE>
      <LAST_NAME>HUQUE</LAST_NAME>
      <FIRST_NAME>ALAYNA</FIRST_NAME>
      <CITY>WASHINGTON</CITY>
      <ADDRESS_LINE1>1600 PENNSYLVANIA AVENUE</ADDRESS_LINE1>
      <ADDRESS_LINE2>N/A</ADDRESS_LINE2>
      <STATE>DC</STATE>
      <ZIP>20500</ZIP>
      <PHONE>2024561414</PHONE>
      <EMAIL>ALAHUQUE&#64GMAIL.COM</EMAIL>
      <COUNTY>WASHINGTON DC</COUNTY>
      <PRIMARY_CARE_PROVIDER>HERSH, WILLIAM R.</PRIMARY_CARE_PROVIDER>
      <BIOLOGY_SAMPLE_OPT_OUT>N</BIOLOGY_SAMPLE_OPT_OUT>
      <GENETIC_OPT_OUT>N</GENETIC_OPT_OUT>
    </DATA_RECORD>
  </main>
```

Python Program

1. Parse the OHSU XML source files
2. Store the EHR patient data into a list
3. Add all lists to a queue
4. Remove one list at a time from the queue
5. Index the list and assign the information to the OMOP field
6. Print a new XML file that follows the OMOP CDM standards

Steps 1 and 2

```python
def demographics(data_set):
    '''parses the xml file for the demographics tables of patients and
    stores only the information that maps to the OMOP CDM in a list'''
    base_path = os.path.dirname(os.path.realpath(__file__)) # path to this file
    xml_file = os.path.join(base_path, "C:\\Users\\huque\\Desktop\\Data\\{}".format(data_set))
    tree = et.parse(xml_file) # parses xml file
    root = tree.getroot()

    # creates a list for each DATA_RECORD
    collected_data = []
    for d in root.findall('DATA_RECORD'):
        collected_data.append([d.find(x).text for x in ['OHSU_PATIENT_ID',
            'BIRTH_DATE',
            'GENDER',
            'ETHNICITY',
            'RACE',
            'DEATH_DATE',
            'LAST_NAME',
            'FIRST_NAME',
            'CITY',
            'ADDRESS_LINE1',
            'ADDRESS_LINE2',
            'STATE',
            'ZIP',
            'COUNTY',
            'PRIMARY_CARE_PROVIDER']])

    return collected_data
```

Steps 3 and 4

```python
def q_demographics(data_set):
    '''adds the list containing data records of demographics
    to a queue and continuously removes a list and prints an OMOP PERSON
    and LOCATION XML document until the queue is empty'''
    demographics_queue = Queue()
    for i in range(len(pohsu.demographics(data_set))):
        demographics_queue.enqueue(pohsu.demographics(data_set)[i]) # adds lists to queue
    while demographics_queue.isEmpty() != True:
        data_record = demographics_queue.dequeue() # removes list from queue
        pxml.print_demographics_PERSON_LOCATION(data_record) # prints OMOP xml document
```

Step 5

```python
def demographics_PERSON_elements(root, collected_data):
    '''indexes the demographics list and maps the OHSU data fields to
    the OMOP structure for the PERSON table'''
    record = et.SubElement(root, 'DATA_RECORD')
    et.SubElement(record, "person_id").text = collected_data[0]
    et.SubElement(record, "gender_concept_id").text = None # map from Athena
    et.SubElement(record, "year_of_birth").text = collected_data[1][6:10]
    et.SubElement(record, "month_of_birth").text = collected_data[1][0:2]
    et.SubElement(record, "day_of_birth").text = collected_data[1][3:5]
    et.SubElement(record, "birth_datetime").text = collected_data[1]
    et.SubElement(record, "death_datetime").text = collected_data[5]
    et.SubElement(record, "race_concept_id").text = None # map from Athena
    et.SubElement(record, "ethnicity_concept_id").text = None # map from Athena
    et.SubElement(record, "location_id").text = None # unique ID from location table
    et.SubElement(record, "provider_id").text = None # from provider table
    et.SubElement(record, "care_site_id").text = None
    et.SubElement(record, "person_source_value").text = collected_data[7] + ' ' + collected_data[8]
    et.SubElement(record, "gender_source_value").text = collected_data[2]
    et.SubElement(record, "gender_source_concept_id").text = None
    et.SubElement(record, "race_source_value").text = collected_data[4]
    et.SubElement(record, "race_source_concept_id").text = None
    et.SubElement(record, "ethnicity_source_value").text = collected_data[3]
    et.SubElement(record, "ethnicity_source_concept_id").text = None


def demographics_LOCATION_elements(root, collected_data):
    '''indexes the demographics list and maps the OHSU data fields to
    the OMOP structure for the LOCATION table'''
    record = et.SubElement(root, 'DATA_RECORD')
    et.SubElement(record, "location_id").text = None # unique system generated ID
    et.SubElement(record, "address_1").text = collected_data[9]
    et.SubElement(record, "address_2").text = collected_data[10]
    et.SubElement(record, "city").text = collected_data[8]
    et.SubElement(record, "state").text = collected_data[11]
    et.SubElement(record, "zip").text = collected_data[12]
    et.SubElement(record, "county").text = collected_data[13]
    et.SubElement(record, "country").text = None # find from state
    et.SubElement(record, "location_source_value").text = None
    et.SubElement(record, "latitude").text = None
    et.SubElement(record, "longitude").text = None
```

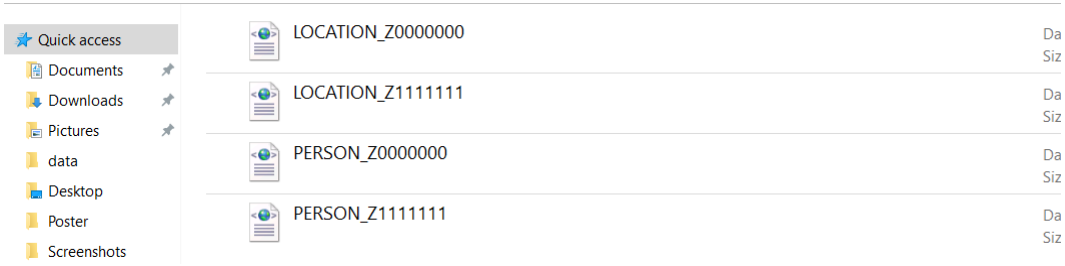
Step 6

```python
def print_demographics_PERSON_LOCATION(collected_data):
    '''takes a list of demographics from OHSU data records and
    reformats to a PERSON and LOCATION table that meets OMOP standards'''
    root = et.Element('LOCATION')
    xmlc.demographics_PERSON_elements(root, collected_data) # indexes list

    tree = et.ElementTree(root)
    tree.write(f"C:\\Users\\huque\\Desktop\\data\\PERSON_{collected_data[0]}.xml")

    root = et.Element('PERSON')
    xmlc.demographics_LOCATION_elements(root, collected_data) # indexes list

    tree = et.ElementTree(root)
    tree.write(f"C:\\Users\\huque\\Desktop\\data\\LOCATION_{collected_data[0]}.xml")
```



Example OMOP EHR XML Document

PERSON and LOCATION


```xml
<?xml version="1.0"?>
- <PERSON>
  - <DATA_RECORD>
      <person_id>Z0000000</person_id>
      <gender_concept_id/>
      <year_of_birth>1999</year_of_birth>
      <month_of_birth>04</month_of_birth>
      <day_of_birth>21</day_of_birth>
      <birth_datetime>04/21/1999</birth_datetime>
      <death_datetime>12/31/1999</death_datetime>
      <race_concept_id/>
      <ethnicity_concept_id/>
      <location_id>1</location_id>
      <provider_id/>
      <care_site_id/>
      <person_source_value>ALYSSA HUQUE</person_source_value>
      <gender_source_value>FEMALE</gender_source_value>
      <gender_source_concept_id/>
      <race_source_value>ASIAN</race_source_value>
      <race_source_concept_id/>
      <ethnicity_source_value>NOT HISPANIC OR LATINO</ethnicity_source_value>
      <ethnicity_source_concept_id/>
    </DATA_RECORD>
  </PERSON>
```

```xml
<?xml version="1.0"?>
- <LOCATION>
  - <DATA_RECORD>
      <location_id>2</location_id>
      <address_1>1600 PENNSYLVANIA AVENUE</address_1>
      <address_2>N/A</address_2>
      <city>WASHINGTON</city>
      <state>DC</state>
      <zip>20500</zip>
      <county>WASHINGTON DC</county>
      <country/>
      <location_source_value/>
      <latitude/>
      <longitude/>
    </DATA_RECORD>
  </LOCATION>
```

The Two Programs

- Similar approaches in both programs
  - Was able to map and standardize to OMOP CDM
- Programs can be used in tandem
  - Print OMOP XML documents
  - Move into database (such as SQL)
  - Implement standardized codes and relationality of tables

Future Directions

- Refine OHSU to OMOP mappings
- Implement standardized codes used in the OMOP CDM _concept_id fields
- Improve the efficiency of our programs to handle a few million data records

Funded by Grant 2T15LM007088, Research Training in Biomedical Informatics and Data Science at Oregon Health and Science University
National Library of Medicine

Special thanks to Dr. William Hersh, Dr. Aaron Cohen, Dr. Steven Bedrick, and Dr. Steven Chamberlin

Works Cited

Belenkaya, R., Blacketer, C., Hripcsak, G., Natajan, K., O'Hara, D., Reich, D., Roa, G., Torok, D., Zandt, M., Velez, M., Voss, E.. *OMOP Common Data Model and Standardized Vocabularies*. https://www.ohdsi.org/wp-content/uploads/2017/10/OHDSI-Vocabulary-CDM-Tutorial-2017-2017.10.11.pdf. PowerPoint Presentation.

Ohdsi. "OHDI/CommonDataModel." *GitHub*, github.com/OHDSI/CommonDataMode/wiki/Standardized-Clinical-Data-Tables.

"OMOP Common Data Model." *OHDSI*, www.ohdsi.org/data-standardization/the-common-data-model. Accessed 26 Jun. 2019.

"Rodriguez, Alison. "HITECH Act Resulted in Significant Gains in EHR Adoption in Hospitals." *AJMC*, 15 Aug. 2017, www.ajmc.com/newsroom/hitech-act-resulted-in-significant-gains-in-ehr-adoption-in-hospitals.