OHDSI Cohort Queries via DataKnots.jl

a preliminary sketch of a domain specific language

Clark C. Evans ⟨cce@clarkevans.com⟩,
Kyrylo Simonov ⟨xi@resolvent.net⟩

Tue, 30th July, 2019
OHDSI Community Call

Prometheus Research, LLC

Scope of Completed Research Effort

Expected Result Generation

Extracted 3 cohort queries from Atlas, ran SQL to identify 10 patients from those cohorts in SynPUF 5% dataset.

Sample Data Preparation

Created a "`sp10`" CDMv5 extraction with those 10 patients. Verified SQL generates same results on `sp10`.

Domain Specific Query Language

Defined a minimal domain specific language using `DataKnots.jl` which could cover those 3 cohort queries.

Cohort Query Translation

Translated the 3 cohort queries to the DSQL and verified they produce expected results on `sp10`.

Evaluate Query Clarity

Examine translations for readability and correctness.

Preliminary Results

Adding Data Navigation

Needed 120 lines of code (1 days) used to fill introspection gaps; e.g. a primitive for start & end dates.

Adding Interval DataType

Needed 130 lines of code (3 days) to add interval type and interval collapsing logic for final cohort step).

Adding Concept Combinators

Needed 20 lines of code (1 day) to support concept logic.

1770674, Acute myocardial infarction events

Query is 25 lines of code (or five parts totaling 40 lines).

1770675, New users of ACE inhibitors...

Seven parts, each 7 lines of code, plus 10 line assembly.

1770676, New users of Thiazide diuretics...

Seven parts, each 7 lines of code, plus 10 line assembly.

SynPUF-HCFU : 10 Patient Fair Use SynPUF Extract

Created a CDMv5 "fair use" extract from SynPUF 5% to directly support three cohort defintions.

- 1770674: Acute myocardial infarction events... 6 patients.
- 1770675: New users of ACE inhibitors... 5 patients.
- 1770676: New users of Thiazide-like diuretics... 5 patients.

This database has 723 concepts (13 SNOMED, 22 RxNorm, 22 ICD, etc.) required by 26 concepts at 10 locations over 27 visits for 10 patients, as well as 13 drug exposure and era records.

The ACE and Thiazide cohorts do not overlap. The 6 patients with AMI are evenly distributed, 3 for each of ACE and Thiazine. Each of ACE and Thiazine have 2 patients lacking a AMI. This sample lacks an example that matches neither ACE nor Thiazine.

1770674 : Myocardial Infarction (cohort)

People having any of the following: *a condition occurrence of Acute myocardial Infarction* with continuous observation of at least 0 days prior and 0 days after event index date, and limit initial events to: all events per person.

For people matching the Primary Events, include: Having all of the following criteria: *at least 1 occurrences of a visit occurrence of Inpatient or ER visit* where event starts between all days Before and 0 days After index start date and event ends between 0 days Before and all days After index start date

This cohort defintion end date will be the index event's start date plus 7 days Collapse cohort by era with a gap size of 180 days.

1770674 : Myocardial Infarction (1of2)

```
@query sp10 begin
  person.keep(it)
  collapse_intervals(180days, begin
      condition
      keep(index_date => start_date)
      keep(continuous_observation =>
            person.observation_period.
               filter(includes(index_date)).
               is0to1())
      keep(acute_visit =>
            person.visit.filter(
                concept.iscoded("Visit", "ERIP", "ER", "IP
                includes(index_date) &&
                during(continuous_observation)))
      filter(concept.iscoded("SNOMED", 22298006, 1755008)
             exists(acute_visit))
```

1770674 : Myocardial Infarction (2of2)

```
                date_interval(index_date,
                    min(index_date + 7days,
                        continuous_observation.end_date))
    end)
    myocardial_infarction_cohort =>
        { person,
          cohort_entry => start_date,
          cohort_exit => end_date}
end
#=>
   myocardial_infarction_cohort
   person  cohort_entry  cohort_exit

1  1780    2008-04-10    2008-04-17
2  30091   2009-08-02    2009-08-09
3  69985   2010-07-22    2010-07-29
...
```

1770675 : New users of ACE inhibitors (cohort 1of2)

People having any of the following: *a drug exposure of ACE inhibitors for the first time in the person's history* with continuous observation of at least 365 days prior and 0 days after event index date, and limit initial events to: *earliest event per person*.

Inclusion Criteria #1: has hypertension diagnosis in 1 yr prior to treatment, *at least 1 occurrences of a condition occurrence of Hypertensive disorder where event starts between 365 days Before and 0 days After index start date*.

Inclusion Criteria #2: has no prior antihypertensive drug exposures in medical history, *exactly 0 occurrences of a drug exposure of Hypertension drugs where event starts between all days Before and 1 days Before index start date*.

1770675 : New users of ACE inhibitors (cohort 2of2)

Inclusion Criteria #3: Is only taking ACE as monotherapy, with no concomitant combination treatments: *exactly 1 distinct occurrences of a drug era of Hypertension drugs2 where event starts between 0 days Before and 7 days After index start date*.

Limit qualifying cohort to: earliest event per person.

Custom Drug Era Exit Criteria: This strategy creates a drug era from the codes found in the specified concept set. If the index event is found within an era, the cohort end date will use the era's end date. Otherwise, it will use the observation period end date that contains the index event. Use the era end date of ACE inhibitors, allowing 30 days between exposures adding 0 days after exposure end.

Collapse cohort by era with a gap size of 0 days.

1770675 : Concept Sets

This cohort defines 3 concept sets: (a) a diagnosis of hypertension, (b) exposure to a hypertension drug, and (c) exposure to ace inhibitor.

```
@define is_hypertensive =
    iscoded("SNOMED", 38341003)
@define is_hypertension_drug =
    iscoded("RxNorm", 149, 325646, 17767, 1091643, 11170,
        644, 1202, 18867, 1520, 19484, 1808, 214354, 1998,
        2409, 2599, 3443, 49276, 3827, 298869, 83515, 4316,
        4603, 40114, 5470, 5487, 5764, 83818, 33910, 6185,
        52175, 6876, 6916, 6918, 6984, 30131, 7226, 31555,
        7417, 7435, 321064, 7973, 54552, 8332, 8629, 8787,
        35296, 9997, 73494, 37798, 38413, 38454, 10763, 697
@define is_ace_inhibitor =
    iscoded("RxNorm", 18867, 1998, 3827, 50166, 29046,
        30131, 54552, 35208, 35296, 38454)
```

1770675 : Initial Events

The initial event is described as: *People having any of the following: a drug exposure of ACE inhibitors for the first time in the person's history*

```
@define candidate_events = begin
    person.keep(it)
    first(begin
            drug_exposure
            filter(concept.is_ace_inhibitor)
            sort(start_date)
    end)
    keep(index_date => start_date)
end
```

1770675 : Continuous Observation

The query continues: *with continuous observation of at least 365 days prior and 0 days after event index date, and limit initial events to: earliest event per person*

This can be written by creating a new interval from the index date which includes the prior 365 days.

```
@define with_continuous_observation = begin
    keep(continuous_observation =>
          person.observation_period.
             filter(includes(
                index_date.and_previous(365days))).
             is0to1())
    filter(exists(continuous_observation))
end
```

1770675 : Intermediate Results

```
@query sp10 begin
     candidate_events
     with_continuous_observation
     { person,
       period_start => continuous_observation.start_date,
       index_date,
       period_end => continuous_observation.end_date }
 end
 #=>
    person  period_start  index_date  period_end

 1  30091   2008-02-09    2009-03-28  2010-07-20
 2  42383   2008-01-04    2009-11-06  2010-08-28
 3  69985   2008-02-07    2009-05-05  2010-11-14
 4  82328   2008-05-01    2009-08-24  2010-06-19
 5  110862  2008-01-04    2010-04-05  2010-09-13
 =#
```

1770675 : Inclusion #1

Inclusion Criteria #1: *has hypertension diagnosis in 1 yr prior to treatment having all of the following criteria: at least 1 occurrences of a condition occurrence of Hypertensive disorder where event starts between 365 days Before and 0 days After index start date*

```
@define with_hypertension_diagnoses = begin
  keep(hypertension_diagnoses =>
        person.condition.filter(
          concept.is_hypertensive &&
          start_date.during(
              index_date.and_previous(365days))))
  filter(exists(hypertension_diagnoses))
end
```

1770675 : Inclusion #2

Inclusion Criteria #2: *Has no prior antihypertensive drug exposures in medical history having all of the following criteria: exactly 0 occurrences of a drug exposure of Hypertension drugs where event starts between all days Before and 1 days Before index start date*

```
@define no_prior_antihypertensive =
    filter(!exists(
            person.drug_exposure.filter(
              concept.is_hypertension_drug &&
              start_date < index_date)))
```

1770675 : Inclusion #3

Inclusion Criteria #3: *Is only taking ACE as monotherapy, with no concomitant combination treatments having all of the following criteria: exactly 1 distinct occurrences of a drug era of Hypertension drugs where event starts between 0 days Before and 7 days After index start date*

```
@define with_monotherapy_7day_era = begin
    keep(monotherapy_7day_era =>
            person.drug_era.filter(
              concept.is_hypertension_drug &&
              start_date.during(
                  index_date.and_subsequent(7days))))
    filter(1 == count(monotherapy_7day_era))
    keep(monotherapy_7day_era =>
            monotherapy_7day_era.is1to1())
    filter(monotherapy_7day_era.concept.is_ace_inhibitor)
end
```

1770675 : Intermediate Results

```
@query sp10 begin
    candidate_events
    with_continuous_observation
    with_hypertension_diagnoses
    no_prior_antihypertensive
    with_monotherapy_7day_era
    { person, index_date,
      era_start => monotherapy_7day_era.start_date,
      era_end => monotherapy_7day_era.end_date }
end
#=>
```

	person	index_date	era_start	era_end
1	30091	2009-03-28	2009-03-28	2009-04-27
2	42383	2009-11-06	2009-11-06	2009-12-06
3	69985	2009-05-05	2009-05-05	2009-06-04
4	82328	2009-08-24	2009-08-24	2009-09-23

1770675 : Custom Era (1of2)

Custom Cohort Strategy: *This strategy creates a drug era from the codes found in the specified concept set. If the index event is found within an era, the cohort end date will use the era's end date. Otherwise, it will use the observation period end date that contains the index event. Use the era end date of ACE inhibitors: allowing 30 days between exposures, adding 0 days after exposure end. Then, collapse cohort by era with a gap size of 0 days*

The logic specified in the textual description above is somewhat unclear, by reverse engineering the SQL source, we come up with something like...

1770675 : Custom Era (2of2)

```
@define with_custom_era =
    keep(custom_era => begin
          person.drug_exposure
          filter((concept.is_ace_inhibitor ||
                  source_concept.is_ace_inhibitor) &&
                  start_date >= index_date)
          { start_date,
            end_date => coalesce(end_date,
                                 start_date + days_supply,
                                 start_date + 1days) }
          collapse_intervals(30days)
          first()
     end)
```

1770675 : Final Query Result

```
@query sp10 begin
    candidate_events
    with_continuous_observation
    with_hypertension_diagnoses
    no_prior_antihypertensive
    with_monotherapy_7day_era
    with_custom_era
    { person, cohort_enter_date => custom_era.start_date,
      cohort_exit_date => custom_era.end_date }
end
#=>
```

	person	cohort_enter_date	cohort_exit_date
1	30091	2009-03-28	2009-04-27
2	42383	2009-11-06	2009-12-06
3	69985	2009-05-05	2009-06-04
4	82328	2009-08-24	2009-09-23
5	110862	2010-04-05	2010-05-05

Observations of Queries

1. Path-like semantics, navigating towards information
2. Queries are incrementally constructed & tested
3. Intermediate operations can be defined & named
4. Common notions can be named and reused
5. Semantics of each operator can be clearly defined

Scope of Next Phase Research Effort

Documentation & Regression Tests

Document OHDSI combinators, add edgecase data to `sp10` and corresponding tests to verify accuracy.

Create Atlas OHDSI Combinators

Define, document and test higher-level combinators that exactly match Atlas screens, create queries from JSON definition, and a web-service implementation.

Polish Temporal Functions

Review, test, document, and package temporal functions to be compatible with the Clinical Quality Language.

Experiment /w Cyclops

Wrap Cyclops with `CXX.jl` and see how higher level queries could be integrated with both data & statistics.

Experiment /w ETL

This framework could be used for ETL, to wrap FHIR (Synthea) and other data sources to build CDM.

Unscheduled efforts

Create SQL "push down" mechanism
: Currently, `DataKnots.jl` pulls back data cell-by-cell, this is our fallback. We need to convert query operators into SQL code that is validated.

Support various database backends
: Depending upon backend, generate SQL custom to that service, for Redshift, etc. Potentially needing low-level Julia adapters since some databases arn't yet supported.

Custom push-downs for OHDSI
: To be converted to SQL, temporal operators will need to have backend-specific push-down to 'INTERVAL' and other types. Even so, the Julia fallback will work.

Integration /w Atlas, etc. Besides documentation, testing, and others, we'll need to ensure that Atlas and document `JuliaCall` so that this could be used from "R" programs.