Bostat 218 Problem Set 1

Due Feb 07 @ 11:59PM in PDF by email

Adolfo Jacobo 006333111

January 23, 2025

R Setup

```r
## Load libraries
library(DatabaseConnector)

## Clean Environment
rm(list = ls())

## Force garbage collection
gc()
```

```
##           used (Mb) gc trigger (Mb) max used (Mb)
## Ncells  745283 39.9    1266861 67.7  1266861 67.7
## Vcells 1356867 10.4    8388608 64.0  2123768 16.3
```

```r
## Set file path for synthetic database
absoluteFileName <- file.path(getwd(), "../data", "synthetic.duckdb")
```

OMOP CDM

1. John is an African American man born on August 4, 1974. Define an entry in the `PERSON` table that encodes this information.

Column name	Value	Explanation
PERSON_ID	1	Unique identifier
GENDER_CONCEPT_ID	8507	Value for 8507 refers to “Male”.
YEAR_OF_BIRTH	1974	Year of birth
MONTH_OF_BIRTH	8	Month of birth
DAY_OF_BIRTH	4	Day of birth
BIRTH_DATETIME	1974-08-04 00:00:00	When time of birth is is unknown value will default to midnight.
RACE_CONCEPT_ID	8516	Value for 8516 refers to “Black or African American”.
ETHNICITY_CONCEPT_ID	38003564	Value for 38003564 refers to “Not hispanic”.

Column name	Value	Explanation
LOCATION_ID		Address of the person is not known
PROVIDER_ID		PCP is not known
CARE_SITE_ID		PCP site is not known
PERSON_SOURCE_VALUE		Value is not known
GENDER_SOURCE_VALUE	Male	Text description of GENDER_CONCEPT_ID
GENDER_SOURCE_CONCEPT_ID	0	GENDER_CONCEPT_ID from source data. When this is not known, value will default to 0.
RACE_SOURCE_VALUE	African American	Text description of RACE_CONCEPT_ID
RACE_SOURCE_CONCEPT_ID	0	RACE_CONCEPT_ID from source data. When this is not known, value will default to 0.
ETHNICITY_SOURCE_VALUE	Not hispanic	Text description of ETHNICITY_CONCEPT_ID.
ETHNICITY_SOURCE_CONCEPT_ID	0	EHTNICITY_CONCEPT_ID from source data. When this is not known, value will default to 0.

2. John enrolled in his current insurance on January 1st, 2015. The data from his insurance database were extracted on July 1st, 2019. Define an entry in the `OBSERVATION_PERIOD` table that encodes this information.

Column name	Value	Explanation
OBSERVATION_PERIOD_ID	1	Unique identifier
PERSON_ID	1	Unique identifier
OBSERVATION_PERIOD_START_DATE	2015-01-01 00:00:00	Start date of observation period
OBSERVATION_PERIOD_END_DATE	2019-07-01 00:00:00	End date of observation period
PERIOD_TYPE_CONCEPT_ID	44814722	Value for 44814722 refers to "Period while enrolled in insurance".
PERIOD_TYPE_SOURCE_VALUE	Period while enrolled in insurance	Text description of PERIOD_TYPE_CONCEPT_ID
PERIOD_TYPE_SOURCE_CONCEPT_ID	0	PERIOD_TYPE_CONCEPT_ID from source data. When this is not known, value will default to 0.

3. John was prescribed a 30-day supply of Ibuprofen 200 MG Oral tablets (NDC code: 76168009520) on May 1st, 2019. Define an entry in the `DRUG_EXPOSURE` table that encodes this information.

Column name	Value	Explanation
DRUG_EXPOSURE_ID	1	Unique identifier
PERSON_ID	1	Unique identifier
DRUG_CONCEPT_ID	19078461	Value for 19078461 refers to "Ibuprofen 200 MG Oral Tablet".
DRUG_EXPOSURE_START_DATE	2019-05-01	Start date of drug exposure
DRUG_EXPOSURE_START_DATETIME	2019-05-31 00:00:00	Start date and time of drug exposure
DRUG_EXPOSURE_END_DATE	2019-05-31	End date of drug exposure

Column name	Value	Explanation
DRUG_EXPOSURE_END_DATETIME	2019-05-31 00:00:00	End date and time of drug exposure
VERBATIM_END_DATE	2019-05-31	End date of drug exposure as it appears in the source data
DRUG_TYPE_CONCEPT_ID	38000175	Value for 38000175 refers to “Prescription dispensed in pharmacy”.
STOP_REASON		Reason for stopping drug exposure
REFILLS	0	Number of refills allowed
QUANTITY	30	Quantity of drug exposure
DAYS_SUPPLY	30	Days supply of drug exposure
SIG	Take 1 tablet by mouth once daily	Instructions for taking the drug
ROUTE_CONCEPT_ID	0	Value for 0 refers to “Unknown”.
LOT_NUMBER		Lot number of the drug
PROVIDER_ID		Prescribing provider
VISIT_OCCURRENCE_ID		Visit occurrence
VISIT_DETAIL_ID		Visit detail
DRUG_SOURCE_VALUE	76168009520	NDC code of the drug
DRUG_SOURCE_CONCEPT_ID	583945	DRUG_CONCEPT_ID from source data. When this is not known, value will default to 0.
ROUTE_SOURCE_VALUE	0	ROUTE_SOURCE_VALUE from source data. When this is not known, value will default to 0.
DOSE_UNIT_SOURCE_VALUE	0	DOSE_UNIT_SOURCE_VALUE from source data. When this is not known, value will default to 0.

4. Using SQL and R, retrieve all records of the condition “Gastrointestinal hemorrhage” (with concept ID 192671) from the `Eunomia` dataset.

```r
# Using Eunomia -- will download with each R session
connection <- connect(Eunomia::getEunomiaConnectionDetails())
```

```
## attempting to download GiBleed
```

```
## attempting to extract and load: C:\Users\ajaco\AppData\Local\Temp\RtmpGsAC9w/GiBleed_5.3.zip to: C:\
```

```
## Connecting using SQLite driver
```

```r
# Get list of tables
# getTableNames(connection,databaseSchema = 'main')

querySql(connection = connection,
         sql = "
         SELECT *
         FROM concept
         WHERE CONCEPT_ID = 192671;
         ")
```

```
##   CONCEPT_ID                CONCEPT_NAME DOMAIN_ID VOCABULARY_ID
## 1     192671 Gastrointestinal hemorrhage Condition        SNOMED
##   CONCEPT_CLASS_ID STANDARD_CONCEPT CONCEPT_CODE VALID_START_DATE
## 1 Clinical Finding                S     74474003       1970-01-01
##   VALID_END_DATE INVALID_REASON
## 1     2099-12-31           <NA>
```

```
disconnect(connection)
```

5. Using SQL and R, retrieve all records of the condition "Gastrointestinal hemorrhage" using source codes. This database uses ICD-10, and the relevant ICD-10 code is "K92.2" from the `Eunomia` dataset.

```
# Using Eunomia -- will download with each R session
connection <- connect(Eunomia::getEunomiaConnectionDetails())
```

```
## attempting to download GiBleed
```

```
## attempting to extract and load: C:\Users\ajaco\AppData\Local\Temp\RtmpGsAC9w/GiBleed_5.3.zip to: C:\
```

```
## Connecting using SQLite driver
```

```
# Get list of tables
# getTableNames(connection,databaseSchema = 'main')

querySql(connection = connection,
         sql = "
         SELECT *
         FROM concept
         WHERE CONCEPT_CODE = 'K92.2';
         ")
```

```
##   CONCEPT_ID                             CONCEPT_NAME DOMAIN_ID VOCABULARY_ID
## 1   35208414 Gastrointestinal hemorrhage, unspecified Condition       ICD10CM
##      CONCEPT_CLASS_ID STANDARD_CONCEPT CONCEPT_CODE VALID_START_DATE
## 1 4-char billing code             <NA>        K92.2       2007-01-01
##   VALID_END_DATE INVALID_REASON
## 1     2099-12-31           <NA>
```

```
disconnect(connection)
```

6. Using SQL and R, retrieve the observation period of the person with `PERSON_ID` 61 from the `Eunomia` dataset.

```
# Using Eunomia -- will download with each R session
connection <- connect(Eunomia::getEunomiaConnectionDetails())
```

```
## attempting to download GiBleed
```

```
## attempting to extract and load: C:\Users\ajaco\AppData\Local\Temp\RtmpGsAC9w/GiBleed_5.3.zip to: C:\
```

```
## Connecting using SQLite driver
```

```
# Get list of tables
# getTableNames(connection,databaseSchema = 'main')

querySql(connection = connection,
         sql = "
         SELECT *
         FROM observation_period
         WHERE PERSON_ID = 61;
         ")
```

```
##   OBSERVATION_PERIOD_ID PERSON_ID OBSERVATION_PERIOD_START_DATE
## 1                    61        61                    1968-01-21
##   OBSERVATION_PERIOD_END_DATE PERIOD_TYPE_CONCEPT_ID
## 1                  2019-01-06               44814724
```

```
disconnect(connection)
```

Standardize vocabularies

7. What is the standard concept ID for "Gastrointestinal hemorrhage"?

- The standard concept ID for "Gastrointestinal hemorrhage" is `192671`.

8. Which ICD-10CM codes map to the standard concept for "Gastrointestinal hemorrhage"? Which ICD-9CM codes map to this Standard Concept?

- The ICD-10CM codes that map to the standard concept for "Gastrointestinal hemorrhage" are `K92.2` and `K92.9`. The ICD-9CM codes that map to this standard concept are `578.9` and `578.0`.

~~9. What are the MedDRA preferred terms that are equivalent to the standard concept for "Gastrointestinal hemorrhage"?~~

Advanced SQL

10. What is the minimum, maximum, and mean length (in days) of observation from the `synthetic` dataset? (Hint: you can use the `DATEDIFF` function to compute the time between two dates.)

```
syn_connection <- connect(dbms = "duckdb", server = absoluteFileName)
```

```
## Connecting using DuckDB driver
```

```
querySql(syn_connection,
         sql = "SELECT MIN(DATEDIFF('day', OBSERVATION_PERIOD_START_DATE, OBSERVATION_PERIOD_END_DATE))
                     , MAX(DATEDIFF('day', OBSERVATION_PERIOD_START_DATE, OBSERVATION_PERIOD_END_DATE))
                     , AVG(DATEDIFF('day', OBSERVATION_PERIOD_START_DATE, OBSERVATION_PERIOD_END_DATE))
                FROM OBSERVATION_PERIOD;")
```

```
##   MIN_OBSERVATION_PERIOD_START_DATE MAX_OBSERVATION_PERIOD_END_DATE
## 1                                 0                           40509
##   AVG_OBSERVATION_DAYS
## 1             13683.69
```

```
disconnect(syn_connection)
```

11. How many people have at least one prescription of celecoxib from the `synthetic` dataset? (Note: there's an easy way to do this, using `DRUG_ERA`, and a harder way using `DRUG_EXPOSURE` and `CONCEPT_ANCESTOR`. Can you do both?)

```
syn_connection <- connect(dbms = "duckdb", server = absoluteFileName)
```

```
## Connecting using DuckDB driver
```

```
querySql(syn_connection,
         sql = "SELECT COUNT(DISTINCT PERSON_ID) AS TOTAL_CELECOXIB_PRESCRIPTIONS
                FROM DRUG_ERA de
                LEFT JOIN CONCEPT c ON de.DRUG_CONCEPT_ID = c.CONCEPT_ID
                WHERE LOWER(c.CONCEPT_NAME) LIKE '%cele%';")
```

```
##   TOTAL_CELECOXIB_PRESCRIPTIONS
## 1                             0
```

```
disconnect(syn_connection)
```

```
syn_connection <- connect(dbms = "duckdb", server = absoluteFileName)
```

```
## Connecting using DuckDB driver
```

```
querySql(syn_connection,
         sql = "SELECT COUNT(DISTINCT PERSON_ID) AS TOTAL_CELECOXIB_PRESCRIPTIONS
                FROM DRUG_EXPOSURE a
                LEFT JOIN CONCEPT_ANCESTOR b ON a.DRUG_CONCEPT_ID = b.DESCENDANT_CONCEPT_ID
                LEFT JOIN CONCEPT c ON b.ANCESTOR_CONCEPT_ID = c.CONCEPT_ID
                WHERE LOWER(c.CONCEPT_NAME) LIKE '%celecoxib%';")
```

```
##   TOTAL_CELECOXIB_PRESCRIPTIONS
## 1                             0
```

```
disconnect(syn_connection)
```

12. During which period in time (calender start and end date) did people start a celecoxib prescription from the `synthetic` dataset?

```
syn_connection <- connect(dbms = "duckdb", server = absoluteFileName)
```

```
## Connecting using DuckDB driver
```

```r
querySql(syn_connection,
         sql = "SELECT MIN(DRUG_ERA_START_DATE) AS MIN_CELECOXIB_PRESCRIPTION_DATE
                     , MAX(DRUG_ERA_END_DATE) AS MAX_CELECOXIB_PRESCRIPTION_DATE
                FROM DRUG_ERA
                WHERE DRUG_CONCEPT_ID = 1118084;")
```

```
##   MIN_CELECOXIB_PRESCRIPTION_DATE MAX_CELECOXIB_PRESCRIPTION_DATE
## 1                            <NA>                            <NA>
```

```r
disconnect(syn_connection)
```