

# Exploring MIMIC-III Critical Care Data Using Postgres/tidyverse/dbplyr/dplyr







Earl F Glynn

**Kansas City R Users Group**



2019-01-12

# Files on GitHub

<https://github.com/EarlGlynn/MIMIC-III-Getting-Started>

 <a href="#">000-Download-Files</a>	
 <a href="#">010-Count-Characters</a>	Quality Checks
 <a href="#">020-Count-Lines-Fields-Records</a>	
 <a href="#">040-Load-MIMIC-into-PostgreSQL</a>	Loading database
 <a href="#">050-Querying-MIMIC-III</a>	SQL / dplyr example
 <a href="#">060-Nature-Scientific-Data</a>	Jupyter notebook example

<https://github.com/EarlGlynn/MIMIC-III-Exploration>

 <a href="#">admissions</a>	Database tables
 <a href="#">diagnoses_icd</a>	
 <a href="#">patients</a>	

Future: MIMIC-III-Predictive-Analytics

# MIMIC-III Critical Care Data Outline

- Getting Started with MIMIC-III
  - Background / Motivation
  - Training Requirements
  - Loading Postgres Database
  - Querying MIMIC-III with SQL or dplyr
- MIMIC-III Exploration with tidyverse/dplyr
  - Patients Table
  - Admissions Table
  - Diagnoses Tables
- Take Home

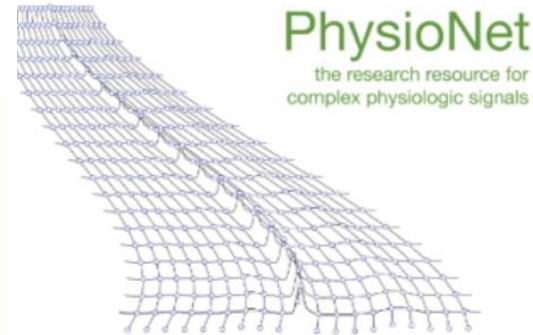
# Getting Started with MIMIC-III

## Background / Motivation

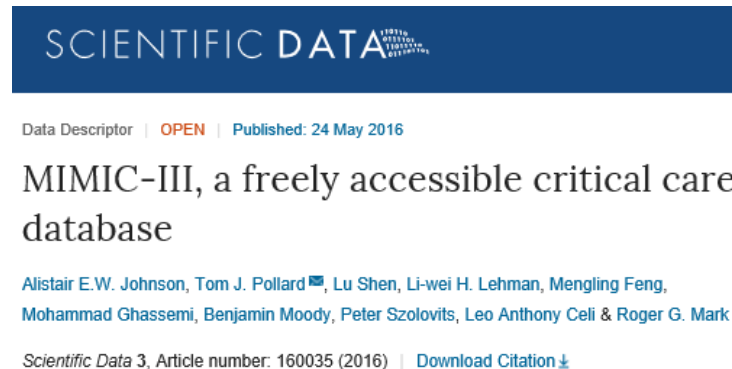
- <https://mimic.physionet.org/>

### Collaborative research

MIMIC is an openly available dataset developed by the MIT Lab for Computational Physiology, comprising deidentified health data associated with ~40,000 critical care patients. It includes demographics, vital signs, laboratory tests, medications, and more.



- <https://www.nature.com/articles/sdata201635>



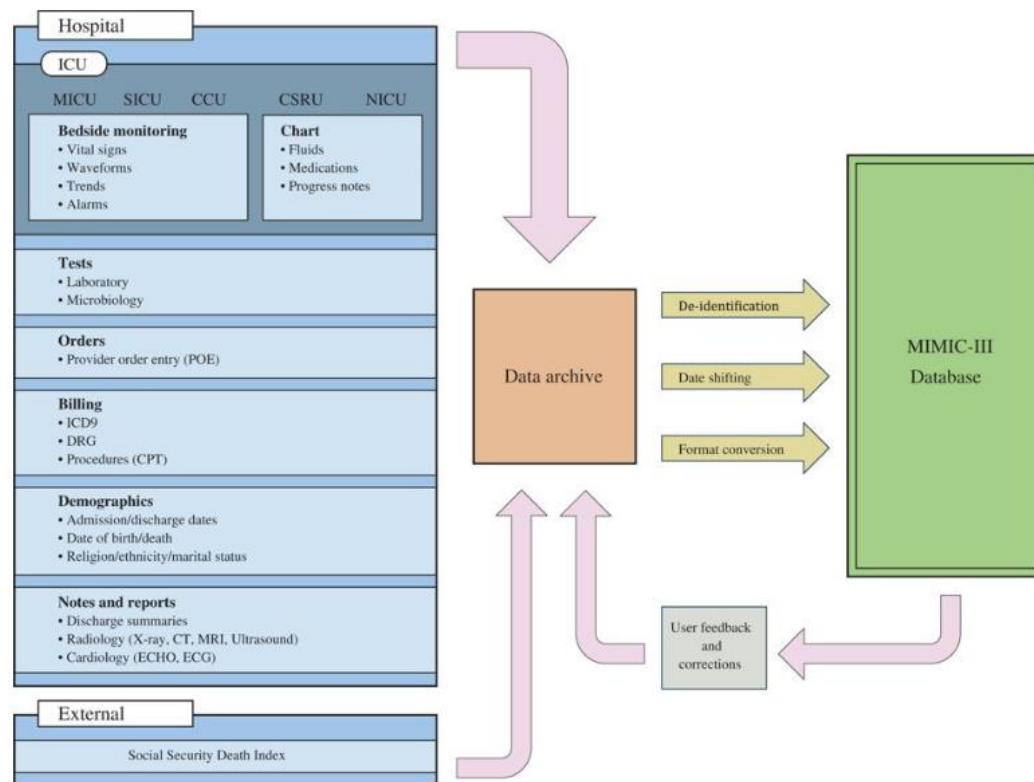
- MIMIC-III supports applications including academic and industrial research, quality improvement initiatives, and higher education coursework.
- MIMIC-III is great source for data science experiments, including predictive analytics.

# Getting Started with MIMIC-III Background / Motivation

- <https://www.nature.com/articles/sdata201635>

Figure 1 : Overview of the MIMIC-III critical care database.

From: MIMIC-III, a freely accessible critical care database



# Getting Started with MIMIC-III

## Training Requirements

- <https://mimic.physionet.org/gettingstarted/access/>

### Complete the required training course

Prior to requesting access to MIMIC, you will need to complete the CITI "Data or Specimens Only Research" course:

- First register on the CITI program website, selecting "Massachusetts Institute of Technology Affiliates" as your organization affiliation (**not** "independent learner"):  
<https://www.citiprogram.org/index.cfm?pageID=154&icat=0&ac=0>
- Follow the links to add a Massachusetts Institute of Technology Affiliates course. In the Human Subjects training category, select the "Data or Specimens Only Research" course
- Complete the course and save a copy of your completion report. The completion report lists all modules completed, with dates and scores.

## COLLABORATIVE INSTITUTIONAL TRAINING INITIATIVE (CITI PROGRAM)

### REQUIRED AND ELECTIVE MODULES ONLY

History and Ethics of Human Subjects Research (ID: 498)  
Basic Institutional Review Board (IRB) Regulations and Review Process (ID: 2)  
Records-Based Research (ID: 5)  
Genetic Research in Human Populations (ID: 6)  
Populations in Research Requiring Additional Considerations and/or Protections (ID: 16680)  
Research and HIPAA Privacy Protections (ID: 14)  
Conflicts of Interest in Human Subjects Research (ID: 17464)  
Massachusetts Institute of Technology (ID: 1290)

3 to 7  
quiz questions  
per module

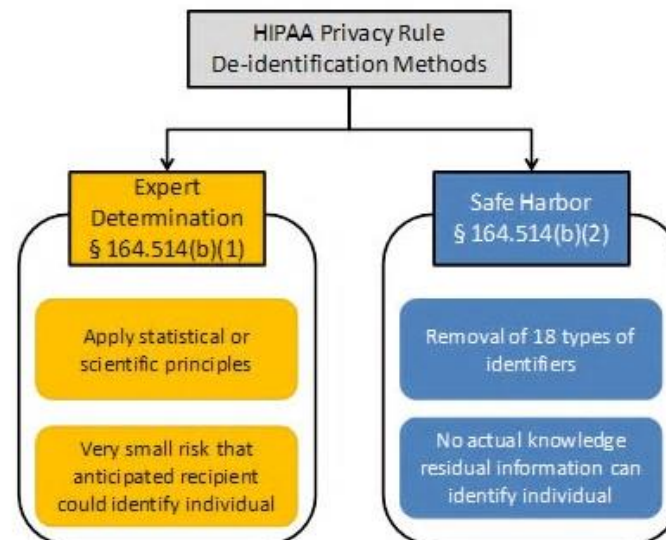
# Getting Started with MIMIC-III

## Training Requirements

### Purpose of Training: Protected Health Information (PHI)

<https://www.hhs.gov/hipaa/for-professionals/privacy/special-topics/de-identification/index.html>

### Guidance Regarding Methods for De-identification of Protected Health Information in Accordance with the Health Insurance Portability and Accountability Act (HIPAA) Privacy Rule



# Getting Started with MIMIC-III

## Loading Postgres Database

- <https://mimic.physionet.org/gettingstarted/dbsetup/>
- 000-Download-Files/DownloadMIMIC-III.nb.html

### Download MIMIC-III Files

efg | 2018-07-03

Start at [PhysioNet web page](#) and select [MIMIC-III Database](#).

Select **Data** and complete all the requirements to access MIMIC-III.

Once you have a username and password, run this script ...

- Quality checks
  - 010-Count-Characters
  - 020-Count-Lines-Fields-Records/MIMIC-III-Will-Files-Parse.nb.html
  - 030-Files-Fields-MetaStats



# Getting Started with MIMIC-III

## Loading Postgres Database

MIMIC-III-Will-Files-Parse.nb.html

Filename	Lines	Records	Fields
ADMISSIONS.csv	58,977	58,977	19
CALLOUT.csv	34,500	34,500	24
CAREGIVERS.csv	7,568	7,568	4
CHARTEVENTS.csv	330,712,484	330,712,484	15
CPTEVENTS.csv	573,147	573,147	12
D_CPT.csv	135	135	9
D_ICD_DIAGNOSES.csv	14,568	14,568	4
D_ICD_PROCEDURES.csv	3,883	3,883	4
D_ITEMS.csv	12,488	12,488	10
D_LABITEMS.csv	754	754	6
DATETIMEEVENTS.csv	4,485,938	4,485,938	14
DIAGNOSES_ICD.csv	651,048	651,048	5
DRGCODES.csv	125,558	125,558	8

Filename	Lines	Records	Fields
ICUSTAYS.csv	61,533	61,533	12
INPUTEVENTS_CV.csv	17,527,936	17,527,936	22
INPUTEVENTS_MV.csv	3,618,992	3,618,992	31
LABEVENTS.csv	27,854,056	27,854,056	9
MICROBIOLOGYEVENTS.csv	631,727	631,727	16
NOTEEVENTS.csv	91,692,309	2,083,181	11
OUTPUTEVENTS.csv	4,349,219	4,349,219	13
PATIENTS.csv	46,521	46,521	8
PRESCRIPTIONS.csv	4,156,451	4,156,451	19
PROCEDUREEVENTS_MV.csv	258,067	258,067	25
PROCEDURES_ICD.csv	240,096	240,096	5
SERVICES.csv	73,344	73,344	6
TRANSFERS.csv	261,898	261,898	13

# Getting Started with MIMIC-III

## Loading Postgres Database

### 040-Load-MIMIC-into-PostgreSQL

#### **Installing-PostgreSQL-on-Windows-for-MIMIC-III.docx**

Download and use BigSQL graphical installer to install Postgres on Windows in directory D:\PostgreSQL.

To use the odbc R package below, be sure to install the Windows Postgres ODBC Driver discussed in Section 4.

#### **Loading-MIMIC-III-into-PostgreSQL.docx**

Run psql shell to ...

- create mimic database,
- load tables with data,
- build indexes for faster access,
- test the build.

The final section shows a link to online MIMIC-III database schema documentation.

#### **MIMIC-Install-on-Postgres.html**

Additional details about running SQL command in psql from previous step.

# Getting Started with MIMIC-III R Database Driver Packages

040-Load-MIMIC-into-PostgreSQL/MIMIC-III-First-Look\*.html

## 2 Open database with RPostgres

```
MimicDB <- dbConnect(RPostgres::Postgres(),  
  host      = "localhost",  
  dbname    = "mimic",  
  user      = Sys.getenv("MIMIC_User"),  
  password  = Sys.getenv("MIMIC_Password"),  
  bigint    = "integer",  
  options   = "-c search_path=mimiciii")
```

```
> Sys.getenv("R_USER")  
[1] "C:/Users/efg/Documents"
```

This PC > OS (C:) > Users > efg > Documents

Name

.Renvron

```
MIMIC_User = postgres  
MIMIC_Password = NotSoSecure
```

## 2 Open mimic database with RPostgreSQL

```
MimicDB <- dbConnect(PostgreSQL(),  
  dbname="mimic",  
  user      = Sys.getenv("MIMIC_User"),  
  password  = Sys.getenv("MIMIC_Password"))
```

## 2 Open database with odbc

```
MimicDB <- dbConnect(odbc::odbc(),  
  driver    = "PostgreSQL Unicode(x64)",  
  database  = "mimic",  
  user      = Sys.getenv("MIMIC_User"),  
  password  = Sys.getenv("MIMIC_Password"),  
  bigint    = "integer")
```

# Getting Started with MIMIC-III

## Querying MIMIC-III with SQL or dplyr

### 050-Querying-MIMIC-III

- <https://mimic.physionet.org/tutorials/intro-to-mimic-iii/>
- Online examples numbered 3 – 7,  
plus a tutorial problem in 8 steps
- Examples using SQL code chunks  
Querying-MIMIC-III-SQL.html  
Querying-MIMIC-III-Tutorial-Problem-SQL.html
- Equivalent examples with r dply code chunks  
Querying-MIMIC-III-dplyr.html  
Querying-MIMIC-III-Tutorial-Problem-Tidyverse-dplyr.html
- These sets of notebooks should help understand equivalent SQL and dplyr solutions

# Getting Started with MIMIC-III

## Querying MIMIC-III with SQL or dplyr

### Example 3. Patient Numbers

#### SQL Chunk

```
```{sql, connection=MimicDB}  
SELECT  gender, COUNT(*)  
FROM    patients  
GROUP BY gender  
```
```

```
SELECT  gender, COUNT(*)  
FROM    patients  
GROUP BY gender
```

2 records

| gender | count |
|--------|-------|
| M      | 26121 |
| F      | 20399 |

050-Querying-MIMIC-III/Querying-MIMIC-III-SQL.html  
050-Querying-MIMIC-III/Querying-MIMIC-III-dplyr.html

#### R Chunk

```
```{r}  
patients  
  group_by(gender) %>%  
  count() %>%  
  Show()  
```
```

```
library(tidyverse)
```

```
patients <- tbl(MimicDB, in_schema("mimiciii", "patients"))
```

```
patients %>%  
  group_by(gender) %>%  
  count() %>%  
  Show()
```

| gender | n     |
|--------|-------|
| M      | 26121 |
| F      | 20399 |

# Getting Started with MIMIC-III

Experimental “kable” function to “Show” results

```
library(kableExtra)
```

Helper function: Common formatting mostly for data.frames/tibbles below

```
Show <- function(data)
{
  data          %>%
  kable("html")  %>%
  kable_styling(bootstrap_options = c("striped", "bordered", "condensed"),
                position = "left", full_width = FALSE, font_size=12)
}
```

# Getting Started with MIMIC-III

## Querying MIMIC-III with SQL or dplyr

dbplyr “magic” automatically creates SQL  
“show\_query” reveals generated SQL

```
{r}  
patients %>%  
  group_by(gender) %>%  
  count() %>%  
  show_query()
```

```
<SQL>  
SELECT "gender", COUNT(*) AS "n"  
FROM mimiciii.patients  
GROUP BY "gender"
```

```
{r}  
dobByYearMonth <-  
  patients %>%  
  mutate(Year = date_part("year", dob),  
         Month = date_part("month", dob)) %>%  
  select(Year, Month) %>%  
  group_by(Year, Month) %>%  
  count() %>%  
  ungroup() %>%  
  show_query()
```

```
<SQL>  
SELECT "Year", "Month", COUNT(*) AS "n"  
FROM (SELECT "Year", "Month"  
FROM (SELECT "row_id", "subject_id", "gender", "dob", "dod", "dod_hosp",  
"dod_ssn", "expire_flag", DATE_PART('year', "dob") AS "Year",  
DATE_PART('month', "dob") AS "Month"  
FROM mimiciii.patients) "ityyvyejfh") "spjpupahrs"  
GROUP BY "Year", "Month"
```

# Getting Started with MIMIC-III

## Querying MIMIC-III with SQL or dplyr

### Tutorial. Step 2

#### SQL

```
```{sql, connection=MimicDB, output.var="SQLresults"}
```

#### Step 2

Using the patients table retrieve the calculated age of patients.

```
SELECT
  ie.subject_id,
  ie.hadm_id,
  ie.icustay_id,
  ie.intime,
  ie.outtime,
  ROUND((cast(ie.intime as date) - cast(pat.dob as date))/365.242, 2) AS age_years
FROM
  icustays ie

  INNER JOIN patients pat
    ON ie.subject_id = pat.subject_id;
```

```
dim(SQLresults)
```

```
[1] 61532      6
```

#### dplyr

```
icustays <- tbl(MimicDB, in_schema("mimiciii", "icustays"))
patients <- tbl(MimicDB, in_schema("mimiciii", "patients"))
```

#### Step 2

Using the patients table retrieve the calculated age of patients.

```
results2 <-
  icustays
  inner_join(patients, by="subject_id")
  select(subject_id, hadm_id, icustay_id, intime, outtime, dob)
  collect()
  mutate(ageYears = (as.numeric(floor_date(intime, unit="day") -
    floor_date(dob, unit="day"),
    units="days") / 365.242) %>% round(2)) %>%
  select(-dob)
```

```
dim(results2)
```

```
[1] 61532      6
```

Calculated ages

subject_id	hadm_id	icustay_id	intime	outtime	ageYears
268	110404	280836	2198-02-14	2198-02-18	65.98
269	106296	208613	2170-11-05	2170-11-08	40.10
270	188028	220345	2128-06-24	2128-06-27	80.08



# Getting Started with MIMIC-III

## Querying MIMIC-III with SQL or dplyr

### Tutorial. Step 5

## SQL

### Step 5

Next find the date of the patient's death if applicable.

```
SELECT
  ie.subject_id,
  ie.hadm_id,
  ie.icustay_id,
  ie.intime,
  ie.outtime,
  adm.deathtime,

  ROUND((cast(ie.intime as date) - cast(pat.dob as date)) / 365.242, 2) as age_years,
  (cast(ie.intime as date) - cast(adm.admittime as date)) as preiculus_days,

CASE
  WHEN ROUND((cast(ie.intime as date) - cast(pat.dob as date)) / 365.242, 2) <= 1
  THEN 'neonate'

  WHEN ROUND((cast(ie.intime as date) - cast(pat.dob as date)) / 365.242, 2) <= 14
  THEN 'middle'

  -- all ages > 89 in the database were replaced with 300
  WHEN ROUND((cast(ie.intime as date) - cast(pat.dob as date)) / 365.242, 2) > 100
  THEN '>89'

  ELSE 'adult'
END AS ICUSTAY_AGE_GROUP

FROM
  icustays ie

  INNER JOIN patients pat
  ON ie.subject_id = pat.subject_id

  INNER JOIN admissions adm
  ON ie.hadm_id = adm.hadm_id
```

```
dim(SQLresults)
```

```
[1] 61532    9
```

## dplyr

```
icustays <- tbl(MimicDB, in_schema("mimiciii", "icustays"))
patients  <- tbl(MimicDB, in_schema("mimiciii", "patients"))
admissions <- tbl(MimicDB, in_schema("mimiciii", "admissions"))
```

### Step 5

Next find the date of the patient's death if applicable.

```
results5 <-
  icustays %>%
  inner_join(patients, by = "subject_id") %>%
  inner_join(admissions, by = c("subject_id", "hadm_id")) %>%
  select(subject_id, hadm_id, icustay_id, intime, outtime,
         deathtime, dob, admittime) %>%
  collect() %>%
  mutate(ageYears = (as.numeric(floor_date(intime, unit="day") -
                                floor_date(dob, unit="day"),
                                units="days") / 365.242) %>% round(2),

         preiculusDays = (as.numeric(floor_date(intime, unit="day") -
                                floor_date(admittime, unit="day"),
                                units="days"))

  ) %>%
  select(-dob, -admittime) %>%
  mutate(icuStayAgeGroup =
    case_when
    (
      ageYears <= 1 ~ "neonate",
      ageYears <= 14 ~ "middle",
      ageYears > 100 ~ ">89",
      TRUE ~ "adult"
    )
  )
```

```
dim(results5)
```

```
[1] 61532    9
```

# MIMIC-III Critical Care Data Outline

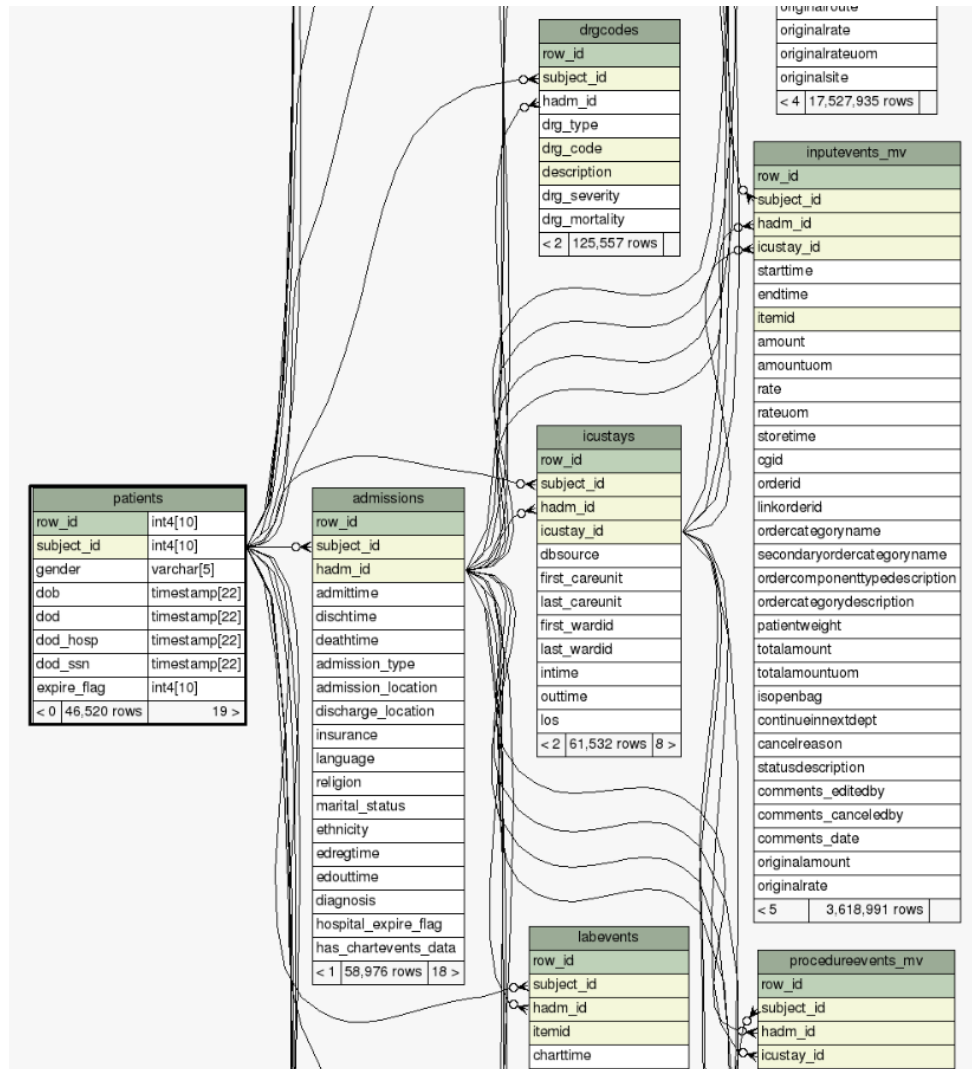
- Getting Started with MIMIC-III
  - Background / Motivation
  - Training Requirements
  - Loading Postgres Database
  - Querying MIMIC-III with SQL or dplyr
- MIMIC-III Exploration with tidyverse/dplyr
  - Patients Table
  - Admissions Table
  - Diagnoses Tables
- Take Home

# MIMIC-III Exploration

- Important to understand data fields before using in analysis projects
- Avoid “garbage in, garbage out”
- A data dictionary or code book can be helpful but often can be inadequate, out-of-date
- Database schema can help with relationships among tables, but provides little information about data fields
- Want a “statistical abstract” as starting point
- Goal is “clean” and understandable data to build predictive models

# MIMIC-III Exploration

[Online Schema](#) Shows Complicated Relationships



# MIMIC-III Exploration

[Online Schema](#) Shows Terse Field Information

## Table mimic.mimiciii.patients

Patients associated with an admission to the ICU.

☐ Related columns ☐ Constraints ☒ Comments ☒ Legend

Column	Type	Size	Nulls	Auto	Default	Children	Parents	Comments
row_id	int4	10						Unique row identifier.
subject_id	int4	10				<a href="#">admissions</a> <a href="#">callout</a> <a href="#">chartevents</a> <a href="#">cptevents</a> <a href="#">datetimeevents</a> <a href="#">diagnoses_icd</a> <a href="#">drgcodes</a> <a href="#">icustays</a> <a href="#">inputevents_cv</a> <a href="#">inputevents_mv</a> <a href="#">labevents</a> <a href="#">microbiologyevents</a> <a href="#">noteevents</a> <a href="#">outputevents</a> <a href="#">prescriptions</a> <a href="#">procedureevents_mv</a> <a href="#">procedures_icd</a> <a href="#">services</a> <a href="#">transfers</a>		Primary key. Identifies the patient.
gender	varchar	5						Gender.
dob	timestamp	22						Date of birth.
dod	timestamp	22	✓		null			Date of death. Null if the patient was alive at least 90 days post hospital discharge.
dod_hosp	timestamp	22	✓		null			Date of death recorded in the hospital records.
dod_ssn	timestamp	22	✓		null			Date of death recorded in the social security records.
expire_flag	int4	10						Flag indicating that the patient has died.

Table contained 46,520 rows at Thu Oct 12 12:29 EDT 2017

# MIMIC-III Exploration

## Patients Table

### MIMIC-III Patients Table

- 1 Setup
  - 1.1 Packages
  - 1.2 Helper function
  - ➔ 1.3 Open database
- 2 List of fields in a patients table
- ➔ 3 Sample patients
- ➔ 4 Record count
- 5 Fields
  - 5.1 row\_id
  - ➔ 5.2 subject\_id
  - 5.3 gender
  - ➔ 5.4 dob (date of birth) counts
  - 5.5 expire\_flag
  - 5.6 dod (date of death) counts
  - 5.7 dod\_hosp and dod\_ssn
  - ➔ 5.8 Computed: Age at Death [INCORRECT results with RPostgres]
- 6 Close database
- 7 Use RPostgreSQL package
  - ➔ 7.1 Computed: Age at Death [CORRECT results with RPostgreSQL]

### YAML Markdown Header

```
1 ---
2 title: "MIMIC-III Patients Table"
3 output:
4   html_document:
5     toc: yes
6     number_sections: yes
7   html_notebook:
8     toc: yes
9 ---
10
11 <style type="text/css">
12 div#TOC li {
13   list-style:none;
14   background-image:none;
15   background-repeat:none;
16   background-position:0;
17 }
18 </style>
```

# MIMIC-III Exploration

## Patients Table

### 1.3 Open database

```
MimicDB <- dbConnect(RPostgres::Postgres(),
                     dbname = "mimic",
                     user   = Sys.getenv("MIMIC_User"),
                     password = Sys.getenv("MIMIC_Password"),
                     bigint  = "integer",
                     options = "-c search_path=mimiciii")

plotCaptionLeft <- "MIMIC-III v1.4"
plotCaptionRight <- paste("efg", format(Sys.time(), "%Y-%m-%d"))
```

### 3 Sample patients

```
patients <- tbl(MimicDB, in_schema("mimiciii", "patients"))

patients %>%
  head(10) %>%
  Show()
```

row_id	subject_id	gender	dob	dod	dod_hosp	dod_ssn	expire_flag
234	249	F	2075-03-13	NA	NA	NA	0
235	250	F	2164-12-27	2188-11-22	2188-11-22	NA	1

### 4 Record count

```
patients %>%
  summarize(n = n()) %>%
  Show()
```

n
46520

# MIMIC-III Exploration

## Patients Table

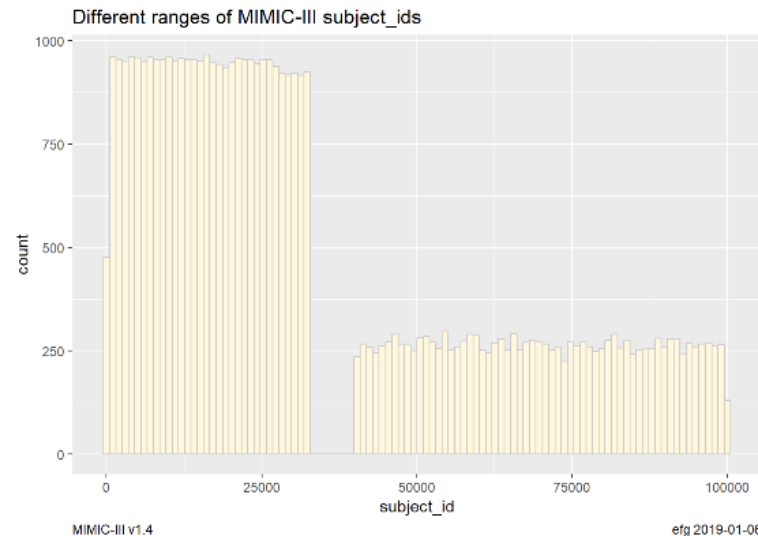
### 5.2 subject\_id

```
patients %>%  
  summarize(minSubjectID = min(subject_id, na.rm=TRUE),  
            maxSubjectID = max(subject_id, na.rm=TRUE)) %>%  
  Show()
```

minSubjectID	maxSubjectID
2	99999

There are two groups of `subject_id` ranges.

```
patients %>%  
  select(subject_id) %>%  
  collect %>%  
  ggplot(aes(x=subject_id)) +  
    geom_histogram(fill="cornsilk", color="grey80", bins=100) +  
    labs(title = "Different ranges of MIMIC-III subject_ids",  
         caption=c(plotCaptionLeft, plotCaptionRight)) +  
    theme(plot.caption = element_text(hjust=c(0.0,1.0)))
```



Why do some patients have more data than others?

Why aren't counts more ragged?



# MIMIC-III Exploration

## Patients Table

### 5.4 dob (date of birth) counts

DB  
R

```
dobByYearMonth <-  
  patients %>%  
  mutate(Year = date_part("year", dob),  
         Month = date_part("month", dob)) %>%  
  select(Year, Month) %>%  
  group_by(Year, Month) %>%  
  count() %>%  
  ungroup() %>%  
  collect() %>%  
  arrange(Year, Month) %>%  
  spread(Month, n, fill=0)
```

Counts of MIMIC-III dates of birth (dob) by year and month

```
dobByYearMonth %>%  
  Show() %>%  
  scroll_box(height = "500px")
```

Year	1	2	3	4	5	6	7	8	9	10	11	12
1800	0	0	0	0	0	0	4	2	3	2	1	1
1801	1	4	2	0	1	0	3	3	2	3	0	1
1802	5	2	4	3	0	3	3	3	1	2	4	3
1803	1	3	5	1	2	0	3	0	2	4	0	1
1804	2	5	1	3	3	3	1	0	4	0	2	2
...	.	.	.	.	.	.	.	.	.	.	.	.

To de-identify patients, dates have been shifted over a 300 year range. [Unusual convention]

Date-of-birth values range from 1800 to 1901 and 2012 to 2201.

Date-of-death values range from 2100 to 2211.

Year-over-year trends will be impossible to observe, which is often a good self-consistency check.

# MIMIC-III Exploration

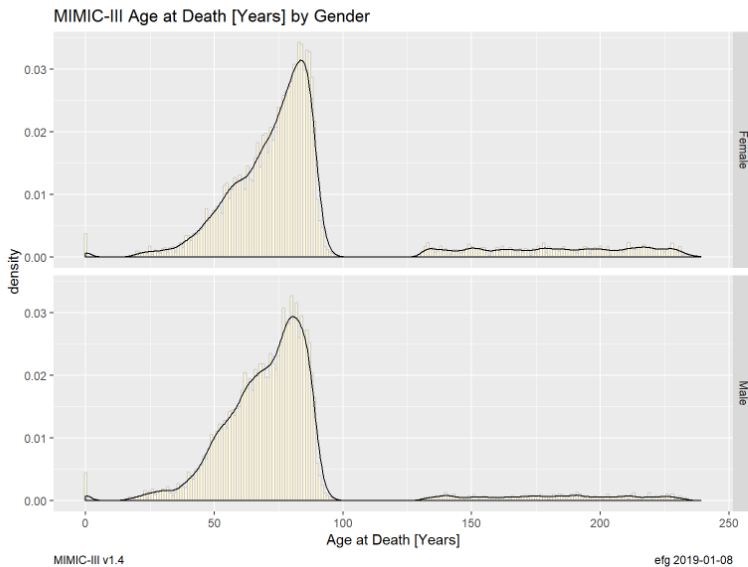
## Patients Table

### 5.8 Computed: Age at Death [INCORRECT results with RPostgres]

```
facetLabels <- c('F' = "Female",
                'M' = "Male")

patients %>%
  filter(expire_flag == 1) %>%
  select(-dod_hosp, -dod_ssn, -expire_flag) %>%
  collect() %>%
  mutate(AgeAtDeathYears = as.double(dod - dob) / (86400 * 365.25)) %>%

ggplot(aes(x = AgeAtDeathYears, y = ..density..)) +
  geom_histogram(fill="cornsilk", color="grey80", binwidth=1) +
  geom_density() +
  facet_grid(gender ~ ., labeller = as_labeller(facetLabels)) +
  labs(title="MIMIC-III Age at Death [Years] by Gender",
       x = "Age at Death [Years]",
       caption=c(plotCaptionLeft, plotCaptionRight)) +
  theme(plot.caption = element_text(hjust=c(0.0,1.0)))
```

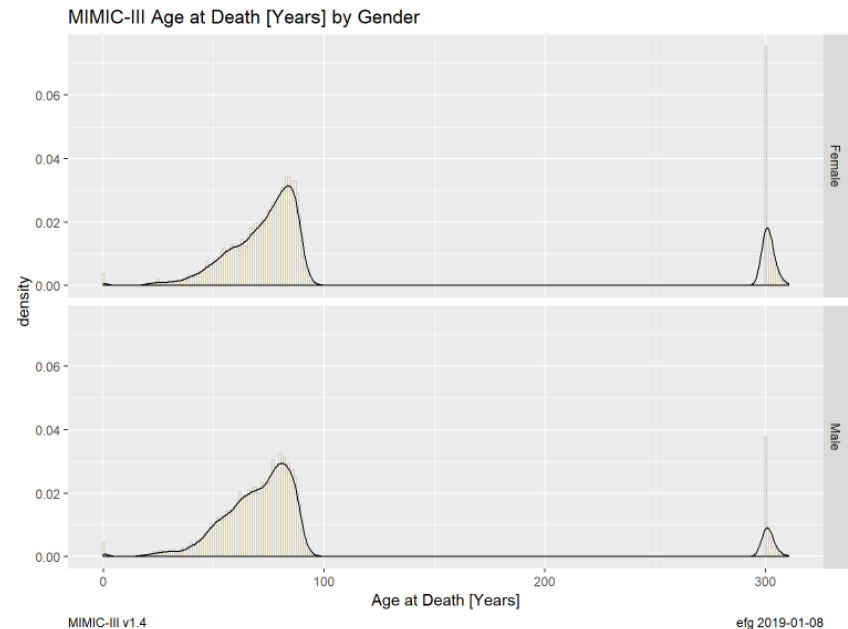


Confirmed that this is a bug in RPostgres. Should be fixed with the next update, planned end of February.

### 7.1 Computed: Age at Death [CORRECT results with RPostgreSQL]

```
patients %>%
  filter(expire_flag == 1) %>%
  select(-dod_hosp, -dod_ssn, -expire_flag) %>%
  collect() %>%
  mutate(AgeAtDeathYears = as.double(dod - dob) / (86400 * 365.25)) %>%

ggplot(aes(x = AgeAtDeathYears, y = ..density..)) +
  geom_histogram(fill="cornsilk", color="grey80", binwidth=1) +
  geom_density() +
  facet_grid(gender ~ ., labeller = as_labeller(facetLabels)) +
  labs(title="MIMIC-III Age at Death [Years] by Gender",
       x = "Age at Death [Years]",
       caption=c(plotCaptionLeft, plotCaptionRight)) +
  theme(plot.caption = element_text(hjust=c(0.0,1.0)))
```



# MIMIC-III Exploration

## Admissions Table

### MIMIC-III Admissions Table

- 1 Setup
  - 1.1 Packages
  - 1.2 Helper function
  - 1.3 Open database
- 2 List tables in database
- 3 List of fields in admissions table
- 4 Admissions table record count
- 5 Sample admissions records
- 6 Fields
  - 6.1 row\_id
  - 6.2 subject\_id
  - 6.3 hadm\_id
  - 6.4 admittance, dischtime, deathtime
  - 6.5 admission\_type
  - 6.6 admission\_location
  - 6.7 discharge\_location
  - 6.8 insurance
  - 6.9 language
  - 6.10 religion
  - 6.11 marital\_status
  - 6.12 ethnicity
  - 6.13 edregtime, edouttime
  - 6.14 diagnosis
  - 6.15 hospital\_expire\_flag
  - 6.16 has\_chartevents\_data
- 7 Close database

# MIMIC-III Exploration

## Admissions Table

### 2 List tables in database

This works with RPostgres but not RPostgreSQL.

```
dbListTables(MimicDB)
```

```
[1] "admissions"      "callout"          "caregivers"
[4] "chartevents_1"   "chartevents_2"    "chartevents_3"
[7] "chartevents_4"   "chartevents_5"    "chartevents_6"
[10] "chartevents_7"   "chartevents_8"    "chartevents_9"
[13] "chartevents_10"  "chartevents_11"   "chartevents_12"
[16] "chartevents_13"  "chartevents_14"   "chartevents_15"
[19] "chartevents_16"  "chartevents_17"   "chartevents"
[22] "cptevents"       "datetimeevents"   "diagnoses_icd"
[25] "drgcodes"        "d_cpt"            "d_icd_diagnoses"
[28] "d_icd_procedures" "d_items"          "d_labitems"
[31] "icustays"         "inputevents_cv"    "inputevents_mv"
[34] "labevents"        "microbiologyevents" "noteevents"
[37] "outputevents"     "patients"          "prescriptions"
[40] "procedureevents_mv" "procedures_icd"    "services"
[43] "transfers"
```

### 3 List of fields in admissions table

```
dbListFields(MimicDB, "admissions")
```

```
[1] "row_id"          "subject_id"       "hadm_id"
[4] "admittime"       "dischtime"        "deathtime"
[7] "admission_type"  "admission_location" "discharge_location"
[10] "insurance"       "language"         "religion"
[13] "marital_status"  "ethnicity"        "edregtime"
[16] "edouttime"       "diagnosis"        "hospital_expire_flag"
[19] "has_chartevents_data"
```

# MIMIC-III Exploration

## Admissions Table

### 5 Sample admissions records

```
admissions %>%
  head(5) %>%
  Show()
```

row_id	subject_id	hadm_id	admittime	disctime	deathtime	admission_type	admission_location	discharge_location	insurance	language	religion
21	22	165315	2196-04-09 12:26:00	2196-04-10 15:54:00	NA	EMERGENCY	EMERGENCY ROOM ADMIT	DISC-TRAN CANCER/CHLDRN H	Private	NA	UNOBTAINABLE
22	23	152223	2153-09-03 07:15:00	2153-09-08 19:10:00	NA	ELECTIVE	PHYS REFERRAL/NORMAL DELI	HOME HEALTH CARE	Medicare	NA	CATHOLIC

marital_status	ethnicity	edregtime	edouttime	diagnosis	hospital_expire_flag	has_chartevents_data
MARRIED	WHITE	2196-04-09 10:06:00	2196-04-09 13:24:00	BENZODIAZEPINE OVERDOSE	0	1
MARRIED	WHITE	NA	NA	CORONARY ARTERY DISEASEARTERY BYPASS GRAFT/SDA	0	1

# MIMIC-III Exploration

## Admissions Table

### 6.4 admittance, dischtime, deathtime

- Time of admission to the hospital.
- Time of discharge from the hospital.
- Time of death.

sample values

Unclear why time differences are in minutes here.

```
admissions %>%  
  select(admittime, dischtime, deathtime) %>%  
  collect() %>%  
  mutate(StayDays = as.double(dischtime - admittime) / 1440) %>%  
  head(10) %>%  
  Show()
```

admittime	dischtime	deathtime	StayDays
2196-04-09 12:26:00	2196-04-10 15:54:00	NA	1.144444
2153-09-03 07:15:00	2153-09-08 19:10:00	NA	5.496528

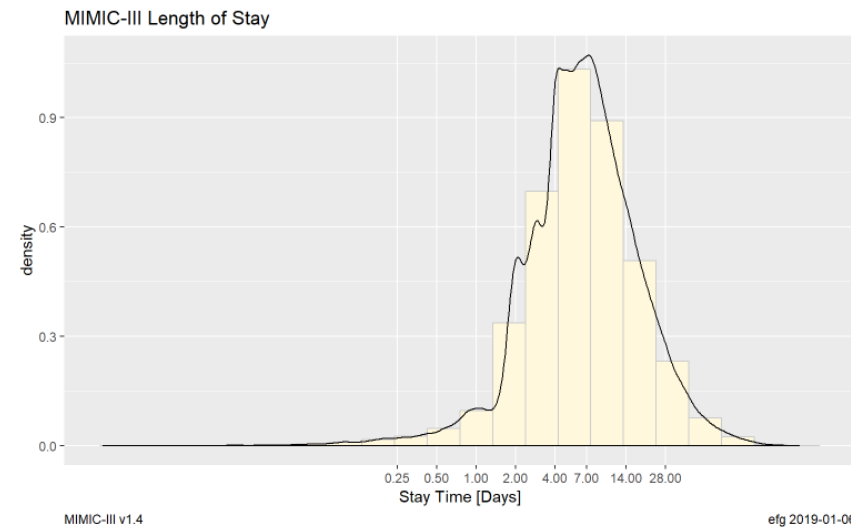
# MIMIC-III Exploration

## Admissions Table

How does visit time vary?

Length of Stay = Discharge Time - Admit Time

```
admissions %>%  
  select(admittime, dischtime, deathtime) #>%  
  collect() #>%  
  mutate(StayDays = as.double(dischtime - admittime) / 1440) #>%  
  filter(StayDays > 0) #>%  
  
  ggplot(aes(x = StayDays, y = ..density..)) +  
    geom_histogram(fill="cornsilk", color="grey80", binwidth=0.25) +  
    geom_density() +  
    scale_x_log10(breaks = c(0.25, 0.50, 1, 2, 4, 7, 14, 28),  
                 minor_breaks = NULL) +  
    labs(title="MIMIC-III Length of Stay",  
         x = "Stay Time [Days]",  
         caption=c(plotCaptionLeft, plotCaptionRight)) +  
    theme(plot.caption = element_text(hjust=c(0.0,1.0)))
```



Note log scale

Length-of-Stay often used in predictive modeling

# MIMIC-III Exploration

## Admissions Table

### 6.7 discharge\_location

```
admissions %>%  
  group_by(discharge_location) %>%  
  summarize(n = n()) %>%  
  Show()
```

discharge_location	n
ICF	47
HOSPICE-MEDICAL FACILITY	153
DISC-TRAN CANCER/CHLDRN H	669
DEAD/EXPIRED	5854
HOME	18962
LEFT AGAINST MEDICAL ADVI	365
SNF-MEDICAID ONLY CERTIF	1
REHAB/DISTINCT PART HOSP	6429
HOME WITH HOME IV PROVIDR	67
DISC-TRAN TO FEDERAL HC	11
LONG TERM CARE HOSPITAL	2305
SNF	7705
SHORT TERM HOSPITAL	1534
OTHER FACILITY	63
HOSPICE-HOME	402
HOME HEALTH CARE	13963
DISCH-TRAN TO PSYCH HOSP	446

### 6.8 insurance

```
admissions %>%  
  group_by(insurance) %>%  
  summarize(n = n()) %>%  
  Show()
```

insurance	n
Government	1783
Self Pay	611
Medicare	28215
Private	22582
Medicaid	5785



# MIMIC-III Exploration

## Diagnoses Tables

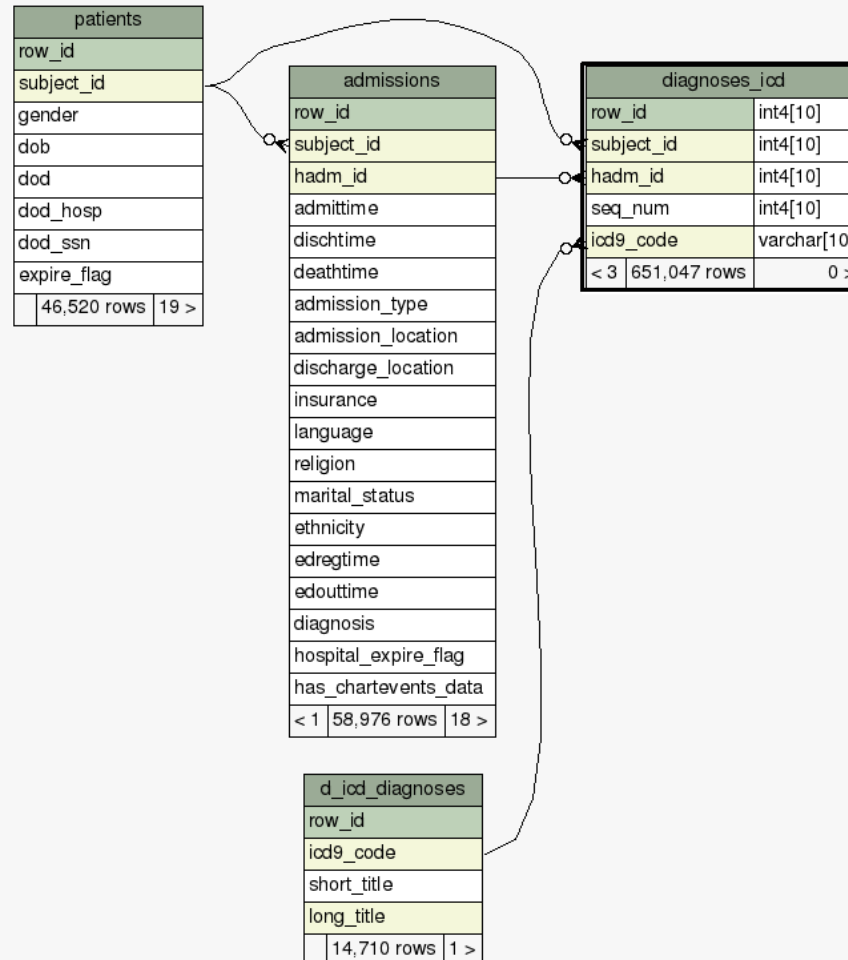
### MIMIC-III Diagnoses Tables

- 1 Fact and Dimension tables
- 2 Setup
  - 2.1 Packages
  - 2.2 Helper function
  - 2.3 Open database
- 3 Tables
  - 3.1 Dimension table: d\_icd\_diagnoses
    - 3.1.1 Fields
    - 3.1.2 Record count
    - 3.1.3 Sample records
  - 3.2 Fact table: diagnoses\_icd
    - 3.2.1 Fields
    - 3.2.2 Record count
    - 3.2.3 Sample records
    - 3.2.4 Fields
  - 3.3 Join diagnoses tables
    - 3.3.1 inner\_join
    - 3.3.2 left\_join
- 4 Summaries
  - 4.1 Summarize Diagnoses Counts
    - 4.1.1 Top 10
  - 4.2 Summarize Diagnoses as Primary or Secondary
    - 4.2.1 Top 10 Primary Diagnoses
  - 4.3 Summarize Diagnoses by Age Intervals
    - 4.3.1 Admit Age Density
    - 4.3.2 Counts by diagnosis and age interval

# MIMIC-III Exploration

## Diagnoses Tables

Close relationships within ☒ one ☐ two degrees of separation:



# MIMIC-III Exploration

## Diagnoses Tables

### Fact Table

### Dimension Table

#### 3.2.3 Sample records

```
factDiagnoses %>%  
  arrange(subject_id, row_id) %>% # control order  
  head(5) %>%  
  Show(caption = "Sample Records from Diagnoses Fact Table,  
diagnoses_icd")
```

Sample Records from Diagnoses Fact Table,  
diagnoses\_icd

row_id	subject_id	hadm_id	seq_num	icd9_code
1	2	163353	1	V3001
2	2	163353	2	V053
3	2	163353	3	V290
4	3	145834	1	0389
5	3	145834	2	78559

#### 3.1.3 Sample records

```
dimDiagnoses %>%  
  arrange(icd9_code) %>%  
  head(10) %>%  
  Show(caption = "Sample Records from Diagnoses Dimension Table, d_icd_diagnoses")
```

Sample Records from Diagnoses Dimension Table, d\_icd\_diagnoses

row_id	icd9_code	short_title	long_title
233	0010	Cholera d/t vib cholerae	Cholera due to vibrio cholerae
234	0011	Cholera d/t vib el tor	Cholera due to vibrio cholerae el tor
235	0019	Cholera NOS	Cholera, unspecified
236	0020	Typhoid fever	Typhoid fever
237	0021	Paratyphoid fever a	Paratyphoid fever A
238	0022	Paratyphoid fever b	Paratyphoid fever B
239	0023	Paratyphoid fever c	Paratyphoid fever C
240	0029	Paratyphoid fever NOS	Paratyphoid fever, unspecified
241	0030	Salmonella enteritis	Salmonella gastroenteritis
242	0031	Salmonella septicemia	Salmonella septicemia

# MIMIC-III Exploration

## Diagnoses Tables

### Inner Join

Join dimension table to fact table

#### 3.3.1 inner\_join

```
factDiagnoses                                >>>
  inner_join(dimDiagnoses, by = "icd9_code")  >>>
  head()                                     >>>
  collect()                                 >>>
  str()
```

```
Classes 'tbl_df', 'tbl' and 'data.frame':  6 obs. of  8 variables:
 $ row_id.x : int  307770 497329 513229 398507 579783 408925
 $ subject_id : int  27367 65733 69141 44437 83908 46693
 $ hadm_id    : int  136441 168926 136542 181557 180356 183574
 $ seq_num    : int    7  9  1 10  3  4
 $ icd9_code   : chr  "0030" "0030" "0031" "0038" ...
 $ row_id.y : int  241 241 242 249 249 250
 $ short_title: chr  "Salmonella enteritis" "Salmonella enteritis" "Sa
pticemia" "Salmonella infection NEC" ...
 $ long_title : chr  "Salmonella gastroenteritis" "Salmonella gastroen
almonella septicemia" "Other specified salmonella infections" ...
```

# MIMIC-III Exploration

## Diagnoses Tables

### Inner Join

```
factDiagnoses                                     %>%
  inner_join(dimDiagnoses, by = "icd9_code")      %>%
  arrange(subject_id, row_id.x)                  %>% # control order
  head(5)   %>%
  collect()                                       %>%
  Show(caption = "Sample Diagnoses Records (Fact + Dimension)")
```

Sample Diagnoses Records (Fact + Dimension)

row_id.x	subject_id	hadm_id	seq_num	icd9_code	row_id.y	short_title	long_title
1	2	163353	1	V3001	13695	Single lb in-hosp w cs	Single liveborn, born in hospital, delivered by cesarean section
2	2	163353	2	V053	12202	Need prphyl vc vrl hepat	Need for prophylactic vaccination and inoculation against viral hepatitis
3	2	163353	3	V290	13688	NB obsrv suspct infect	Observation for suspected infectious condition
4	3	145834	1	0389	660	Septicemia NOS	Unspecified septicemia
5	3	145834	2	78559	12992	Shock w/o trauma NEC	Other shock without mention of trauma

Fact

Dimension

# MIMIC-III Exploration

## Diagnoses Tables

### Inner Join: Missing Records?

#### 3.2.2 Record count

```
factDiagnoses %>%  
  summarize(n = n(),  
            diagnosesCodes = n_distinct(icd9_code)) %>%  
  Show()
```

n	diagnosesCodes
651047	6984

Why were so many fact records lost in join?

```
factDiagnoses %>%  
  inner_join(dimDiagnoses, by = "icd9_code") %>%  
  count() %>%  
  Show()
```

n
634709

$651,047 - 634,709 = 16,338$  missing records

Dimension table does not have entries for ~144 icd9\_codes in fact table.

One way to fix: update the dimension table to have all possible codes.

# MIMIC-III Exploration

## Diagnoses Tables

### Left Join fills in NULL values

#### 3.2.2 Record count

```
factDiagnoses %>%
  summarize(n = n(),
            diagnosesCodes = n_distinct(icd9_code)) %>%
  Show()
```

n	diagnosesCodes
651047	6984

#### 3.3.2 left\_join

Keep all diagnosis codes even if not in dimension table using a `left_join`. This will create NULLs for diagnosis descriptions.

```
factDiagnoses %>%
  left_join(dimDiagnoses, by = "icd9_code") %>%
  count() %>%
  Show()
```

n
651047

Examples with missing short\_titles

```
factDiagnoses %>%
  left_join(dimDiagnoses, by = "icd9_code") %>%
  filter(is.na(short_title)) %>%
  head(2) %>%
  Show()
```

row_id.x	subject_id	hadm_id	seq_num	icd9_code	row_id.y	short_title	long_title
1524	117	140784	2	7895	NA	NA	NA
1529	117	140784	7	2765	NA	NA	NA

# MIMIC-III Exploration

## Summarize Diagnoses Counts

### 4.1 Summarize Diagnoses Counts

```
diagnosesCounts <-  
  factDiagnoses                                %>%  
  filter(!is.na(icd9_code))                    %>%  
  group_by(icd9_code)                          %>%  
  count()                                       %>%  
  ungroup()                                    %>%  
  left_join(dimDiagnoses,  
            by = "icd9_code")                  %>%  
  select(n, everything(), -row_id)             %>%  
  arrange(desc(n))                             %>%  
  collect()  
  
nrow(diagnosesCounts)
```

```
[1] 6984
```

Many of the ICD 9 diagnoses codes in the dimension table are never referenced.



# MIMIC-III Exploration

## Summarize Diagnoses Counts: Top 10

### 4.1.1 Top 10

```
diagnosesCounts %>% head(10) %>% Show()
```

n	icd9_code	short_title	long_title
20703	4019	Hypertension NOS	Unspecified essential hypertension
13111	4280	CHF NOS	Congestive heart failure, unspecified
12891	42731	Atrial fibrillation	Atrial fibrillation
12429	41401	Crnry athrsc1 native vssl	Coronary atherosclerosis of native coronary artery
9119	5849	Acute kidney failure NOS	Acute kidney failure, unspecified
9058	25000	DMII wo cmp nt st uncntr	Diabetes mellitus without mention of complication, type II or unspecified type, not stated as uncontrolled
8690	2724	Hyperlipidemia NEC/NOS	Other and unspecified hyperlipidemia
7497	51881	Acute respiratry failure	Acute respiratory failure
6555	5990	Urin tract infection NOS	Urinary tract infection, site not specified
6326	53081	Esophageal reflux	Esophageal reflux

# MIMIC-III Exploration

## Summarize Diagnoses by Age Intervals

### 4.3 Summarize Diagnoses by Age Intervals

```
admitAges <-  
  factDiagnoses %>%  
  inner_join(factPatients,  
    by = "subject_id") %>%  
  inner_join(factAdmissions,  
    by = c("subject_id",  
      "hadm_id")) %>%  
  select(subject_id, hadm_id, icd9_code, dob, admittance) %>%  
  collect() %>%  
  mutate(  
    # collect from database and use R for mutate  
    delta = admittance - dob,  
    admitAgeYears = as.numeric((admittance - dob) / ( 365.25 * 86400)), # seconds  
    ageInterval = pmin(90, 10 * admitAgeYears %/% 10) # age decades; 90 is 90+  
  )  
  
nrow(admitAges)
```

```
[1] 651047
```

Sample records

```
admitAges %>%  
  head(2) %>%  
  Show()
```

subject_id	hadm_id	icd9_code	dob	admittime	delta	admitAgeYears	ageInterval
109	172335	40301	2117-08-07	2141-09-18 10:32:00	761049120 secs	24.11619	20
109	172335	486	2117-08-07	2141-09-18 10:32:00	761049120 secs	24.11619	20

# MIMIC-III Exploration

## Summarize Diagnoses by Age Intervals

Counts by age intervals

```
table(admitAges$ageInterval)
```

0	10	20	30	40	50	60	70	80	90
46791	2821	17857	25990	56645	101425	131705	131789	101652	34372

Interval percentages

```
round(100 * table(admitAges$ageInterval) / nrow(admitAges), 2)
```

0	10	20	30	40	50	60	70	80	90
7.19	0.43	2.74	3.99	8.70	15.58	20.23	20.24	15.61	5.28

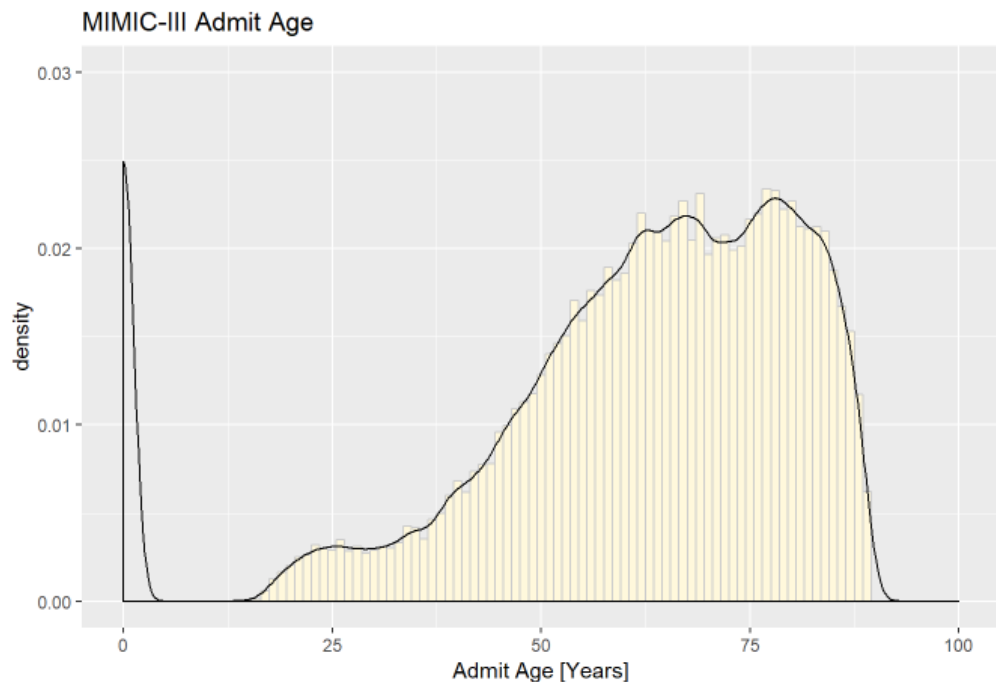
Here "0" means 0-9, "10" means 10-19, ..., "80" means 80-89, "90" means 90+

# MIMIC-III Exploration

## Summarize Diagnoses by Age Intervals

### 4.3.1 Admit Age Density

```
ggplot(admitAges, aes(x = admitAgeYears, y = ..density..)) +  
  geom_histogram(fill="cornsilk", color="grey80", binwidth=1) +  
  geom_density() +  
  xlim(0, 100) +  
  ylim(0.00, 0.03) +  
  labs(title="MIMIC-III Admit Age",  
       x = "Admit Age [Years]",  
       caption=c(plotCaptionLeft, plotCaptionRight)) +  
  theme(plot.caption = element_text(hjust=c(0.0,1.0)))
```



MIMIC-III v1.4

efg 2019-01-11

# MIMIC-III Exploration

## Summarize Diagnoses by Age Intervals

### 4.3.2 Counts by diagnosis and age interval

```
admitAgesCounts <-  
  factDiagnoses %>%  
    inner_join(factPatients,  
               by = "subject_id") %>%  
    inner_join(factAdmissions,  
               by = c("subject_id",  
                      "hadm_id")) %>%  
    select(subject_id, hadm_id, icd9_code, dob, admittance) %>%  
    collect() %>% # extract data from DB  
    mutate( # normal R processing  
      delta = admittance - dob,  
      admitAgeYears = as.numeric((admittance - dob)) / (365.25 * 86400), # seconds  
      ageInterval = pmin(90, 10 * admitAgeYears %/% 10) # age decades; 90 is 90+  
    ) %>%  
    group_by(icd9_code, ageInterval) %>%  
    count() %>% # counts in long format  
    ungroup() %>%  
    arrange(icd9_code, ageInterval) %>%  
    spread(ageInterval, n, fill=0) %>% # long to wide format  
    mutate(RowTotal = rowSums(.[-1], na.rm=TRUE)) %>%  
    select(icd9_code, RowTotal, everything()) %>% # reorder variables  
    arrange(desc(RowTotal)) %>% # descending order  
    left_join(dimDiagnoses, by = "icd9_code", # add code labels  
              copy = TRUE) %>% # table to tibble  
    select(-row_id) %>% # remove a variable  
    rename("0s" = "0", "10s" = "10", "20s" = "20", "30s" = "30",  
           "40s" = "40", "50s" = "50", "60s" = "60", "70s" = "70",  
           "80s" = "80", "90+" = "90") # slightly better names  
  
nrow(admitAgesCounts)
```

[1] 6985

# MIMIC-III Exploration

## Summarize Diagnoses by Age Intervals

```
admitAgesCounts %>%
  head(5) %>%
  Show()
```

icd9_code	RowTotal	0s	10s	20s	30s	40s	50s	60s	70s	80s	90+	short_title	long_title
4019	20703	13	6	109	433	1489	3479	5035	5184	3744	1211	Hypertension NOS	Unspecified essential hypertension
4280	13111	13	7	67	208	644	1500	2722	3527	3161	1262	CHF NOS	Congestive heart failure, unspecified
42731	12891	0	2	19	71	316	1138	2652	3961	3496	1236	Atrial fibrillation	Atrial fibrillation
41401	12429	0	0	6	135	614	1815	3241	3488	2431	699	Coronary atherosclerosis of native coronary artery	Coronary atherosclerosis of native coronary artery
5849	9119	4	7	169	307	707	1424	1777	2088	1837	799	Acute kidney failure NOS	Acute kidney failure, unspecified

# MIMIC-III Exploration

## Summarize Diagnoses by Age Intervals

Add total row at bottom

```
admitAgesCounts <-
  bind_rows(admitAgesCounts,
    bind_cols(icd9_code = "ColumnTotal",
      admitAgesCounts %>%
        summarize_if(is.numeric, sum, na.rm=TRUE)))
```

```
admitAgesCounts %>%
  tail(3) %>%
  Show()
```

icd9_code	RowTotal	0s	10s	20s	30s	40s	50s	60s	70s	80s	90+	short_title	long_title
V9089	1	0	0	0	1	0	0	0	0	0	0	Retain FB NEC	Other specified retained foreign body
V9103	1	0	0	0	1	0	0	0	0	0	0	Twin gest- dich/diamniotc	Twin gestation, dichorionic/diamniotic (two placentae, two amniotic sacs)
ColumnTotal	651047	46791	2821	17857	25990	56645	101425	131705	131789	101652	34372	NA	NA

# Take Home

- MIMIC-III is a great data source for data science experiments and future predictive analytics projects involving electronic health records.
- MIMIC-IV to appear in 2019?
- There is also waveform data to explore via a separate download

## THE MIMIC-III WAVEFORM DATABASE

