

Iniciación práctica al análisis de datos OMOP

Estudios de caracterización (PatientProfiles and CohortCharacteristics)



Real World Epidemiology

Marzo 2025

A la hora de hacer un estudio, una vez tenemos la población definida, ¿cuál es el **primer paso** que debemos hacer?

A la hora de hacer un estudio, una vez tenemos la población definida, ¿cuál es el **primer paso** que debemos hacer?

Hacer una **descriptiva** de nuestra **población de estudio**:
Características demográficas, condiciones previas, historia previa

❖ Tabla 1

Calcular edad, calcular historia previa, juntar tablas.

Pueden ser **muchas líneas de código...**



Iniciación práctica al análisis de datos OMOP

PatientProfiles



Real World Epidemiology

Marzo 2025

- El objetivo del paquete ***PatientProfiles*** es **simplificar el código** necesario para **caracterizar individuos** (características demográficas, historia previa, etc.)



- Se ha desarrollado para el **Proyecto de Darwin EU®** por el equipo OxInfer.
- Está disponible en CRAN y se puede instalar fácilmente en R.
- ¿Cómo podemos **descargarlo**?

```
>install.packages("PatientProfiles")  
>library(PatientProfiles)
```

<https://cran.r-project.org/web/packages/PatientProfiles/index.html>

Contiene un **conjunto de funciones** que permiten **añadir las características individuales** más comunes en **cualquier tabla del OMOP CDM** que contiene datos a **nivel de paciente** (por ejemplo, aparición de enfermedades, exposición a fármacos, etc.)

Contiene un conjunto de funciones que permiten añadir las características individuales más comunes en cualquier tabla del OMOP CDM que contiene datos a nivel de paciente (por ejemplo, aparición de enfermedades, exposición a fármacos, etc.)

¿Cuáles son estas **tablas**?

Contiene un conjunto de funciones que permiten añadir las características individuales más comunes en cualquier tabla del OMOP CDM que contiene datos a nivel de paciente (por ejemplo, aparición de enfermedades, exposición a fármacos, etc.)

¿Cuáles son estas **tablas**?

Person, conditions, drug_exposures...

Contiene un conjunto de funciones que permiten añadir las características individuales más comunes en cualquier tabla del OMOP CDM que contiene datos a nivel de paciente (por ejemplo, aparición de enfermedades, exposición a fármacos, etc.)

¿Cuáles son estas tablas?

Person, conditions, drug_exposures...

¿Cuáles son las **características** que queremos añadir?

Contiene un conjunto de funciones que permiten añadir las características individuales más comunes en cualquier tabla del OMOP CDM que contiene datos a nivel de paciente (por ejemplo, aparición de enfermedades, exposición a fármacos, etc.)

¿Cuáles son estas tablas?

Person, conditions, drug_exposures...

¿Cuáles son las **características** que queremos añadir?

Edad, sexo, historia previa, intersecciones entre cohortes...

Funciones del paquete *PatientProfiles*

addAge	3
addAgeQuery	4
addCategories	5
addCdmName	6
addCohortIntersectCount	7
addCohortIntersectDate	8
addCohortIntersectDays	9
addCohortIntersectFlag	11
addCohortName	12
addConceptIntersectCount	13
addConceptIntersectDate	14
addConceptIntersectDays	16
addConceptIntersectField	17
addConceptIntersectFlag	19
addDateOfBirth	20
addDateOfBirthQuery	21
addDeathDate	22
addDeathDays	23
addDeathFlag	24
addDemographics	25
addDemographicsQuery	27
addFutureObservation	29
addFutureObservationQuery	30
addInObservation	31
addInObservationQuery	32
addObservationPeriodId	33
addObservationPeriodIdQuery	34
addPriorObservation	35
addPriorObservationQuery	36
addSex	37
addSexQuery	37
addTableIntersectCount	38
addTableIntersectDate	39
addTableIntersectDays	40
addTableIntersectField	42
addTableIntersectFlag	43
availableEstimates	44
benchmarkPatientProfiles	45
endDateColumn	46
filterCohortId	46

filterInObservation	47
mockDisconnect	47
mockPatientProfiles	48
sourceConceptIdColumn	49
standardConceptIdColumn	49
startDateColumn	50
summariseResult	50
variableTypes	52

Si queremos **añadir la edad** a cualquier tabla que contenga la información a nivel del paciente.

```
> cdm$curs_omop_covid <- cdm$curs_omop_covid %>%  
  addAge()
```

Si queremos añadir la edad a cualquier tabla que contenga la información a nivel del paciente.

```
>cdm$curs_omop_covid <- cdm$curs_omop_covid %>%  
  addAge()
```

¿Qué **problemas** pueden surgir a la hora de calcular la edad?

Si queremos añadir la edad a cualquier tabla que contenga la información a nivel del paciente.

```
> cdm$curs_omop_covid <- cdm$curs_omop_covid %>%  
  addAge()
```

¿Qué **problemas** pueden surgir a la hora de calcular la edad?

Puede ser que nos falte información sobre el mes o el día de nacimiento.

addAge()

```
addAge(  
  x,  
  indexDate = "cohort_start_date",  
  ageName = "age",  
  ageGroup = NULL,  
  ageMissingMonth = 1,  
  ageMissingDay = 1,  
  ageImposeMonth = FALSE,  
  ageImposeDay = FALSE,  
  missingAgeGroupValue = "None",  
  name = NULL  
)
```

addAge()

```
> cdm$curso_omop_covid
# Source:   table<curso_omop_covid> [?? x 4]
# Database: DuckDB v0.10.0 [apalomar@windows 10 x64:R 4.2.3/C:\Users\
  cohort_definition_id subject_id cohort_start_date cohort_end_date
      <int>          <dbl> <date>          <date>
1           1           493 2020-10-06      2020-10-16
2           1          3447 2021-02-21      2021-03-03
3           1          3596 2021-06-05      2021-06-15
4           1          4251 2021-04-19      2021-04-29
5           1          7310 2021-01-20      2021-01-30
6           1          3598 2020-05-20      2020-05-30
7           1          8462 2020-07-25      2020-08-04
8           1          7813 2020-12-09      2020-12-19
9           1          9848 2021-01-09      2021-01-19
10          1          2764 2020-08-26      2020-09-05
```

addAge()

```
> cdm$curso_omop_covid %>% addAge()
# Source:   table<og_008_1712431175> [?? x 5]
# Database: DuckDB v0.10.0 [apalomar@Windows 10 x64:R 4.2.3/C:\Users\apaloma]

  cohort_definition_id subject_id cohort_start_date cohort_end_date    age
      <int>          <dbl> <date>          <date>          <dbl>
1             1           493 2020-10-06      2020-10-16         15
2             1          3447 2021-02-21      2021-03-03         13
3             1          3596 2021-06-05      2021-06-15         34
4             1          4251 2021-04-19      2021-04-29         36
5             1          7310 2021-01-20      2021-01-30         89
6             1          3598 2020-05-20      2020-05-30         70
7             1          7813 2020-12-09      2020-12-19          1
8             1          9848 2021-01-09      2021-01-19          7
9             1          2764 2020-08-26      2020-09-05         35
10            1          7269 2021-01-12      2021-01-22         12
```

addAge()

También se pueden crear **grupos de edades** y darles un nombre

```
cdm$curso_omop_covid <- cdm$curso_omop_covid |>  
  addAge(ageGroup = list("18 años o menos"=c(0,18),  
                          "19 a 70 años"=c(19,70),  
                          "Más de 70 años"=c(71,150)))
```

addAge()

También se pueden crear **grupos de edades** y darles un nombre

	cohort_definition_id	subject_id	cohort_start_date	cohort_end_date	age	age_group
	<int>	<dbl>	<date>	<date>	<int>	<chr>
1	2	8	2021-01-08	2021-01-18	2	18 años o menos
2	2	36	2020-10-09	2020-10-19	4	18 años o menos
3	2	47	2020-11-26	2020-12-06	29	19 a 70 años
4	2	52	2020-12-08	2020-12-18	55	19 a 70 años
5	2	58	2020-07-12	2020-07-22	53	19 a 70 años
6	2	72	2020-12-02	2020-12-12	76	Más de 70 años
7	2	89	2020-11-27	2020-12-07	20	19 a 70 años
8	2	106	2020-08-16	2020-08-26	90	Más de 70 años
9	2	137	2020-09-08	2020-09-18	7	18 años o menos
10	2	156	2021-01-02	2021-01-12	46	19 a 70 años

addSex()

De forma parecida, podemos añadir **información sobre el sexo** de cada individuo.

```
addSex(  
  x,  
  sexName = "sex",  
  missingSexValue = "None",  
  name = NULL  
)
```

addSex()

De forma parecida, podemos añadir **información sobre el sexo** de cada individuo.

```
cdm$curs_omop_covid %>% addSex()
Source:   table<dbplyr_114> [?? x 6]
Database: DuckDB v0.9.2 [apistillo@windows 10 x64:R 4.3.2/C:\Users\APISTI~1\AppData
e568870545f68.duckdb]
  cohort_definition_id subject_id cohort_start_date cohort_end_date   age sex
      <int>          <dbl>   <date>          <date>      <dbl> <chr>
1             1           8 2021-01-08      2021-01-18         2 Male
2             1          36 2020-10-09      2020-10-19         4 Male
3             1          47 2020-11-26      2020-12-06        29 Female
4             1          52 2020-12-08      2020-12-18        55 Female
5             1          58 2020-07-12      2020-07-22        53 Female
6             1          72 2020-12-02      2020-12-12        76 Male
7             1          89 2020-11-27      2020-12-07        20 Male
8             1         106 2020-08-16      2020-08-26        90 Male
9             1         137 2020-09-08      2020-09-18         7 Male
10            1         156 2021-01-02      2021-01-12        46 Male
i more rows
```


addPriorObservation()

Podemos añadir la **historia previa (días o fecha)** en base al periodo de observación correspondiente.

Por ejemplo, podemos añadir la historia previa y después filtrar a los pacientes que tengan al menos un año de historia previa.

addPriorObservation()

Podemos añadir la **historia previa (días o fecha)** en base al periodo de observación correspondiente. Por ejemplo, podemos añadir la historia previa y después filtrar a los pacientes que tengan al menos un año de historia previa.

```
addPriorObservation(  
  x,  
  indexDate = "cohort_start_date",  
  priorObservationName = "prior_observation",  
  priorObservationType = "days",  
  name = NULL  
)
```

addPriorObservation()

Podemos añadir la **historia previa (días)** en base al periodo de observación correspondiente. Por ejemplo, podemos añadir la historia previa y después filtrar a los pacientes que tengan al menos un año de historia previa.

	cohort_definition_id	subject_id	cohort_start_date	cohort_end_date	age	age_group	prior_observation
	<int>	<dbl>	<date>	<date>	<int>	<chr>	<int>
1	2	8	2021-01-08	2021-01-18	2	18 años o menos	872
2	2	36	2020-10-09	2020-10-19	4	18 años o menos	1472
3	2	47	2020-11-26	2020-12-06	29	19 a 70 años	1678
4	2	52	2020-12-08	2020-12-18	55	19 a 70 años	2560
5	2	58	2020-07-12	2020-07-22	53	19 a 70 años	1980
6	2	72	2020-12-02	2020-12-12	76	Más de 70 años	2652
7	2	89	2020-11-27	2020-12-07	20	19 a 70 años	2491
8	2	106	2020-08-16	2020-08-26	90	Más de 70 años	3847
9	2	137	2020-09-08	2020-09-18	7	18 años o menos	2584
10	2	156	2021-01-02	2021-01-12	46	19 a 70 años	2228

addPriorObservation()

Podemos añadir la **historia previa** (fecha) en base al periodo de observación correspondiente. Por ejemplo, podemos añadir la historia previa y después filtrar a los pacientes que tengan al menos un año de historia previa.

```
> cdm$curso_omop_covid |> addPriorObservation(priorObservationType = "date")
# Source:   table<og_017_1741003343> [?? x 7]
# Database: DuckDB v1.1.3 [apa1omar@Windows 10 x64:R 4.4.1/C:\Users\apa1omar\AppData\Local\Temp\RtmpaQbHat\file23434
ac.duckdb]
```

	cohort_definition_id	subject_id	cohort_start_date	cohort_end_date	age	age_group	prior_observation
	<int>	<dbl>	<date>	<date>	<int>	<chr>	<date>
1	2	8	2021-01-08	2021-01-18	2	18 años o menos	2018-08-20
2	2	36	2020-10-09	2020-10-19	4	18 años o menos	2016-09-28
3	2	47	2020-11-26	2020-12-06	29	19 a 70 años	2016-04-23
4	2	52	2020-12-08	2020-12-18	55	19 a 70 años	2013-12-05
5	2	58	2020-07-12	2020-07-22	53	19 a 70 años	2015-02-09
6	2	72	2020-12-02	2020-12-12	76	Más de 70 años	2013-08-29
7	2	89	2020-11-27	2020-12-07	20	19 a 70 años	2014-02-01
8	2	106	2020-08-16	2020-08-26	90	Más de 70 años	2010-02-03
9	2	137	2020-09-08	2020-09-18	7	18 años o menos	2013-08-12
10	2	156	2021-01-02	2021-01-12	46	19 a 70 años	2014-11-27

addDemographics()

También podemos utilizar las tres funciones que acabamos de ver a la vez.

```
cdm$curs_omop_covid %>%  
  addAge() %>%  
  addSex() %>%  
  addPriorObservation()
```

addDemographics()

O podemos usar la función *addDemographics()*

```
addDemographics(  
  X,  
  indexDate = "cohort_start_date",  
  age = TRUE,  
  ageName = "age",  
  ageMissingMonth = 1,  
  ageMissingDay = 1,  
  ageImposeMonth = FALSE,  
  ageImposeDay = FALSE,  
  ageGroup = NULL,  
  missingAgeGroupValue = "None",  
  sex = TRUE,  
  sexName = "sex",  
  missingSexValue = "None",  
  priorObservation = TRUE,  
  priorObservationName = "prior_observation",  
  priorObservationType = "days",  
  futureObservation = TRUE,  
  futureObservationName = "future_observation",  
  futureObservationType = "days",  
  dateOfBirth = FALSE,  
  dateOfBirthName = "date_of_birth",  
  name = NULL  
)
```

Add intersections

Otra característica del paquete es la de poder ver de forma sencilla la **intersección entre dos tablas distintas del CDM**. Por ejemplo, queremos ver cuántas personas de una cohorte específica también tienen alguna otra condición o bien tienen una prescripción de algún medicamento, en una determinada ventana temporal.

Otra característica del paquete es la de poder ver de forma sencilla la intersección con dos tablas distintas del CDM. Por ejemplo, queremos ver cuántas personas de una cohorte específica también tienen alguna otra condición o bien tienen una prescripción de algún medicamento, en una determinada ventana temporal.

Hay 3 funciones principales

- `addCohortIntersectFlag()`
- `addCohortIntersectCount()`
- `addCohortIntersectDays()`

addCohortIntersectFlag()

Esta función crea una nueva columna indicando si el paciente en la cohorte 1 tiene (1) o no tiene (0) un tromboembolismo venoso en cualquier otro momento también.

addCohortIntersectFlag()

La función tiene más opciones

```
addCohortIntersectFlag(  
  x,  
  targetCohortTable,  
  targetCohortId = NULL,  
  indexDate = "cohort_start_date",  
  censorDate = NULL,  
  targetStartDate = "cohort_start_date",  
  targetEndDate = "cohort_end_date",  
  window = list(c(0, Inf)),  
  nameStyle = "{cohort_name}_{window_name}",  
  name = NULL  
)
```

Por ejemplo, se puede definir una, o varias, ventanas temporales.

addCohortIntersectFlag()

cohort_definition_id	subject_id	cohort_start_date	cohort_end_date	venous_thromboembolism_0_to_inf
<int>	<dbl>	<date>	<date>	<dbl>
1	3251	2021-01-27	2021-02-06	0
1	7937	2020-07-11	2020-07-21	0
1	942	2020-11-26	2020-12-06	1
1	3565	2021-01-27	2021-02-06	1
1	8924	2021-08-29	2021-09-08	1
1	6969	2020-12-23	2021-01-02	0
1	5105	2021-08-26	2021-09-05	1
1	8709	2021-01-06	2021-01-16	1
1	454	2021-03-19	2021-03-29	0
1	5055	2021-01-19	2021-01-29	0

addCohortIntersectCount ()

También se pueden **contar las intersecciones por persona.**

```
addCohortIntersectCount(  
  x,  
  targetCohortTable,  
  targetCohortId = NULL,  
  indexDate = "cohort_start_date",  
  censorDate = NULL,  
  targetStartDate = "cohort_start_date",  
  targetEndDate = "cohort_end_date",  
  window = list(c(0, Inf)),  
  nameStyle = "{cohort_name}_{window_name}",  
  name = NULL  
)
```

addCohortIntersectDays()

Calcula el número de días hasta un evento de una cohorte de referencia (target cohort).

```
addCohortIntersectDays(  
  x,  
  targetCohortTable,  
  targetCohortId = NULL,  
  indexDate = "cohort_start_date",  
  censorDate = NULL,  
  targetDate = "cohort_start_date",  
  order = "first",  
  window = c(0, Inf),  
  nameStyle = "{cohort_name}_{window_name}",  
  name = NULL  
)
```

Iniciación práctica al análisis de datos OMOP

Cohort Characteristics



Real World Epidemiology

Marzo 2025

- El objetivo del paquete ***CohortCharacteristics*** es resumir y describir características generales de una cohorte de pacientes.
- Calcula estadísticas descriptivas a nivel de cohorte, como edad media, distribución de género, comorbilidades, etc.



- Se ha desarrollado para el **Proyecto de Darwin EU®** por el equipo OxInfer.
- Está disponible en CRAN y se puede instalar fácilmente en R.
- ¿Cómo podemos **descargarlo**?

```
>install.packages("CohortCharacteristics")  
>library(CohortCharacteristics)
```

- <https://CRAN.R-project.org/package=CohortCharacteristics>

Funciones del paquete CohortCharacteristics

summariseCharacteristics	14
summariseCohortAttrition	17
summariseCohortCount	18
summariseCohortOverlap	18
summariseCohortTiming	19
summariseLargeScaleCharacteristics	20

Resumir información

tableCharacteristics	22
tableCohortAttrition	23
tableCohortCount	24
tableCohortOverlap	25
tableCohortTiming	26
tableLargeScaleCharacteristics	27

Crear tablas en base a los datos resumidos

plotCharacteristics	6
plotCohortAttrition	7
plotCohortCount	8
plotCohortOverlap	9
plotCohortTiming	10
plotComparedLargeScaleCharacteristics	12
plotLargeScaleCharacteristics	13

Representar gráficamente estos datos

- Resumir características en una cohorte.

```
summariseCharacteristics(  
  cohort,  
  cohortId = NULL,  
  strata = list(),  
  counts = TRUE,  
  demographics = TRUE,  
  ageGroup = NULL,  
  tableIntersectFlag = list(),  
  tableIntersectCount = list(),  
  tableIntersectDate = list(),  
  tableIntersectDays = list(),  
  cohortIntersectFlag = list(),  
  cohortIntersectCount = list(),  
  cohortIntersectDate = list(),  
  cohortIntersectDays = list(),  
  conceptIntersectFlag = list(),  
  conceptIntersectCount = list(),  
  conceptIntersectDate = list(),  
  conceptIntersectDays = list(),  
  otherVariables = character(),  
  otherVariablesEstimates = c("min", "q25", "median", "q75", "max",  
                               "percentage")  
)
```

- Formatear un objeto summarised_characteristics en una tabla visual.

```
table <- tableCharacteristics(  
  table1,  
  type = "gt",  
  formatEstimateName = c(`N (%)` = "<count> (<percentage>%)", N = "<count>",  
                          `Median [Q25 - Q75]` = "<median> [<q25> - <q75>]",  
                          `Mean (SD)` = "<mean> (<sd>)",  
                          Range = "<min> to <max>"),  
  header = c("group"),  
  split = c("group", "strata"),  
  groupColumn = NULL,  
  excludeColumns = c("result_id", "estimate_type", "additional_name", "additional_level"),  
  .options = list()  
)
```

- Se puede exportar fácilmente a Word

```
table |> gt::gtsave("table.docx")
```

tableCharacteristics

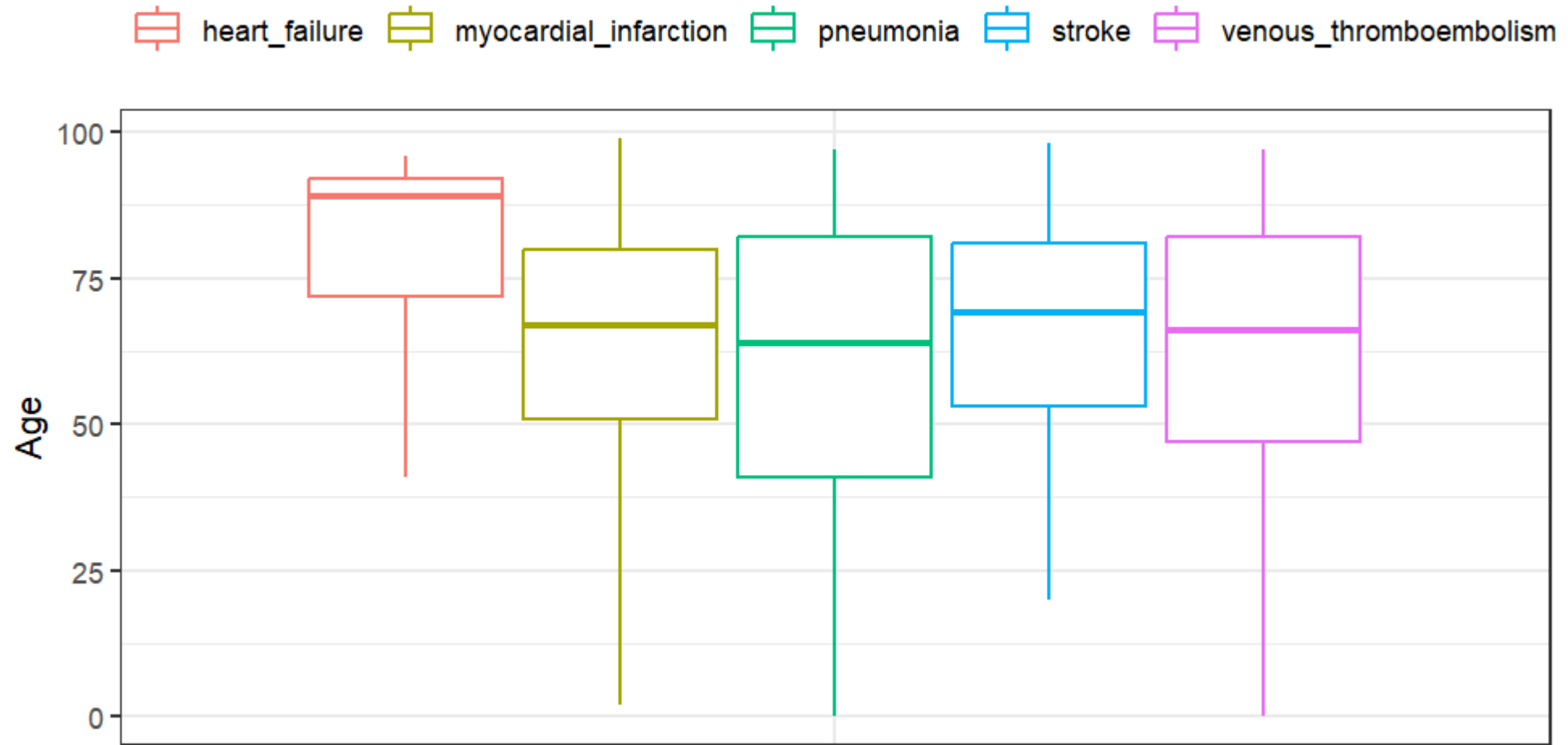
```
result <- cdm$curso_omop_covid |>
  filter(cohort_definition_id==1) |>
  addSex() |>
  summariseCharacteristics(
    strata = list("sex")
  )
result |> tableCharacteristics(header = "strata", groupColumn = "cohort_name")
```

CDM name	Variable name	Variable level	Estimate name	Sex		
				Overall	Female	Male
Covid19 diagnosis						
Synthea	Number records	-	N	964	497	467
	Number subjects	-	N	964	497	467
	Cohort start date	-	Median [Q25 - Q75]	2020-12-13 [2020-10-22 - 2021-01-31]	2020-12-17 [2020-10-19 - 2021-02-01]	2020-12-09 [2020-10-27 - 2021-01-31]
			Range	2020-03-15 to 2021-09-14	2020-03-16 to 2021-09-14	2020-03-15 to 2021-09-08
	Cohort end date	-	Median [Q25 - Q75]	2020-12-23 [2020-11-01 - 2021-02-10]	2020-12-27 [2020-10-29 - 2021-02-11]	2020-12-19 [2020-11-06 - 2021-02-10]
			Range	2020-03-25 to 2021-09-24	2020-03-26 to 2021-09-24	2020-03-25 to 2021-09-18
	Sex	Female	N (%)	497 (51.6%)	497 (100.0%)	-
		Male	N (%)	467 (48.4%)	-	467 (100.0%)

- Crea un ggplot a partir de la salida de summariseCharacteristics.

```
result <- cdm$condiciones |>  
  summariseCharacteristics()  
  
result |>  
  filter(variable_name == "Age") |>  
  plotCharacteristics(plotStyle = "boxplot", colour = "group_level")
```

plotCharacteristics



summariseLargeScaleCharacteristics

- Esta función se utiliza para resumir las características a gran escala de una tabla de cohorte.

```
summariseLargeScaleCharacteristics(  
  cohort,  
  strata = list(),  
  window = list(c(-Inf, -366), c(-365, -31), c(-30, -1), c(0, 0), c(1, 30), c(31, 365),  
                c(366, Inf)),  
  eventInWindow = NULL,  
  episodeInWindow = NULL,  
  indexDate = "cohort_start_date",  
  censorDate = NULL,  
  includeSource = FALSE,  
  minimumFrequency = 0.005,  
  excludedCodes = c(0)  
)
```

- Formatear un objeto summarised_characteristics en una tabla visual.

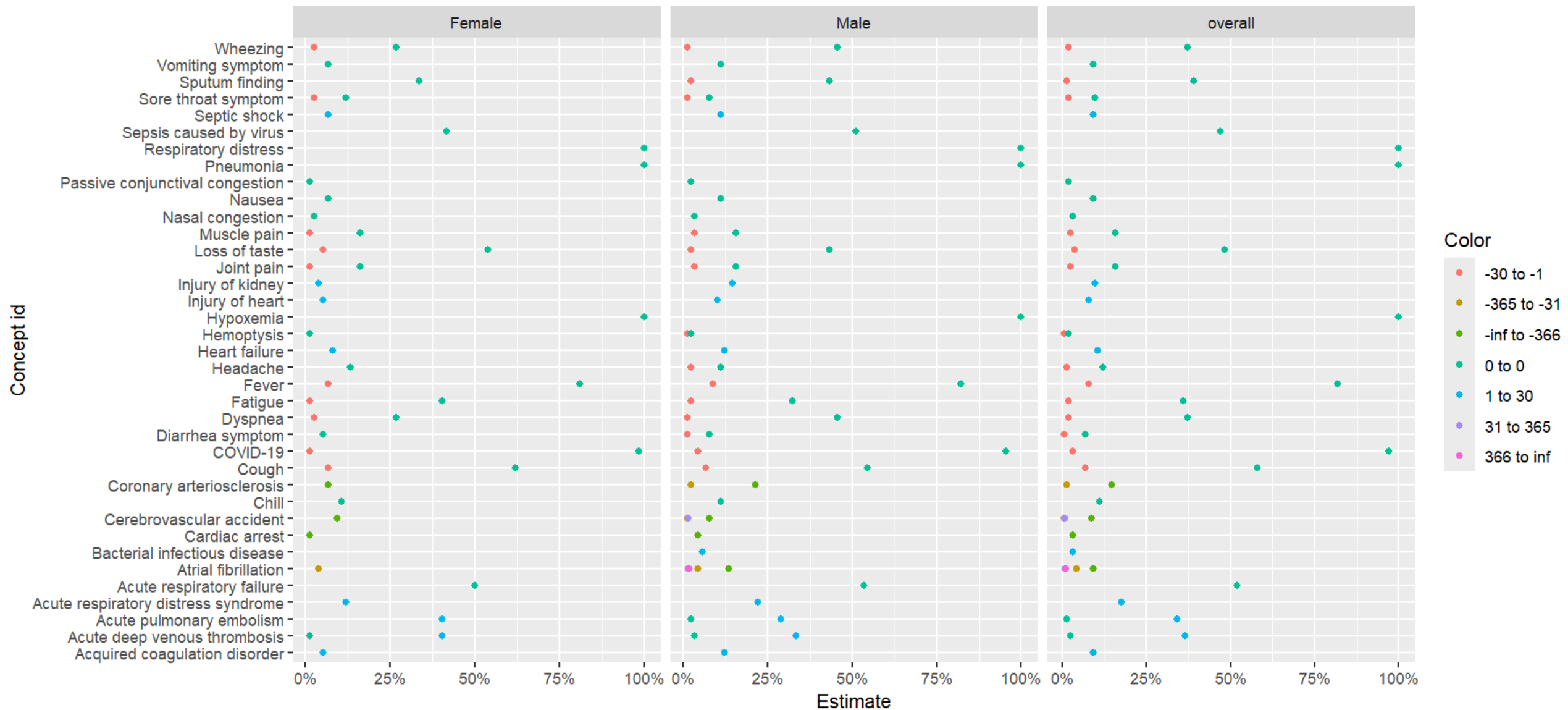
```
tableLargeScaleCharacteristics(  
  result,  
  type = "gt",  
  formatEstimateName = c(`N (%)` = "<count> (<percentage>%)"),  
  splitStrata = TRUE,  
  header = c("cdm name", "cohort name", "strata", "window name"),  
  topConcepts = NULL  
)
```







plotLargeScaleCharacteristics

- Crea un ggplot a partir de la salida de summariseLargeScaleCharacteristics.

```
plotLargeScaleCharacteristics(  
  data,  
  position = "horizontal",  
  splitStrata = FALSE,  
  facet = NULL,  
  colorVars = "variable_level"  
)
```

plotLargeScaleCharacteristics



-  Cohortes
-  1_PatientProfiles
-  PatientProfiles
-  patientprofiles_practica
-  patientprofiles_soluciones

Objetivo

El objetivo de esta práctica es caracterizar una población a partir de varias tablas de una base de datos en formato OMOP. Veremos cómo añadir de forma sencilla información demográfica, combinar tablas, seleccionar según criterios (flowchart) y hacer una tabla descriptiva.

¿Cómo funciona?

En este fichero encontrarás una serie de ejercicios que te proponemos, acompañados de teoría y pistas para su resolución. Puedes crear un script.R en el mismo directorio donde está este fichero, e ir resolviéndolos en el script. En cada ejercicio hay pistas, puesto que el curso es abierto a muchos perfiles y habrá gente que necesitará indicaciones sobre programación en R, otras sobre dónde encontrar las cosas en OMOP, y otras con todo. Además, en este mismo directorio, hay un fichero adicional con las soluciones a los ejercicios, para que puedas autocorregirte la práctica. Si tienes cualquier duda, pregunta a los docentes de apoyo en las prácticas, ¡e intenta utilizar las soluciones solo para corregir!