La BD del proyecto de Bayesiana

Chávez Santiago, Rafael. Barra Melendrez, Enki Alonso.
Jeshua Romero, Guadarrama. Montaño Castro, David.

2022-05-17

R Markdown

Importacion de las posibles paqueterias a utilizar

```
library(MASS)
library(tidyverse)
library(datos)
library(htmlwidgets)
library(survival)
library(AUC)
library(gdata)
library(dplyr)
library(DescTools)
library(openxlsx)
library(rjags)
```

Datos a utilizar

Mandamos a llamar los datos que vamos a ocupar para el modelo

```
tratamiento_art <- read.csv("/Users/enki/Documents/Modelo_Cox_Bayesiano/DBs/art_sim.csv")
informacion_basica <- read.csv("/Users/enki/Documents/Modelo_Cox_Bayesiano/DBs/basic_sim.csv")
seguimiento_paciente <- read.csv("/Users/enki/Documents/Modelo_Cox_Bayesiano/DBs/follow_sim.csv")
conteo_cd4 <- read.csv("/Users/enki/Documents/Modelo_Cox_Bayesiano/DBs/lab_cd4_sim.csv")
carga_viral <- read.csv("/Users/enki/Documents/Modelo_Cox_Bayesiano/DBs/lab_rna_sim.csv")
seguimiento_visitas <- read.csv("/Users/enki/Documents/Modelo_Cox_Bayesiano/DBs/visit_sim.csv")
```

Variables disponibles por cada archivo excel

— DATOS DE art_sim.csv tratamiento_art patient <– paciente site <– lugar de donde viene art_id <– tratamiento que sigue art_sd <– fecha de inicio de tratamiento art_ed <– fecha de termino del tratamiento - si no hay fecha de termino el paciente continuo con dicho tratamiento - art_rs <– razon de cambio de tratamiento

— DATOS DE basic_sim.csv informacion_basica baseline_d <– fecha de enrrolamiento del paciente male <– 1 si es hombre, 0 si es mujer age <– edad del paciente birth_d <– fecha de nacimiento hivdiagnosis_d <– fecha de diagnostico mode <– modo de transmision de la enfermedad birth_d_a <– exactitud de la fecha registrada

— DATOS DE follow_sim.csv seguimiento_paciente l_alive_d <– ultima fecha en la que sabemos que el paciente aun se encuentra vivo, en contacto con el sistema death_y <– registro de su muerte, 1 si murio 0 si no murio death_d <– fecha de muerte

Sobre el CD4 Y RNA(CV) Lo normal es que existan entre 500 y 1600 celulas CD4 por milimetro cubico de sangre rna es el numero de copias del virus por ml de sangre o por c/ml

— DATOS DE lab_cd4_sim.csv conteo_cd4 cd4_d <– fecha en la que se realizo el conteo de cd4 cd4_v <– valor de cd4 correspondiente a su fecha en la que se realizo el conteo

— DATOS DE lab_rna_sim.csv carga_viral rna_d <– fecha del conteo de la carga viral rna_v <– valor de rna correspondiente a su fecha en la que se realizo el conteo si nos marca un rna_v = -40 nos informa que en ese momento era indectable

— DATOS DE visit_sim.csv visit_d <– fechas de visita del paciente

Fechas

Cambiamos las fechas para que aparezcan de una forma que nos sea facil tratarlas, i.e, cambiamos el formato en el cual aparecen y revisamos que esten en el formato actualizado

```
tratamiento_art$art_sd <- as.Date(tratamiento_art$art_sd, "%Y-%m-%d")
class(tratamiento_art$art_sd)
```

```
## [1] "Date"
```

```
tratamiento_art$art_ed <- as.Date(tratamiento_art$art_ed, "%Y-%m-%d")
class(tratamiento_art$art_ed)
```

```
## [1] "Date"
```

```
informacion_basica$baseline_d <- as.Date(informacion_basica$baseline_d, "%Y-%m-%d")
class(informacion_basica$baseline_d)
```

```
## [1] "Date"
```

```
informacion_basica$birth_d <- as.Date(informacion_basica$birth_d, "%Y-%m-%d")
class(informacion_basica$birth_d)
```

```
## [1] "Date"
```

```
informacion_basica$hivdiagnosis_d <- as.Date(informacion_basica$hivdiagnosis_d, "%Y-%m-%d")
class(informacion_basica$hivdiagnosis_d)
```

```
## [1] "Date"
```

```
seguimiento_paciente$l_alive_d <- as.Date(seguimiento_paciente$l_alive_d,  "%Y-%m-%d")
class(seguimiento_paciente$l_alive_d)
```

```
## [1] "Date"
```

```
seguimiento_paciente$death_d <- as.Date(seguimiento_paciente$death_d,  "%Y-%m-%d")
class(seguimiento_paciente$death_d)
```

```
## [1] "Date"
```

```
conteo_cd4$cd4_d <- as.Date(conteo_cd4$cd4_d, "%Y-%m-%d")
class(conteo_cd4$cd4_d)
```

```
## [1] "Date"
```

```
carga_viral$rna_d <- as.Date(carga_viral$rna_d, "%Y-%m-%d")
class(carga_viral$rna_d)
```

```
## [1] "Date"
```

```
seguimiento_visitas$visit_d <- as.Date(seguimiento_visitas$visit_d, "%Y-%m-%d")
class(seguimiento_visitas$visit_d)
```

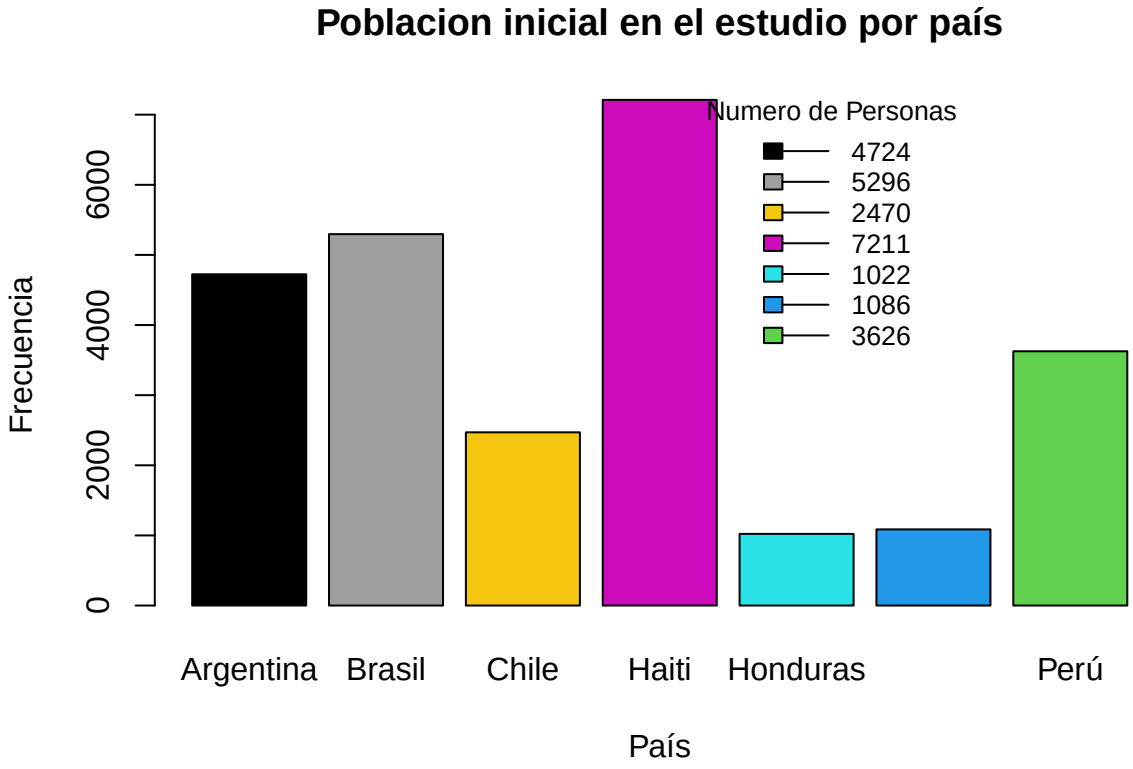
```
## [1] "Date"
```

Ya todas las variables estan en formato “Date”

Analisis de los datos base sin filtros. Lugares de origen de la poblacion de estudio.

Primero veremos de que lugares tenemos datos disponibles, dichos datos se encuentran en la tabla nombrada “informacion_basica”

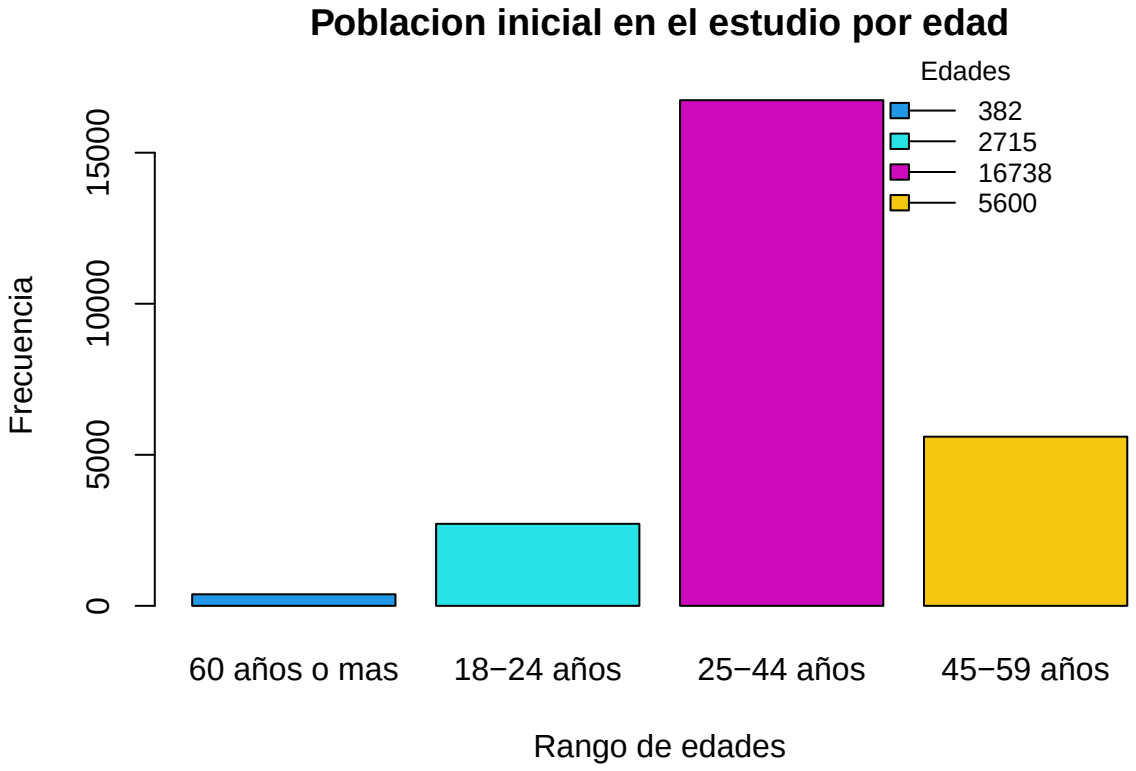
```
# View(informacion_basica)

# Hacemos una tabla sobre la informacion por paises
paises <- table(informacion_basica$site)
 View(paises)

# cambiamos los nombres de la tabla para que estos esten a corde a cada país
names(paises)<- c("Argentina", "Brasil", "Chile", "Haiti", "Honduras", "México", "Perú")

# Generamos un grafico para visualizar la informacion
barplot(paises, col = 9:2, main = "Poblacion inicial en el estudio por país", ylab = "Frecuencia",
        xlab = "País")

# veamos que clasificaiones generamos.-
legend("topright", legend = paises, fill = 9:2, title = "Numero de Personas", cex = .8, xpd = TRUE,
       inset = c(.2, -.02), bty = "n", lwd = 1)
```



Analisis de los datos base sin filtros. Edades de la poblacion de estudio categorizadas.

Veamos que rangos de edades son los que tenemos disponibles para el estudio, datos disponibles en "informacion_basica"

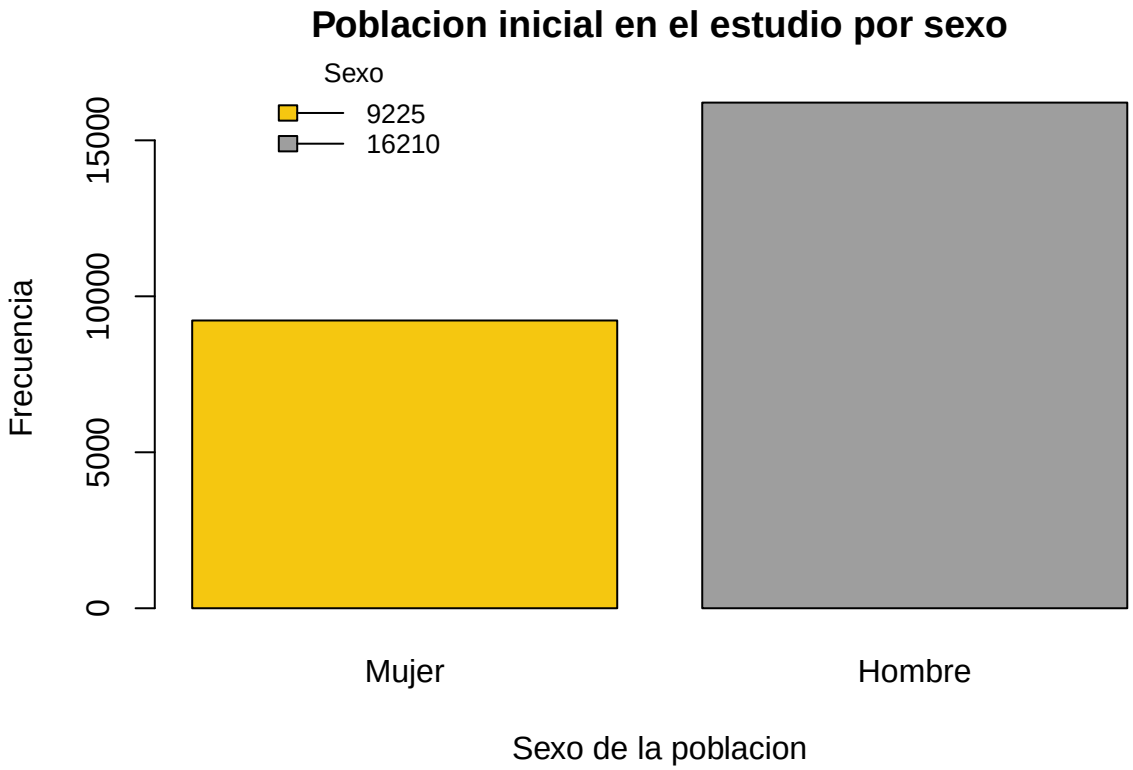
```
# View(informacion_basica)
edades <- table(informacion_basica$age)
# la tabla anmtes generada nos muestra edades diferentes, se hara una agrupacion por edades clasificand
# articulo.- http://www.conapo.gob.mx/work/models/CONAPO/Resource/1342/1/images/02introduccion.pdf

# Primero añadimos una variable vacioa llena de nulos para empezar la clasificacion
informacion_basica$age_c = as.Date(NA)
# Procedemos a realizar la separacion por grupos
informacion_basica$age_c = factor(ifelse(informacion_basica$age<=24, "age18-24", ifelse(informacion_bas

#generamos la tabla de las edades categorizadas
categoria_edades <- table(informacion_basica$age_c)
# View(categoria_edades)

# Asignamos nombres a la tabla
names(categoria_edades) <- c("60 años o mas", "18-24 años", "25-44 años", "45-59 años")

# realizamos un grafico que nos permita ver los rangos de las edades
barplot(categoria_edades, col = 4:8, main = "Poblacion inicial en el estudio por edad", ylab = "Frecuen
        xlab = "Rango de edades")

# veamos las clasificaciones generadas
legend("topright", legend = categoria_edades, fill = 4:8, title = "Edades", cex = .8, xpd = T, inset = 
       bty = "n",, lwd = 1)
```



Analisis de los datos base sin filtros. Sexo de la poblacion de estudio.

Veamos que porcentaje de hombres y mujeres se encuentran dentro de nuesta poblacion de estudio, datos disponibles en “informacion_basica”

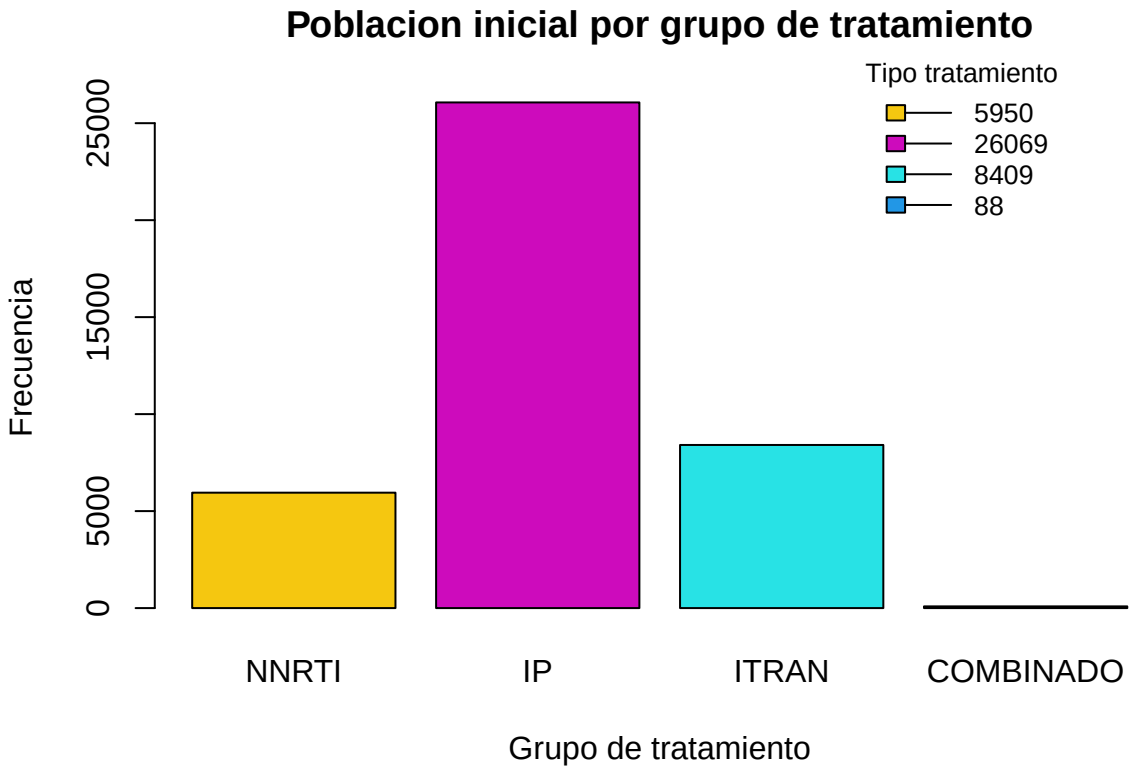
```
#View(informacion_basica)

# recordemos
#male               <-- 1 si es hombre, 0 si es mujer

# generamos la tabla de los sexos disponibles
sexo_poblacion <- table(informacion_basica$male)
# View(sexo_poblacion)

# asignamos clasificaion a los sexos
names(sexo_poblacion) <- c("Mujer", "Hombre")

# graficamos la tabla de sexos para visualizar la informacion
barplot(sexo_poblacion, col = 7:8, main = "Poblacion inicial en el estudio por sexo", ylab = "Frecuencia
        xlab = "Sexo de la poblacion")

# veamos las clasificaciones generadas
legend("topleft", legend = sexo_poblacion, fill = 7:8, title = "Sexo", cex = .8, xpd = T, inset = c(.1,
       bty = "n", lwd = 1)
```



Analisis de los datos base sin filtros. Tipo tratamiento de la poblacion de estudio categorizados.

Anteriormente separamos a la poblacion por grupo de edades y ahora lo haremos en base al tipo de tratamiento, estos grupos de tratamiento se separan en base al activo principal.” Datos siponibles en “tratamiento_art”

```r
#Generamos variables dummy para separar los tipos de tratamiento disponibles.

# tratamientos
value_EFV = "EFV"
tratamiento_art$dummy_EFV <- grepl(value_EFV, tratamiento_art$art_id, fixed = TRUE)
value_NVP = "NVP"
tratamiento_art$dummy_NVP <- grepl(value_NVP, tratamiento_art$art_id, fixed = TRUE)
tratamiento_art$groupNNRTI <- ifelse(tratamiento_art$dummy_EFV==1 | tratamiento_art$dummy_NVP==1,1,0)

value_LPV = "LPV"
tratamiento_art$dummy_LPV <- grepl(value_NVP, tratamiento_art$art_id, fixed = T)
value_RTV = "RTV"
tratamiento_art$dummy_RTV <- grepl(value_RTV, tratamiento_art$art_id, fixed = T)
value_ATV = "ATV"
tratamiento_art$dummy_ATV <- grepl(value_ATV, tratamiento_art$art_id, fixed = T)
value_SQV = "SQV"
tratamiento_art$dummy_SQV <- grepl(value_SQV, tratamiento_art$art_id, fixed = T)
value_DRV = "DRV"
tratamiento_art$dummy_DVR <- grepl(value_DRV, tratamiento_art$art_id, fixed = T)
tratamiento_art$groupIP <- ifelse(tratamiento_art$dummy_LPV | tratamiento_art$dummy_RTV | tratamiento_ar

value_DLG ="DLG"
tratamiento_art$dummy_DLG <- grepl(value_DLG, tratamiento_art$art_id, fixed = T)
value_RAL = "RAL"
tratamiento_art$dummy_RAL <- grepl(value_RAL, tratamiento_art$art_id, fixed = T)
tratamiento_art$groupITRAN <- ifelse(tratamiento_art$dummy_DLG == 1 | tratamiento_art$dummy_RAL == 1, 1

tratamiento_art$group_art <- ifelse(tratamiento_art$groupNNRTI==1,1,
                        ifelse(tratamiento_art$groupIP==1,2,
                               ifelse(tratamiento_art$groupITRAN==1,3,0)))

# los que tienen EFV o NVP son un grupo (NNRTI),
# los que tienen LPV/RTV o ATV/RTV ó SQV ó DRV son otro grupo (IP).
# medicamentos como DLG o RAL, si es así ellos conforman otro grupo (ITRAN).
# Los restantes se asignan a tratamiento combinado

# hacemos la tabulacion de los datos por grupo de tratamiento
tipos_tratamiento <- table(tratamiento_art$group_art)
# View(tipos_tratamiento)

#asignamos nombres de cada grupo de tratamiento
names(tipos_tratamiento) <-c("NNRTI", "IP", "ITRAN", "COMBINADO")

# graficamos la tabla de los tratamientos
barplot(tipos_tratamiento, col = 7:3, main = "Poblacion inicial por grupo de tratamiento",
        ylab = "Frecuencia", xlab = "Grupo de tratamiento")

#veamos las clasificaciones generadas
legend("topright", legend = tipos_tratamiento, fill = 7:3, title = "Tipo tratamiento", cex = .8, xpd =
```



Limpieza_1 de variables que no usaremos en el modelo.

Comenzaremos con la limpieza de las tablas con las que contamos eliminando variables que no vamos a usar y la seleccion de los individuos que participaran en el estudio.

```
# Eliminacion de variables innecesarias.

# Empezamos con la base de datos de tratamientos

tratamiento_art$pi <- NULL
tratamiento_art$nnrti1 <- NULL
tratamiento_art$nnrti2 <- NULL
tratamiento_art$nnrti <- NULL
tratamiento_art$nrti <- NULL
tratamiento_art$t20 <- NULL
tratamiento_art$ccr5 <- NULL
tratamiento_art$ii1 <- NULL
tratamiento_art$ii2 <- NULL
tratamiento_art$rtv_drug <- NULL
tratamiento_art$numdrugs <- NULL
tratamiento_art$art_class <- NULL
tratamiento_art$X <- NULL
tratamiento_art$dummy_ATV <- NULL
tratamiento_art$dummy_DLG <- NULL
tratamiento_art$dummy_DVR <- NULL
tratamiento_art$dummy_EFV <- NULL
tratamiento_art$dummy_LPV <- NULL
tratamiento_art$dummy_NVP <- NULL
tratamiento_art$dummy_RAL <- NULL
tratamiento_art$dummy_RTV <- NULL
```

```
tratamiento_art$dummy_SQV <- NULL
tratamiento_art$groupIP <- NULL
tratamiento_art$groupITRAN <- NULL
tratamiento_art$groupNNRTI <- NULL

# Seguimos con la base de datos de la informacion basica de cada paciente
informacion_basica$aids_y <- NULL
informacion_basica$aids.miss <- NULL
informacion_basica$aids_cl_d <- NULL
informacion_basica$aids_cl_y<-NULL
informacion_basica$recart_y<-NULL
informacion_basica$aids_d<-NULL
informacion_basica$mode_oth<-NULL
informacion_basica$clinicaltrial_y<-NULL
informacion_basica$baseline_d_num <- NULL
informacion_basica$hivdiagnosis_d_num <- NULL

# Seguimos con los datos de seguimiento de los pacientes
seguimiento_paciente$drop_rs_oth<-NULL
seguimiento_paciente$drop_rs<-NULL
seguimiento_paciente$death_d_a<-NULL
seguimiento_paciente$death_d <- NULL

# Seguimos con las cuentas de CD4
seguimiento_visitas$cdcstage<-NULL
seguimiento_visitas$whostage<-NULL



# Visualizamos las variables de cada tabla que nos restan
head(carga_viral)
```

```
##   patient      site      rna_d  rna_v
## 1    ar.1 argentina 2007-01-19  74724
## 2    ar.1 argentina 2013-02-04    -40
## 3    ar.1 argentina 2011-11-03    399
## 4    ar.1 argentina 2010-11-16 407000
## 5    ar.1 argentina 2009-08-26    -50
## 6    ar.1 argentina 2008-05-07    -50
```

```
head(conteo_cd4)
```

```
##   patient      site      cd4_d cd4_v time
## 1    ar.1 argentina 2007-01-19   405    0
## 2    ar.1 argentina 2008-05-07   490  474
## 3    ar.1 argentina 2009-08-26   238  950
## 4    ar.1 argentina 2010-11-16   451 1397
## 5    ar.1 argentina 2011-11-03   811 1749
## 6    ar.1 argentina 2013-02-04   238 2208
```

```
head(informacion_basica)
```

```
##   patient      site baseline_d male      age    birth_d hivdiagnosis_d
## 1    ar.1 argentina 2007-04-13    1 34.16329 1973-02-12     2007-04-13
## 2    ar.2 argentina 2010-07-06    0 45.89359 1964-08-13     1999-01-07
```

```
## 3    ar.3 argentina 2011-03-28    1 47.23421 1964-01-02     1999-08-13
## 4    ar.4 argentina 2002-04-19    0 31.46904 1970-10-30     1996-09-14
## 5    ar.5 argentina 2004-12-27    1 36.88613 1968-02-07     2002-10-23
## 6    ar.6 argentina 2008-08-19    1 35.46968 1973-03-01     2008-08-19
##                                 mode birth_d_a    age_c
## 1               Homosexual contact         D age25-44
## 2              Injecting drug user         D age45-59
## 3             Heterosexual contact         D age45-59
## 4                          Unknown         D age25-44
## 5 Transfusion nonhemophilia related         D age25-44
## 6             Heterosexual contact         D age25-44
```

```
head(seguimiento_paciente)
```

```
##   patient      site  l_alive_d death_y
## 1    ar.1 argentina 2013-02-04       0
## 2   ar.10 argentina 2013-02-13       0
## 3  ar.100 argentina 2013-07-12       0
## 4 ar.1000 argentina 2012-11-10       0
## 5 ar.1001 argentina 2013-06-21       0
## 6 ar.1002 argentina 2014-01-25       0
```

```
head(seguimiento_visitas)
```

```
##   patient      site    visit_d
## 1    ar.1 argentina 2007-04-13
## 2    ar.2 argentina 2010-07-06
## 3    ar.3 argentina 2011-03-28
## 4    ar.4 argentina 2002-04-19
## 5    ar.5 argentina 2004-12-27
## 6    ar.6 argentina 2008-08-19
```

```
head(tratamiento_art)
```

```
##   patient      site      art_id     art_sd     art_ed                art_rs
## 1    ar.1 argentina 3TC,AZT,NVP 2007-05-16 2007-05-28 Toxicity Dermatologic
## 2    ar.1 argentina 3TC,AZT,EFV 2007-05-30 2007-07-04               Unknown
## 3    ar.1 argentina 3TC,ABC,AZT 2007-08-03       <NA>
## 4   ar.10 argentina 3TC,AZT,EFV 2002-02-07       <NA>
## 5  ar.100 argentina 3TC,ABC,AZT 2006-06-01 2006-07-16               Unknown
## 6  ar.100 argentina 3TC,AZT,EFV 2006-07-16 2006-09-26               Unknown
##   group_art
## 1         1
## 2         1
## 3         0
## 4         1
## 5         0
## 6         1
```

Union para generar los datos con los que trabajaremos

Uniremos las "bases" que tenemos para generar un archivo en el cual todo este juto para posteriormente comenzar con la limpieza de los datos

```
# comprovacion rapida de existencia de na's
sum(c(is.na(tratamiento_art$patient),
is.na(tratamiento_art$site),
```

```
is.na(tratamiento_art$art_id),
is.na(tratamiento_art$art_sd),
is.na(tratamiento_art$art_ed),
is.na(tratamiento_art$art_rs),
is.na(tratamiento_art$group_art)))
```

```
## [1] 16466
```

```
# 0

sum(c(is.na(informacion_basica$patient),
is.na(informacion_basica$site),
is.na(informacion_basica$baseline_d),
is.na(informacion_basica$male),
is.na(informacion_basica$age),
is.na(informacion_basica$birth_d),
is.na(informacion_basica$hivdiagnosis_d),
is.na(informacion_basica$mode),
is.na(informacion_basica$birth_d_a),
is.na(informacion_basica$age_c)))
```

```
## [1] 0
```

```
# 0

sum(c(is.na(seguimiento_paciente$patient),
is.na(seguimiento_paciente$site),
is.na(seguimiento_paciente$l_alive_d),
is.na(seguimiento_paciente$death_y)))
```

```
## [1] 0
```

```
# 0

sum(c(is.na(carga_viral$patient),
is.na(carga_viral$site),
is.na(carga_viral$rna_d),
is.na(carga_viral$rna_v)))
```

```
## [1] 3998
```

```
# 2183
# menos del 5% de datos faltantes, omitimos na's
carga_viral <- na.omit(carga_viral)

sum(c(is.na(conteo_cd4$patient),
is.na(conteo_cd4$site),
is.na(conteo_cd4$cd4_d),
is.na(conteo_cd4$cd4_v),
is.na(conteo_cd4$time)))
```

```
## [1] 5445
```

```
# 3630
#menos del 5% de datos faltantes, omitimos na's
conteo_cd4 <- na.omit(conteo_cd4)
```

```r
sum(c(is.na(seguimiento_visitas$patient),
is.na(seguimiento_visitas$site),
is.na(seguimiento_visitas$visit_d)))
```

```
## [1] 0
```

```r
#0


# tratamiento_art <- read.csv("art_sim.csv")
# informacion_basica <- read.csv("basic_sim.csv")
# seguimiento_paciente <- read.csv("follow_sim.csv")
# conteo_cd4 <- read.csv("lab_cd4_sim.csv")
# carga_viral <- read.csv("lab_rna_sim.csv")
# seguimiento_visitas <- read.csv("visit_sim.csv")

# union de las tablas
b1 <- merge(carga_viral, conteo_cd4, by=c("patient", "site"), all = T)
sum(is.na(b1))
```

```
## [1] 125216
```

```r
# View(b1)
# b2 <- merge(informacion_basica, b1, by=c("patient", "site"), all = T)
# View(b2)
# b3 <- merge(seguimiento_paciente, b2, by=c("patient", "site"), all = T)
# View(b3)
# datos_a_usar <- merge(tratamiento_art, b3, by=c("patient", "site"), all = T)
```