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)

## ── Attaching packages ─────────────────────────────────────── tidyverse 1.3.1 ──

## ✓ ggplot2 3.3.5 ✓ purrr 0.3.4  
## ✓ tibble 3.1.6 ✓ dplyr 1.0.8  
## ✓ tidyr 1.2.0 ✓ stringr 1.4.0  
## ✓ readr 2.1.2 ✓ forcats 0.5.1

## ── Conflicts ────────────────────────────────────────── tidyverse\_conflicts() ──  
## x dplyr::filter() masks stats::filter()  
## x dplyr::lag() masks stats::lag()  
## x dplyr::select() masks MASS::select()

library(datos)  
library(htmlwidgets)  
library(survival)  
library(AUC)

## AUC 0.3.2

## Type AUCNews() to see the change log and ?AUC to get an overview.

library(gdata)

## gdata: read.xls support for 'XLS' (Excel 97-2004) files ENABLED.

##

## gdata: read.xls support for 'XLSX' (Excel 2007+) files ENABLED.

##   
## Attaching package: 'gdata'

## The following objects are masked from 'package:dplyr':  
##   
## combine, first, last

## The following object is masked from 'package:purrr':  
##   
## keep

## The following object is masked from 'package:stats':  
##   
## nobs

## The following object is masked from 'package:utils':  
##   
## object.size

## The following object is masked from 'package:base':  
##   
## startsWith

library(dplyr)  
library(DescTools)

## Registered S3 method overwritten by 'DescTools':  
## method from   
## reorder.factor gdata

##   
## Attaching package: 'DescTools'

## The following object is masked from 'package:gdata':  
##   
## reorder.factor

library(openxlsx)  
library(rjags)

## Loading required package: coda

## Linked to JAGS 4.3.1

## Loaded modules: basemod,bugs

## 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)

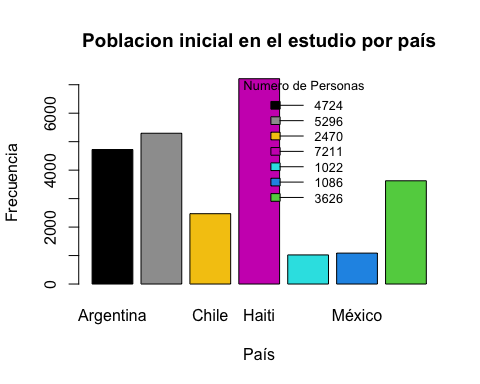
## [1] "Date"

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”

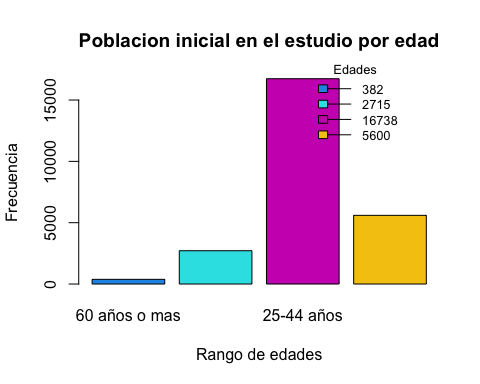
# View(informacion\_basica)  
  
# Hacemos una tabla sobre la informacion por paises   
paises <- table(informacion\_basica$site)  
 View(paises)  
  
# cambiamos los nombres de la tabla para que estos esten a corde a cada país  
names(paises)<- c("Argentina", "Brasil", "Chile", "Haiti", "Honduras", "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”

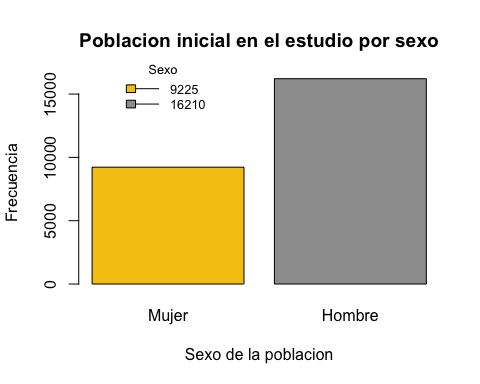
# View(informacion\_basica)  
edades <- table(informacion\_basica$age)  
# la tabla anmtes generada nos muestra edades diferentes, se hara una agrupacion por edades clasificando a jovenes de los 18 a los 24 años, adultos jovenes antes de los entre los 25 y 44 años, adultos maduros entre los 45 y 59 años y finalmente los adultos mayores que tienen mas de 60 años  
# 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$age<=24, "age18-24", ifelse(informacion\_basica$age<=44, "age25-44", ifelse(informacion\_basica$age<=59, "age45-59", "age>=60"))))  
  
#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 = "Frecuencia",  
 xlab = "Rango de edades")  
  
# veamos las clasificaciones generadas   
legend("topright", legend = categoria\_edades, fill = 4:8, title = "Edades", cex = .8, xpd = T, inset = c(.1, -.1),  
 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”

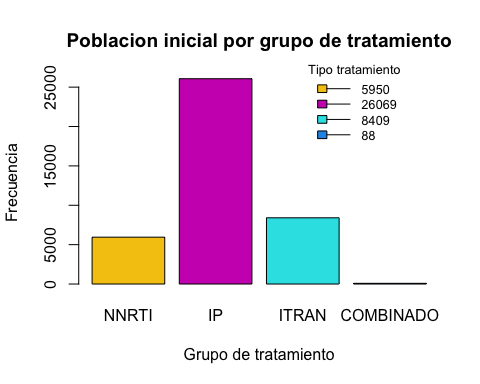
#View(informacion\_basica)  
  
# recordemos   
#male <-- 1 si es hombre, 0 si es mujer  
  
# generamos la tabla de los sexos disponibles  
sexo\_poblacion <- table(informacion\_basica$male)  
# View(sexo\_poblacion)  
  
# asignamos clasificaion a los sexos  
names(sexo\_poblacion) <- c("Mujer", "Hombre")  
  
# 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, -.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”

#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$dummy\_RTV | tratamiento\_art$dummy\_ATV | tratamiento\_art$dummy\_SQV | tratamiento\_art$dummy\_DVR, 1,0)  
  
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, 0)  
  
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 = T, inset = c(.1, -.1), bty = "n", lwd = 1)



## 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)  
  
sum(c(is.na(seguimiento\_visitas$patient),  
is.na(seguimiento\_visitas$site),  
is.na(seguimiento\_visitas$visit\_d)))

## [1] 0

#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

# 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)