Universitat Oberta de Catalunya

Missing data: Imputación múltiple en bases de datos pequeñas

Juan Vicente Hernández Villena

Bioestadística - Análisis de datos
Máster en Bioinformática y Bioestadística

Núria Pérez Álvarez
(Carles Ventura Royo)

junio 2022

FICHA DEL TRABAJO FINAL

Título:	*Missing data:* Imputación múltiple en bases de datos pequeñas
Nombre autor/a:	Juan Vicente Hernández Villena
Nombre PDC:	Núria Pérez Álvarez
Nombre PRA:	Carles Ventura Royo
Fecha de entrega:	junio 2022
Titulación:	Máster en Bioinformática y Bioestadística
Área:	Bioestadística - Análisis de datos
Idioma:	Castellano
Núm. de créditos:	15
Palabras clave:	Missings, Imputación, VIH

Resumen

Un dato faltante o *missing* es información relevante para el análisis pero que, por diversos factores, no pudo ser registrado y como consecuencia, está ausente en las bases de datos de cualquier tipo, incluyendo los registros longitudinales. Los mismos, si no son tomados en cuenta, pueden influir significativamente en el poder estadístico, la integridad del análisis, en estimaciones sesgadas y en la calidad de los resultados, por lo que es necesario un tratamiento correcto sobre los mismos, basado en las características que presenten. Este trabajo buscó poner a prueba diversas alternativas de imputación múltiple mediante la estrategia *PMM*, tratamiento moderno que ha generado buenos resultados, sobre los datos faltantes de una base de datos de pacientes con VIH, en tres escenarios de acuerdo al tamaño muestral. Una vez realizada las imputaciones, los resultados se pusieron a prueba, reproduciendo el análisis realizado en la publicación original de donde provienen los datos (comparación entre tratamientos), obteniendo resultados similares a los descritos. Se describió un flujo de trabajo recomendado para el análisis previo y tratamiento de *missings*, cuando la base de datos es longitudinal, con un tamaño muestral pequeño, sin distribución normal y con altos porcentajes de datos faltantes, características que no suelen ponerse a prueba cuando se evalúan este tipo de métodos, pero que son frecuentes en diversos campos de investigación como la biología y la salud.

Abstract

A missing data is a relevant information to the analysis, but due to different factors could not be recorded, and as consequence, is absent in any kind of dataset, including longitudinal records. If they are not taken into account, they will influence significantly the statistical power, the analysis integrity, the bias, and the quality of the results. So, it is necessary to use the correct treatment on them, based on their characteristics. This study carried out several multiple imputation alternatives by the *PMM* strategy, a modern treatment that has generated good results, on an HIV dataset missing data, in three scenarios according to the sample size. Once the imputations were done, the results were put to test, reproducing the analysis carried out in the original paper where the data comes from (comparison between treatments), obtaining similar results to those described. An improved workflow was described for the missings' previous analysis and treatment, with a longitudinal dataset, with small sample size, non-normal distribution, and with high percentages of missing data, characteristics that are not common when these methods are evaluated, but which are common in the most research fields such as biology and healthcare.

Índice general

Índice de figuras **7**

Índice de cuadros **9**

1. Introducción **10**
1.1. Contexto y justificación 10
1.1.1. Descripción general 10
1.1.2. Justificación del TFM 11
1.2. Objetivos 11
1.3. Enfoque y método 12
1.4. Planificación 12
1.4.1. Tareas 12
1.4.2. Hitos 13
1.4.3. Análisis de riesgo 14
1.4.4. Cronograma 16
1.5. Breve resumen de las contribuciones 17

2. Estado del arte **18**
2.1. Datos longitudinales 18
2.2. Datos faltantes 19
2.2.1. Teoría de los datos faltantes 20
2.2.2. Estrategias para el tratamiento de datos faltantes 23
2.2.3. Impacto en bases de datos pequeñas 27
2.3. Descripción de la base de datos 27

3. Metodología **29**
3.1. Creación de los escenarios de estudio . . . 29
3.2. Análisis descriptivo de los *missings* . . . 30
3.3. Tratamiento de *missings* . . . 31
3.4. Comparación de resultados . . . 32

4. Resultados **34**
4.1. Creación de los escenarios de estudio . . . 34
4.2. Análisis descriptivo de los *missings* . . . 36
4.3. Tratamiento de *missings* . . . 39
4.4. Comparación de resultados . . . 42

5. Discusión **45**

6. Conclusiones **49**
6.1. Líneas de futuro . . . 50
6.2. Seguimiento de la planificación . . . 50

7. Glosario **52**

8. Bibliografía **54**

A. Anexos **58**
A.1. Análisis de *missings* . . . 58
A.1.1. Gráficos QQplot . . . 58
A.1.2. Patrón de *missings* . . . 58
A.1.3. matriz predictora . . . 58
A.2. *Stripplots* . . . 58
A.2.1. Alternativa 1: Por defecto . . . 58
A.2.2. Alternativa 2: *PMM* . . . 58
A.2.3. Alternativa 3: *Midastouch* . . . 58
A.3. *Densityplots* . . . 58
A.3.1. Alternativa 1: Por defecto . . . 58
A.3.2. Alternativa 2: *PMM* . . . 58

B. Código R **69**

Índice de figuras

1.1. Cronograma 16

4.1. Estructura de la base de datos *VIH* 35
4.2. *Boxplot* Valor absoluto de los CD4 (Izq.) y Log_{10} carga viral de ARN del VIH, log_{10}(copias/mL) (Der.) en el tiempo. 35
4.3. Aumento del porcentaje de pérdidas del valor absoluto de CD4 (izq.) y Log_{10}carga viral de ARN del VIH, log_{10}(copias/mL) (der.) en el tiempo. 37
4.4. Patrón de *missings* por variables y observaciones de la base de datos VIH. . . . 38
4.5. *Fluxplots* realizados para las tres bases de datos. 39
4.6. *Stripplot* de los valores imputados de la base de datos VIH con m = 3 41
4.7. *Densityplot* de los valores imputados de la base de datos VIH con m = 3 42
4.8. *Densityplot* de los valores imputados de la base de datos VIH2 con m = 3 43
4.9. *Densityplot* de los valores imputados de la base de datos VIH3 con m = 3 43

A.1. Gráficos QQplot (Arriba) Variable CD4A. (Abajo) log_{10} de Carga viral 59
A.2. Patrón de *missings* por variables y observaciones de la base de datos VIH2 . . . 59
A.3. Patrón de *missings* por variables y observaciones de la base de datos VIH3 . . . 60
A.4. matriz predictora de la función *mice* utilizada con la base de datos VIH3. En rojo los cambios de valor 1 a 0. En azul los cambios de valor 0 a 1. 61
A.5. *Stripplot* de los valores imputados de la base de datos VIH con m = 3 62
A.6. *Stripplot* de los valores imputados de la base de datos VIH2 con m = 3 62
A.7. *Stripplot* de los valores imputados de la base de datos VIH3 con m = 3 63
A.8. *Stripplot* de los valores imputados de la base de datos VIH con m = 3 63
A.9. *Stripplot* de los valores imputados de la base de datos VIH2 con m = 3 64
A.10. *Stripplot* de los valores imputados de la base de datos VIH3 con m = 3 64

A.11. *Stripplot* de los valores imputados de la base de datos VIH2 con m = 3 65
A.12. *Stripplot* de los valores imputados de la base de datos VIH3 con m = 3 65
A.13. *Densityplot* de los valores imputados de la base de datos VIH con m = 3 66
A.14. *Densityplot* de los valores imputados de la base de datos VIH2 con m = 3 66
A.15. *Densityplot* de los valores imputados de la base de datos VIH3 con m = 3 67
A.16. *Densityplot* de los valores imputados de la base de datos VIH con m = 3 67
A.17. *Densityplot* de los valores imputados de la base de datos VIH2 con m = 3 68
A.18. *Densityplot* de los valores imputados de la base de datos VIH3 con m = 3 68

Índice de cuadros

3.1. Variables seleccionadas para las bases de datos 30

4.1. Características de las bases de datos (BD) 36
4.2. Porcentajes de *missings* de cada base de datos 36
4.3. Tabla de resultados obtenidos a un nivel de significancia del 95 % tras comparar los grupos de tratamiento y control de cada base de datos, para las variables CD4A y Carga viral de la semana 48. En negrita, aquellos p valores menores a 0.05. 44

Capítulo 1

Introducción

1.1. Contexto y justificación

1.1.1. Descripción general

Un proyecto de investigación consiste en responder una o más preguntas sobre un tema en particular y para ello es necesario recopilar información para su posterior análisis. Sin embargo, es muy frecuente que por diversas razones no sea posible obtener un registro, considerándose como un valor ausente. Un dato faltante es un valor no observado que, en caso contrario, presentaría un valor significativo para el análisis [1], en otras palabras, un dato faltante es información relevante para el análisis pero que está ausente en la base de datos.

Tal como se ha descrito, los datos faltantes o *missings* pueden estar presentes en cualquier tipo de base de datos y los registros longitudinales no son la excepción. Un estudio longitudinal se basa en colectar información a un conjunto de individuos en el tiempo [1], por lo que es un tipo de bases de datos frecuente en el ámbito de la medicina y salud.

Es importante prestarle atención a los valores ausentes y tratarlos adecuadamente, ya que los mismos pueden influir significativamente en el poder estadístico, la integridad del análisis, estimaciones sesgadas y resultados incorrectos [2, 3]. A través de los años, se han desarrollado una gran cantidad de métodos a emplear sobre los datos faltantes cuando es el momento de analizar la información obtenida, los cuales han sido puestos a prueba por diversos autores tanto en simulaciones como con datos reales, generando conclusiones en cuanto a sus requisitos,

bondades y limitaciones [3, 4, 5].

A pesar de todas las alternativas para el tratamiento correcto de los datos faltantes, no ha sido posible encontrar un método único, principalmente porque no todas las bases de datos gozan de las mismas características [2]. Por otro lado, pocos han sido los trabajos enfocados en poner a prueba las diferentes técnicas sobre bases de datos de pocas observaciones [6, 7, 8, 9] y/o con tasas altas de datos faltantes [5, 10].

1.1.2. Justificación del TFM

Por lo tanto, con base en los trabajos realizados por Barnes [7] y Kleinke [9], así como los pasos claves mencionados por Mirzaei [10], con este trabajo se busca poner a prueba aquellos métodos que en la literatura han demostrado generar buenos resultados ante el tratamiento de datos faltantes, con el fin de explorar sus bondades y limitaciones sobre una base de datos longitudinal a diferentes tamaños muestrales.

1.2. Objetivos

1. Contextualizar el efecto de los datos faltantes en una base de datos longitudinal, así como abordar de manera teórica los diferentes métodos que han tenido mayor impacto en la literatura.
   a) Revisar la bibliografía acerca de los datos faltantes: patrones, mecanismos y métodos.
2. Poner en práctica métodos modernos de Imputación Múltiple enfocados en tratar datos faltantes presentes en una base de datos longitudinal a diferentes tamaños muestrales.
   a) Buscar una base de datos longitudinal publicada.
   b) Aplicar estadística descriptiva sobre los datos, así como identificar patrones y mecanismos de datos faltantes.
   c) Generar las bases de datos a diferentes escenarios de acuerdo al tamaño de datos faltantes.
   d) Poner a prueba los métodos escogidos sobre las bases de datos generadas.
   e) Analizar y comparar los resultados con los resultados de la base de datos original.

1.3. Enfoque y método

La estructura del trabajo final de máster está conformada por dos segmentos principales, teórico y práctico.

El segmento teórico consiste en repasar toda la información relacionada a los datos faltantes en una base de datos, explorando los diferentes patrones, mecanismos y diversos métodos de tratamientos encontrados en la literatura, incluyendo información sobre sus ventajas, supuestos y limitaciones y las diferentes librerías de R que trabajen con estos métodos, así como una breve explicación sobre la base de datos con la que se va a trabajar.

El segmento práctico evalúa la base de datos obtenida con los diferentes tratamientos empleados, mediante estadística descriptiva, así como el análisis de los datos faltantes, mediante la prueba de Little para el mecanismo MCAR [11], y la visualización del patrón de los datos. Seguido a esto, se puso a prueba algunos de los métodos de imputación múltiple mencionados en el primer apartado sobre las bases de datos con los diferentes tratamientos.

Para la búsqueda y descarga bibliográfica se utilizó el motor de búsqueda *Google académico* y la biblioteca de la Universidad Abierta de Cataluña. Como gestor bibliográfico se utilizó el software gratuito *Zotero*.

El tratamiento de la base de datos a realizar en el segundo segmento del TFM se realizó desde el software R [12] con sus respectivos paquetes, los cuales serán mencionados en el escrito y/o código. Para la elaboración del escrito se trabajó con el formato *LaTex* a través de la herramienta *Overleaf*. Por último, el calendario de actividades se realizó en la página web *TeamGantt*.

1.4. Planificación

1.4.1. Tareas

1. Contextualizar el efecto de los datos faltantes en una base de datos longitudinal, así como abordar de manera teórica los diferentes métodos que han tenido mayor impacto en la literatura.

    a) Revisar la bibliografía acerca de los datos faltantes: patrones, mecanismos y métodos.

1) Definición de conceptos, patrones y mecanismos de datos faltantes.

2) Métodos de tratamiento: ventajas y limitaciones.

3) Búsqueda de paquetes de R disponibles para el tratamiento de datos faltantes.

4) Búsqueda de formas de comparación de resultados.

2. Poner en práctica métodos modernos de Imputación Múltiple, enfocados en tratar datos faltantes presentes en una base de datos longitudinal a diferentes tamaños muestrales.

a) Buscar una base de datos longitudinal publicada.

1) Obtención de una base de datos longitudinal.

2) Evaluar y seleccionar las variables a utilizar en el análisis.

b) Generar las bases de datos a diferentes escenarios de acuerdo al tamaño muestral.

1) Análisis descriptivo de la base de referencia en R.

2) Generar escenarios con bases de datos de tamaño pequeño: N/2 y N/3.

3) Analizar el tipo de mecanismo de datos faltantes presente.

4) Analizar las tasas de datos faltantes.

c) Poner a prueba los métodos escogidos sobre las bases de datos generadas.

1) Emplear métodos de Imputación Múltiple de acuerdo a las clases de variables presentes.

d) Analizar y comparar los resultados obtenidos con la base de datos original.

1) Análisis con pruebas de hipótesis de acuerdo al tipo y distribución de los datos.

1.4.2. Hitos

Hito 1: Definición de conceptos, patrones y mecanismos de datos faltantes

Hito 2: Métodos de tratamiento: bondades y limitaciones

Hito 3: Búsqueda de bases de datos: Obtención y análisis

Hito 4: Entrega PEC 2 (11/04/2022)

Hito 5: Generación de escenarios de missings por tamaño muestral

Hito 6: Simulación de análisis previos de missings

Hito 7: Empleo de métodos IM

Hito 8: Análisis de bases de datos imputadas

Hito 9: Comparación de resultados con la publicación de origen

Hito 10: Entrega PEC 3 (16/05/2022)

Hito 11: Elaboración de la memoria

Hito 12: Entrega PEC 4 (02/06/2022)

Hito 13: Elaboración de la presentación (30/05/2022 al 06/06/2022)

Hito 14: Defensa Pública (13/06/2022 al 23/06/2022)

1.4.3. Análisis de riesgo

Durante el desarrollo del trabajo final del máster hubo una serie de contratiempos que dificultaron llevar a cabo los objetivos que se tenían inicialmente previstos. Algunos contratiempos se habían considerado, como la compatibilidad con otras asignaturas del máster, así como la asistencia y presentación de proyectos del trabajo, en un conferencia fuera del país. Así mismo, dificultades informáticas debido a la incompatibilidad de algunos paquetes estadísticos con la versión del software R utilizado.

Por otro lado, hubo algunos contratiempos no considerados, más personales como temas de salud (COVID), o la obtención de un nuevo puesto de trabajo (el cual fue en un área nueva para mi persona, lo que requirió mayor dedicación). Con más relación al TFM, hubo muchos casos de incompatibilidad de ciertas funciones de R con las bases de datos utilizadas, ya sea por paquetes estadísticos desactualizados, o por la naturaleza de los datos.

Ante aquellos contratiempos que exigieron mayor dedicación por mi parte, la solución fue adecuar los objetivos del TFM, a fin de poder alcanzarlos en el tiempo estimado y sin perder la calidad del análisis y aprendizaje. Por ejemplo, inicialmente se pensaba emplear las técnicas de máxima verosimilitud, pero dado su complejidad y mi falta de conocimiento en la estadística multivariante, era imposible cumplir este objetivo en el tiempo establecido. En cambio, se

optó por probar diversas alternativas relacionadas con la imputación múltiple.

Por otro lado, la teoría de datos faltantes es un campo muy amplio y que esta en constante evolución. Es cierto que gran parte de la literatura consultada hacía uso de software estadísticos como SAS, SPSS o Matlab, dado a que son softwares que, por su trayectoria, cuentan con muchas funciones relacionadas con el tratamiento de *missings* en una variedad de escenarios. Sin embargo, no fue problemático conseguir alternativas que se podían emplear en el software R, por lo que la incompatibilidad de paquetes fue solo un inconveniente temporal, pero no una limitante en los análisis realizados.

1.4.4. Cronograma

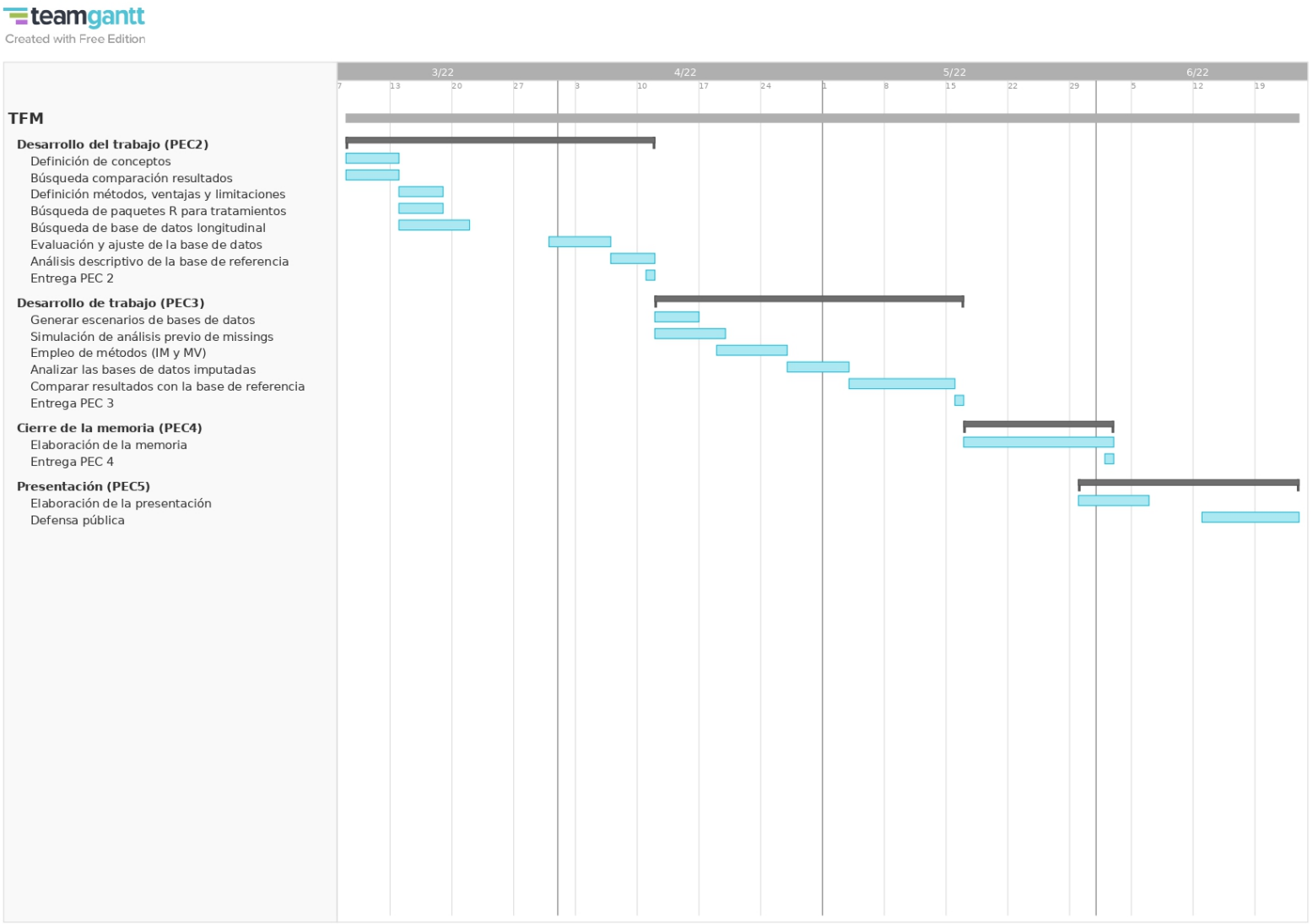

Figura 1.1: Cronograma

1.5. Breve resumen de las contribuciones

Estado del arte: Se abordan todos los conceptos teóricos que competen en este trabajo, describiendo brevemente que son los datos longitudinales y los datos faltantes. Así mismo, se introduce al lector en la teoría de los datos faltantes, explorando que son y los tipos de patrones y mecanismos que existen, así como los diferentes métodos de tratamiento existentes junto a sus bondades y limitaciones. Finalmente, se describe el origen de la base de datos con la que se trabajó.

Metodología: Capítulo enfocado en describir el flujo de trabajo realizado sobre la base de datos. Dicho apartado contiene cuatro bloques principales, i) creación de los escenarios de estudio; ii) Análisis descriptivo de los *missings*, iii) Tratamiento de los *missings*, y iv) Comparación de resultados. Cada bloque no solo describe la metodología utilizada sino que además incluye una breve explicación del porque se emplea la misma.

Resultados: Presentación de los resultados obtenidos tras poner en marcha el flujo de trabajo descrito en el capitulo Metodología. Nuevamente, este capitulo esta separado en cuatro bloques como en el capitulo anterior, a fin de mantener un orden. Además de presentar los resultados, se discute brevemente los mismos.

Discusión: Capítulo enfocado en abordar brevemente el flujo de trabajo realizado, los resultados obtenidos así como una discusión de los mismos. Debido a que mucha información fue discutida en el capitulo anterior, en la discusión se enfoca en describir las bondades y limitaciones de los métodos de imputación puestos en marcha, y como dichas características fueron observadas en los resultados finales.

Conclusiones: Conclusiones finales sobre el trabajo en cuestión, con dos apartados adicionales, i) Líneas a futuro, donde se describe otras metodologías a poner a prueba en futuros trabajos finales; y ii) Seguimiento de la planificación, donde se comentaron los contratiempos sufridos durante la elaboración del trabajo y los cambios realizados a fin de alcanzar el éxito en el mismo.

Capítulo 2

Estado del arte

Como bien se ha mencionado antes, la estructura del Trabajo Final de Máster consta de dos segmentos principales, uno teórico seguido por uno práctico. A continuación se presentará el segmento teórico, con el fin de abordar con mayor detalle aquellos conceptos previamente mencionados.

2.1. Datos longitudinales

Ningún estudio científico estaría completo sin un diseño de estudio de investigación, el cual permitirá de una forma sistemática y objetiva, responder interrogantes y avanzar en el conocimiento de distintos campos de investigación, incluyendo el campo de la salud. Sin embargo, no existe un único tipo de diseño, sino que va a depender del campo de estudio y de las interrogantes que tenga el o los investigadores. Debido a que es muy extenso dar a conocer todas las alternativas de diseño de estudios de investigación, los mismos son clasificados por Reina Ortiz y Sharma [13] en dos grandes grupos, diseños experimentales y observacionales.

Grosso modo, los diseños experimentales son aquellos en los que en principio hay una serie de variaciones controladas, como ensayos clínicos y comunitarios, mientras que en los diseños observacionales no hay variaciones controladas, como los estudios descriptivos y analíticos. A su vez, los estudios analíticos se clasifican de acuerdo a su temporalidad, ya sea porque las observaciones se realizan en un momento determinado (estudios transversales) o porque se realizan en el transcurso del tiempo (estudios longitudinales) [13, 14].

A través del tiempo, la definición de estudio longitudinal ha ido cambiando debido a que ha sido ampliamente utilizado por investigadores de diferentes campos [14, 15]. Actualmente, los estudios longitudinales son definidos como aquellos estudios enfocados en realizar mediciones de forma continua o repetitiva a un grupo de individuos en el tiempo [1, 14], por lo que son muy frecuentes en el campo de la medicina y la salud.

Los diseños experimentales pueden presentarse en diversos estudios observacionales (seguimiento del desarrollo infantil) e incluso como experimentales (evolución de un paciente ante un tratamiento), y los datos pueden ser de origen retrospectivo (obtenidos en el pasado) o prospectivo (a obtenerse en el presente) [15, 16]. Para que un diseño sea longitudinal, este debe estar compuesto de i) un seguimiento en el tiempo, ii) mediciones repetidas (más de dos) y iii) un conjunto de análisis que abarquen todas las mediciones obtenidas [14, 16].

2.2. Datos faltantes

Una de las desventajas principales del estudio longitudinal en el campo de la medicina y salud es su dependencia al recurso tiempo, ya que interviene en diferentes aspectos del estudio. Además de la duración del experimento *per se*, el tiempo puede influir en la colecta de información debido a que el objeto de estudio es el paciente. Por diversos motivos (mudanza, abandono del programa, fallecimiento) el seguimiento u observación es interrumpido [15, 16], lo que conlleva a la aparición de datos faltantes o *missings*.

Little y Rubin definen un dato faltante como un valor no observado que, en caso de que hubiese sido observado, presentaría un valor significativo para el análisis [1], en otras palabras, un dato faltante es información relevante para el análisis que debería haberse registrado pero que está ausente en la base de datos. Es importante recordar que un dato faltante es diferente a registrar cero observaciones de una variable. Tal como se ha descrito, los datos faltantes o *missings* pueden estar presentes en cualquier tipo de base de datos incluyendo los registros longitudinales.

Además del factor tiempo ya mencionado, las causas de los datos faltantes pueden atribuirse a diversos factores, muchos de ellos fuera del control del investigador, que pueden ser el comportamiento de un sujeto de estudio (paciente), fallo de los instrumentos de medición e inclusive el azar [2].

En la literatura se hace mención a un conjunto de buenas prácticas para abordar este problema, inclusive cuando se está planificando el estudio [2]. Sin embargo, independientemente del diseño metodológico empleado y todas las precauciones tomadas, es imposible evitar la presencia de *missings* en la base de datos. Llegado el momento del análisis, es importante prestar atención a los valores ausentes y tratarlos adecuadamente ya que los mismos pueden influir significativamente en el poder estadístico, la integridad del análisis, estimaciones sesgadas y resultados incorrectos [2, 3]; principalmente porque los algoritmos de los análisis de datos fueron originalmente diseñados para bases de datos sin considerar los datos faltantes [17]. Varios autores incluso indican que tasas entre el 5 % y 20 % son aceptables, por lo que tasas superiores son perjudiciales para la interpretación de los resultados si no son debidamente tratados [3, 18, 19].

A pesar de estas consecuencias, son muchas las publicaciones que pasan por alto la pérdida de datos, ignorándolas por completo y/o subestimando su impacto en el análisis [19, 20, 21].

2.2.1. Teoría de los datos faltantes

Debido al impacto que los datos faltantes pueden generar sobre la interpretación de los resultados, y por ende en el éxito del estudio, los mismos han sido ampliamente estudiados con el fin de desarrollar métodos adecuados para su tratamiento. Como resultado, se han podido definir patrones y mecanismos de origen de los datos faltantes [1, 17, 22].

Patrones de datos faltantes

En la literatura es frecuente encontrar dos tipos de datos faltantes [17], traducidos como:

- Unidad sin respuesta (*item nonresponse*)
- Segmento sin respuesta (*wave/unit nonresponse*)
    - Deserción (*attrition*)

Para ejemplificar estos términos, se considera un estudio para evaluar una nueva dieta en pacientes, donde el seguimiento se realiza con encuestas mensuales enviadas por correo electrónico. El investigador al revisar las respuestas de los pacientes, se puede encontrar con alguna pregunta sin contestar en un determinado mes (**unidad sin respuesta**) o sin recibir la encuesta algún mes (**segmento sin respuesta**). Por último, se puede prolongar el caso

del segmento sin respuesta, al no recibir más respuestas por parte del paciente, siendo una **deserción**. Cuando son muchos los datos faltantes de diferentes tipos en una base de datos se habla de patrones.

Un patrón de datos faltantes es definido como la configuración de los datos observados y faltantes en una base de datos. Little y Rubin [1] seguidos por Enders [23] describen seis patrones distintos, los cuales son:

1. General
2. Univariante
3. Monótono
4. Multivariante
5. Coincidencia o faltantes planeados (*File Matching / Planned missing*)
6. Análisis factorial o Variable latente (*Factor Analysis / Latent variable*)

Es frecuente la presencia de unidades sin respuesta en las variables de una base de datos distribuidos sin un patrón definido, por lo que se refiere a un patrón **general**. Tomando el ejemplo anterior, tenemos una base de datos creada a partir de las encuestas recibidas por internet y se observa que no todas las veces la encuesta fue completada. Si en cambio, la mayoría de los pacientes no contestan una pregunta específica, será una variable en particular la que no posea información, presentando un patrón de tipo **univariante**.

Si el investigador deja de recibir las encuestas de pacientes en un tiempo determinado, por deserción por ejemplo, la base de datos va a presentar un patrón **monótono**. Por otro lado, si el paciente, además de realizar la encuesta mensual debe asistir a una consulta médica pero no todos lo cumplen, se tiene un patrón **multivariante** ya que habrán variables específicas sin información.

Un patrón más complejo es el de **coincidencia o faltante planeado**. Si continuamos con el ejemplo utilizado hasta ahora, podemos decir que este tipo de patrón se observaría cuando existen pacientes que únicamente contestan las encuestas y pacientes que solo van a la visita médica. Sin embargo, puede llegar a ser un concepto más complejo. Otra manera de ejemplificar este patrón es pensar en el sistema que, para evitar que una encuesta sea muy larga para un paciente, se fragmenta en subconjuntos, los cuales serán contestados por diferentes pacientes.

Por lo tanto, habrá pacientes que únicamente contesten el primer subconjunto, otros el segundo subconjunto y así sucesivamente.

Por último, **variable latente** se refiere a aquel patrón en el que no se tiene ninguna información para una variable en particular, debido a que es una medida que no puede ser tomada por la encuesta o proviene de otra fuente. Para la variable latente es necesario utilizar técnicas de variable latente, tales como los modelos de ecuaciones estructurales.

Como se ha podido observar, los patrones de datos faltantes son utilizados para describir la localización de los *missings* en una base de datos, sin indagar en su origen o causa [23].

Mecanismos de datos faltantes

Rubin en su trabajo de 1976 explica que cada dato tiene una probabilidad de perderse y que la misma va a estar gobernada por un proceso llamado mecanismo de datos faltantes (*missings data mechanism* en inglés). A partir de esta definición categorizó tres tipos de mecanismos [22]:

- Datos perdidos completamente al azar (MCAR; *missings completely at random*).
- Datos perdidos al azar (MAR; *missings at random*).
- Datos perdidos no debidos al azar (MNAR; *missings not at random*).

Los datos faltantes son considerados **MCAR** cuando la probabilidad de perderse un dato es la misma para todas las variables, es decir, es un evento ocasionado completamente por el azar y no está definido por ningún dato, observado o perdido. Un ejemplo de MCAR sería la ausencia de datos porque los pacientes no pueden asistir a algunas citas que tengan programadas (ya sea por trabajo, por la lejanía del centro de salud, etc.). Entre los tres mecanismos, el MCAR es considerado muy estricto y poco probable a que se cumpla en la práctica, debido a que requiere que la ausencia no esté relacionada con las variables de estudio [3]. Por otro lado, debido a que es el único mecanismo que no depende de datos no observados, se puede validar su presencia en la base de datos mediante pruebas empíricas [4].

Los datos **MAR**, pueden llegar a ocasionar confusión debido a su nombre, ya que son aquellos en que la probabilidad de pérdida de un dato en una variable está definida por otras variables observadas en la base de datos y no meramente al azar. Por ejemplo, un paciente que sigue una dieta para adelgazar pero no es recurrente con el tratamiento debido a que no tiene los medios para pagarlo. La variable económica influye sobre la variable respuesta (peso).

Por último, los datos **MNAR** son aquellos en los que la probabilidad de perderse está definida por la misma variable. Volviendo al ejemplo anterior, un paciente que seguía una dieta para adelgazar abandona el tratamiento debido a que no observa resultados deseados en su peso. En este caso la razón del dato faltante es la misma variable, aunque mayormente es información que desconoce el investigador.

Es importante recalcar que la función de los mecanismos no es buscar las causas u orígenes de los datos faltantes, sino describir posibles relaciones entre las mediciones y la probabilidad de datos faltantes [23].

2.2.2. Estrategias para el tratamiento de datos faltantes

A través de los años, se han desarrollado una gran cantidad de métodos a emplear sobre los datos faltantes cuando es el momento de analizar la información obtenida, los cuales han sido puestos a prueba por diversos autores tanto en simulaciones como con datos reales, generando conclusiones en cuanto a sus requisitos, bondades y limitaciones [3, 4, 5]. Una de las ventajas de tratar con datos faltantes en estudios longitudinales, es poder usar medidas previas en la base de datos para obtener estimaciones de los datos desconocidos con una mayor precisión [19].

Aunque es de sumo interés mencionar cada tipo de tratamiento, en general pueden ser agrupados de acuerdo a como son tratados los datos faltantes [2, 5, 6]:

- Métodos de eliminación
- Métodos de imputación:
  - Métodos no estocásticos
  - Métodos estocásticos

Métodos de eliminación

En términos generales, los métodos de eliminación han sido los más utilizados por la comunidad científica por su fácil implementación, inclusive encontrándose entre las funciones básicas de los *softwares* convencionales. La principal desventaja de estos métodos, es que reducen el poder estadístico al disminuir el tamaño de la base de datos. Entre ellos tenemos los métodos de Análisis de datos completos (*Listwise deletion*) y Análisis de casos disponibles (*Pairwise deletion*).

El método **Listwise** consiste en eliminar todas las observaciones de la base de datos que presentan al menos un dato faltante, sin considerar la pérdida de información que genera sobre las demás variables. Este método solo se recomienda utilizar cuando el porcentaje de datos faltantes es menor al 5 % y en mecanismos MCAR [6]. Por otro lado, el método **Pairwise** se enfoca en una matriz de varianza-covarianza, permitiendo hacer uso de las observaciones presentes en la base de datos y solo descartando aquellas observaciones en los análisis que se incluya la variable con datos faltantes. Sin embargo, esto genera diversos análisis de submuestras con diferentes tamaños muestrales, lo que puede llegar a producir problemas con las medidas de asociación entre las variables [17, 23].

Métodos de imputación no estocásticos

Los métodos de imputación buscan una alternativa más conservadora, evitando eliminar información de la base de datos, a cambio de imputar valores para suplantar los datos faltantes. El uso de los métodos no estocásticos (también llamados métodos de imputación simple) ha sido similar a los métodos de eliminación, aunque pecan por introducir sesgo estadístico, así como pérdida de variabilidad de los datos. Otra desventaja es que asumen que los mecanismos de datos faltantes son de tipo MCAR o MAR [2], y que se deben considerar problemas potenciales con errores tipo I y tipo II [5]. A continuación se presentan los más discutidos en la literatura.

El método de la **Observación previa** (LVCF, por sus siglas en inglés) consiste en reemplazar el valor faltante utilizando el último valor registrado. La desventaja de este método, además de lo ya mencionado, es asumir que entre dos o más tiempos determinados no hubo ningún cambio significativo [3, 5].

La **Imputación o reemplazo por medias** es una opción comúnmente utilizada entre la comunidad científica por su fácil empleo. La sustitución del dato faltante va a ser realizada calculando la media de los datos observados de dicha variable [5]. Una limitante de este método es el sesgo que produce tras la reducción de la varianza y la distorsión de las correlaciones [23].

La **Imputación por regresión** es similar al método anterior, con la diferencia que los valores que van a suplantar al dato faltante son obtenidos a partir de métodos tradicionales de regresión. Aunque este método logra disminuir el sesgo, debido a que permite añadir covariables al modelo, no consigue evitar la reducción de variabilidad estadística [5].

Métodos de imputación estocásticos

Por último, se encuentran los llamados métodos modernos o métodos estocásticos, basados en generar múltiples bases de datos a partir de los valores observados, reduciendo sesgos estadísticos potenciales y maximizando la variabilidad [5]. Entre sus ventajas está la flexibilidad que presentan sus supuestos, en comparación a los métodos previamente mencionados [4]. Son métodos menos utilizados debido a su complejidad, aunque cada vez es más frecuente encontrarlos como paquetes adicionales de software estadísticos recurrentes [4, 19, 21]. Entre ellos se encuentran los métodos de Máxima Verosimilitud e Imputación Múltiple, los cuales proveen resultados no sesgados cuando se emplean sobre mecanismos MCAR o MAR y el menor sesgo posible ante mecanismos MNAR, en comparación de los métodos tradicionales [4, 23].

El método de **Imputación por regresión estocástica** es muy similar a su homólogo no estocástico, con la diferencia de que se añade ruido a la predicciones que sigue una distribución normal, evitando la disminución de variabilidad y el sesgo [6, 23].

Las estimaciones de **Máxima Verosimilitud** (ML, por sus siglas en inglés) son métodos que subyacen bajo un modelo probabilístico. Consisten en generar predicciones a partir de parámetros estimados con los datos observados a partir de la función de Máxima Verosimilitud, sin requerir de imputaciones previas. Dentro de las estimaciones se encuentra el algoritmo Expectación-Maximización (EM), el cual es un algoritmo iterativo diseñado para la obtención de estimadores máximo-verosímiles en bases de datos con muestras incompletas. Cada iteración pasa a través de un paso expectación (E) y un paso maximización (M), los cuales consisten en imputar valores, mediante ecuaciones de regresión, seguido por la obtención de la media aritmética y la matriz de covarianza con la base de datos completa obtenida. El proceso se repite hasta que ocurre la agrupación de todas las estimaciones [23, 24].

Por último, existe una alternativa a las estimaciones de Máxima Verosimilitud llamada **Información Completa de Máxima Verosimilitud** (FIML, por sus siglas en inglés). Este método estima el logaritmo de verosimilitud de cada observación basado en las variables presentes en el modelo, estimando como debería ser el modelo hipotético en un solo paso [17]. Sin embargo, esta alternativa se basa en modelos multivariantes como lo son los modelos de ecuaciones estructurales (SEM, por sus siglas en inglés), por lo que es necesario tamaños muestrales medianos o grandes para un buen desempeño [25].

El método de **Imputación Múltiple** (MI, por sus siglas en inglés) se basa en estimar

posibles opciones para los valores perdidos. Para ello el investigador comienza escogiendo el número de opciones (m) que desea imputar. Al iniciar el método, se generan "m"bases de datos alternativos y, finalmente, los mismos son usados juntos para llevar a cabo análisis estadísticos que generen un único resultado o resumen [1]. Existen diversas formas de realizar la imputación, siendo algunas estrategias el Mínimo cuadrado bayesiano, Emparejamiento predictivo de medias y Residuo aleatorio local (BLS, PMM y LRR, por sus siglas en inglés) [7].

Entre estas estrategias, *PMM* ha sido muy popular debido a que es un procedimiento *hot deck*, es decir, consiste en imputar valores provenientes de donantes o valores observados seleccionados, dentro de una misma variable. De allí viene su versatilidad de poder trabajar con distintos tipos de datos. Sin embargo, cuando se trata de bases de datos pequeñas, se debe tener sumo cuidado en su implementación debido a que, a medida que aumenta el porcentaje de datos faltantes, disminuye el número de donantes adecuados, trayendo como consecuencia la introducción de sesgo [9]. Recientemente, se ha desarrollado un procedimiento llamado *Midastouch* que evita la aparición del sesgo, ya que busca seleccionar donantes basándose en la probabilidad inversamente proporcional a la distancia de la observación con respecto al valor faltante. De esta manera, cualquier observación puede actuar como donante, mejorando la varianza presente [26].

A pesar de todas las alternativas para el tratamiento correcto de los datos faltantes, no ha sido posible encontrar un método único o estandarizado [2], principalmente porque no todas las bases de datos gozan de las mismas características, difiriendo en cuanto a; i) la distribución y tipo de variables así como la covarianza presente entre los mismos; ii) el mecanismo que ocasiona los *missings*; iii) el tamaño de la base de datos; e incluso iv) la tasa de datos perdidos. Además, la mayoría de los métodos (incluso los más recomendados) asumen que los datos faltantes son de tipo MAR o MCAR [2, 5, 18], por lo que si son empleados sobre datos MNAR puede llevar a resultados sesgados [6]. Existen métodos alternativos basados en modelos para los mecanismos MNAR, aunque hasta hace un tiempo no eran adecuados para un uso generalizado [2, 4, 23].

Como se ha mencionado al inicio de este apartado, los métodos expuestos descritos han sido los más utilizados y/o recomendados en la literatura. Sin embargo, existen otros métodos de imputación simple así como métodos basados en modelos (como GEE o Pattern-Mixture) que no son abordados en este trabajo. Por otro lado, es importante mencionar que las bondades de los métodos expuestos han sido principalmente definidas tras su puesta a prueba en grandes bases de datos, reales o simuladas [18, 27], por lo que se debe prestar atención a los resultados

cuando el tratamiento se realiza sobre bases de datos con pocas observaciones [7].

2.2.3. Impacto en bases de datos pequeñas

Como bien es sabido, la literatura siempre ha recomendado que un proyecto de investigación busque obtener el mayor número de observaciones posibles, para tener una base de datos grande que represente la mejor variabilidad posible de una población. Sin embargo, en la práctica muchas veces no es posible alcanzar dicha meta por diversas razones, por lo que el análisis de datos debe realizarse con una base de datos pequeña, es decir, que no posee un gran número de observaciones. No obstante, este concepto es muy ambiguo y va a depender del campo de investigación.

Entre las desventajas de una base de datos de esta índole está la falta de representación de la población, así como la pérdida del poder estadístico, limitando muchas veces al investigador al uso de análisis menos robustos por su flexibilidad ante los supuestos. Si a esto se le añade la presencia de *missings*, la interpretación puede ser inconclusa o muy diferente a la realidad. Por lo tanto, ante esta situación el investigador debe ser precavido y emplear los tratamientos que sean adecuados, acorde a si se cumplen o no los supuestos del método [8]. Nuevamente, la literatura recomienda el uso de métodos de Imputación Múltiple para el tratamiento de datos faltantes en este tipo de bases de datos [6, 27].

2.3. Descripción de la base de datos

El VIH (virus de la inmunodeficiencia humana) es un virus que ataca el sistema inmunitario del cuerpo y si no se trata, puede llegar a causar sida (síndrome de inmunodeficiencia adquirida). A pesar de que la erradicación de la infección es por los momentos un objetivo inalcanzable, están disponibles diversos tratamientos que han sido desarrollados con el paso del tiempo a fin de suprimir el virus en pacientes infectados. Sin embargo, no todos los pacientes consiguen una correcta adherencia a los tratamientos antirretrovirales, por lo que es imperativo desarrollar nuevas alternativas que hagan frente al virus.

La base de datos con la que se va a trabajar en este proyecto es el resultado de un ensayo clínico multicéntrico de 48 semanas realizado a 116 pacientes de VIH sin experiencia previa con tratamientos antirretrovirales. El objetivo del análisis fue poner a prueba un fármaco experimental y para ello, los pacientes fueron distribuidos al azar en grupos equitativos a los que

se les recetaba una combinación antiviral que incluía Efavirenz (tratamiento) o Lopinavir/r (control). La colecta de datos se realizaba mediante encuestas y visitas presenciales cada doce semanas, por lo que el resultado es una base de datos compuesta por variables demográficas como los datos personales del paciente y temporales, valores resultantes de pruebas hematológicas y bioquímicas.

Principalmente, los motivos de los datos faltantes descritos en la publicación fueron causados por reacciones adversas de distintos grados o de hipersensibilidad al tratamiento, seguido por la suspensión o abandono voluntario del paciente [28].

Capítulo 3

Metodología

Tal como se ha mencionado en apartados anteriores, toda la metodología se realizó mediante el software R versión 4.1.2 [12]. En general, los paquetes de R utilizados para gráficos enfocados en los datos faltantes fueron *naniar* y *VIM* [29, 30]. Los paquetes relevantes para el análisis de datos faltantes son mencionados en las diferentes secciones que componen este apartado, mientras que el resto pueden encontrarse al inicio del código R, adjunto al final del documento. La base de datos utilizada fue proporcionada por la tutora Dra. Núria Pérez Álvarez, coautora de la publicación resultante del estudio de investigación donde se obtuvieron los mismos [28].

3.1. Creación de los escenarios de estudio

A partir de la base de datos proporcionada, se procedió a seleccionar únicamente las variables con relevancia en el seguimiento de pacientes con VIH, incluyendo aquellas variables medidas en el tiempo (0, 12, 24, 36 y 48 semanas). De esta manera, se va a contar con una base de datos con más número de observaciones que variables, evitando problemas de multicolinealidad y alteraciones en los grados de libertad durante la imputación [27, 31]. Debido a que el rango de valores de la carga viral puede llegar a ser muy amplio, se procedió a utilizar una transformación de logaritmo en base 10, mediante $log_{10}(x + 1)$, para evitar inconvenientes con los ceros reales. El cuadro 3.1 contiene una breve descripción de las variables seleccionadas.

Entre las variables temporales escogidas están el *CD4A* y *CargaViral*, las cuales son de suma importancia en los estudios de VIH debido a que indican el estado inmunológico del paciente y la carga del virus de VIH en el paciente, respectivamente. Un paciente en mejores

Cuadro 3.1: Variables seleccionadas para las bases de datos

Variable	Tipo	Descripción
Grupo	Categórica	Binaria (EFV / LVP/r)
Sexo	Categórica	Binaria (Masculino / Femenino)
Edad	Numérica	Por año
tpo_vih_meses	Numérica	Tiempo desde el contagio (mes)
factor_riesgo_total	Categórica	Categorías de factor de riesgo (4)
DC4A_	Numérica	Temporal (0,12,24,36,48 semanas)
CargaViral_	Numérica	Temporal (0,12,24,36,48 semanas)

condiciones será aquel que presente, entre otras condiciones, valores altos de CD4A y una carga viral mínima o indetectable, signo de que el tratamiento antirretroviral está teniendo buenos resultados. Por lo tanto, el análisis estuvo enfocado principalmente en estas dos variables.

Los análisis exploratorios fueron enfocados en describir brevemente la estructura de la base de datos, acompañados de gráficos *boxplot* y *QQplot* para determinar la distribución de las variables.

Una vez realizados los análisis exploratorios, se procedió a generar dos bases de datos adicionales, en diferentes escenarios de acuerdo al tamaño N/2 y N/3, siendo N el tamaño muestral original (116 pacientes) de la base de datos en el paso anterior. La obtención fue realizada al azar, utilizando una semilla para que los resultados sean reproducibles, y enfocado en mantener proporciones iguales de pacientes en cada grupo (tratamiento y control).

3.2. Análisis descriptivo de los *missings*

El análisis descriptivo de los datos faltantes se basó en comparar las tres bases de datos generadas en la sección anterior. Adicionalmente, se estudió el patrón de *missings* mediante gráficos, utilizando el paquete *VIM*. De forma cuantitativa, se puede explorar el patrón mediante la comparación del valor *influx* o de flujo de entrada con respecto al *outflux* o flujo de salida. El valor *influx* refleja cómo se conectan los datos faltantes de una variable con los datos observados de otras variables, mientras que el valor *outflux* expresa la situación contraria. Ambos valores están representados en un rango entre 0 y 1 [31]. El resultado puede apreciarse en un gráfico *Fluxplot*.

Por último, se procedió a analizar si el mecanismo de datos faltantes presente en las bases de datos es de tipo MCAR. Debido a que el Test de Little requiere que los datos posean una distribución normal [11], fue necesario utilizar un test alternativo no paramétrico (Test de Jamshidian, Jalal y Jansen) que funciona sin importar el patrón de *missing*, mediante el paquete *MissMech* [32]. Para ello, fue necesario descartar las variables categóricas.

3.3. Tratamiento de *missings*

Para realizar la imputación, se generó una función que permitiera ordenar de izquierda a derecha las variables de acuerdo a su registro temporal, debido a que la imputación se realiza en esta dirección. De esta manera, se establece que la variable X_0 es imputada antes que la variable X_12, y así hasta llegar a la última semana. La imputación fue realizada siguiendo las recomendaciones descritas por Van Buuren y colaboradores, como se describe a continuación, utilizando las funciones del paquete *mice* [27, 31].

Tal como se mencionó anteriormente, existen diversas técnicas de imputación dentro del método FCS, especializadas de acuerdo al tipo de variable. La imputación de los datos fue realizada enfocada en la técnica univariante *Predictive mean matching* (PMM) debido a que ha demostrado buenos resultados en bases de datos con variables categóricas y/o numéricas, con porcentajes de pérdidas hasta del 50 % y que no presentan distribución normal [27, 33], incluso cuando poseen pocas observaciones [6, 7, 9].

Para ello, se exploraron tres alternativas otorgadas por el paquete *mice*; la primera sin definir el método de imputación, permitiendo que la función utilice las técnicas por defecto para cada tipo de variable (*PMM* para variables numéricas, *logreg* para variables binarias y *polyreg* para variables categóricas > 2 niveles) [31]. La segunda alternativa se realizó definiendo el método *PMM* como único método sin discriminar el tipo de variable. La tercera alternativa, similar a la primera, con la diferencia en que se utiliza la técnica *midastouch* en vez de *PMM* sobre las variables numéricas. Para hacer uso del *midastouch* junto a la función *mice*, fue necesario instalar el paquete complemento *midastouch* [26].

Por otro lado, cuando se trata de bases de datos con pocas muestras, la importancia por el número de imputaciones aumenta debido a que tiene un efecto mayor en la precisión de las estimaciones y los errores estándar. Por lo tanto, para obtener estimaciones de parámetros imparciales, fue necesario disminuir el número de imputaciones a medida que se trabajaba

con una base de datos con menor número de muestras, a un mínimo de m = 3, con el fin de evitar errores estándar subestimados [9]. Adicionalmente, la matriz predictora (*predictor matrix*) de la base de datos más pequeña (N/3) tuvo que se adaptada para evitar la aparición de eventos inesperados o *logged events*. Un *logged event* es una advertencia indicando problemas con los datos (como por ejemplo colinearidad entre variables), la cual requiere de la atención del investigador. Por otro lado, la matriz predictora es la que establece la función de predicción de una variable con respecto a las demás. Por defecto, la función *mice* establece que cada variable es predictora de todas las demás [27].

Finalmente, los resultados del número de imputaciones fueron comparadas mediante gráficos de densidad o *densityplot* y *stripplot*, del paquete *mice*. La idea es explorar de forma visual si los datos imputados reflejan una distribución similar a los observados. Por ejemplo, el *stripplot* tiene como ventaja poder estudiar la distribución de los valores imputados de cada imputación m, sobre la nube de puntos observados de variables numéricas, cuando las bases de datos poseen pocos valores. Por otro lado, los *densityplot* consisten en presentar la densidad de kernel de los datos observados e imputados (a diferentes m), a fin de observar si las curvas se asemejan [27]. Este tipo de gráficos es de suma importancia cuando se busca el valor m adecuado para los datos con los que se están trabajando, debido a que el objetivo es la similitud entre curvas.

Una vez hecho esto, los valores imputados fueron añadidos a sus respectivas bases de datos, generando un total de nueve bases de datos (3 bases de datos con 3 alternativas de imputación).

3.4. Comparación de resultados

El primer paso en esta sección fue reproducir brevemente las comparaciones realizadas en la publicación que dio origen a la base de datos [28], comparando los grupos de interés, es decir, entre el tratamiento (Efavirenz) y el control (Lopinavir/r) para todas las bases de datos obtenidas, mediante las variables CD4A y Carga Viral, ambas de la semana 48. El estudio buscaba evaluar la evolución de los pacientes diagnosticados con VIH durante 48 semanas con un nuevo tratamiento antirretroviral. Por lo tanto, uno de los análisis principales era evaluar estas dos variables al final del seguimiento y si los resultados eran iguales o no entre tratamientos.

Dado que las variables presentaban una distribución no normal, se procedió directamente en emplear la prueba de U de Mann-Whitney (*Wilcoxon-Mann-Whitney Test*) para dos muestras independientes mediante la función *wilcox_test* del paquete *coin*, la cual permite calcular el p

valor exacto sin la influencia de valores de empate (tie) [34]. Para ello, se creó una función que permitiera extraer directamente los valores p y Z obtenidos de cada comparación.

Capítulo 4

Resultados

4.1. Creación de los escenarios de estudio

La base de datos original consta de 116 observaciones/pacientes y 219 variables, las cuales 197 eran variables temporales, es decir, mediciones analíticas realizadas a los pacientes en diferentes semanas. Una vez realizada la selección de variables, se obtuvo una base de datos, denominada *VIH*, con 15 variables, tres de ellas categóricas y el resto numéricas, la mayoría de estas últimas con diversos datos faltantes (figura 4.1).

Como se ha descrito antes, la proporción de pacientes, en cuanto al número y al sexo, entre los dos grupos de tratamiento (EFV y LVP/r) era igual, con 58 pacientes cada uno. Sin embargo, la proporción de pacientes masculinos fue mayor (86.36 %). La edad media del paciente era de 37.71 años, y el factor de riesgo mas frecuente fue el homosexual (43.12 %).

Los valores observados de la variable CD4 son bajos en el tiempo 0, en comparación al resto que sí presentan similitud entre ellos, lo cual es señal de un aumento de la respuesta del sistema inmunológico de los pacientes que están siendo tratados. Por otro lado, una situación similar se observa al comparar las cargas virales en el tiempo, las cuales presentan valores mayores al inicio del tratamiento y disminuyen en el tiempo (figura 4.2). Es importante recalcar que uno de los datos del CD4A_12 presentaba un *outlier* incorrecto, con un valor que no correspondía al compararse con otras variables de la base de datos original. Para solucionar esto fue considerado como un valor faltante.

De acuerdo a la forma de los *boxplots* y debido a que la base de datos posee un número

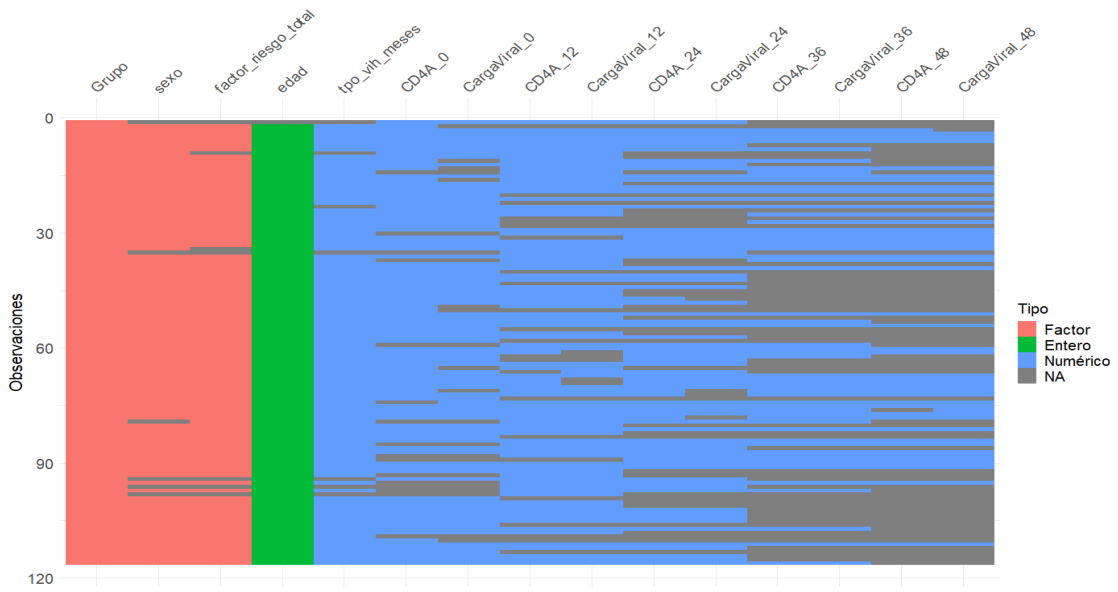

Figura 4.1: Estructura de la base de datos *VIH*

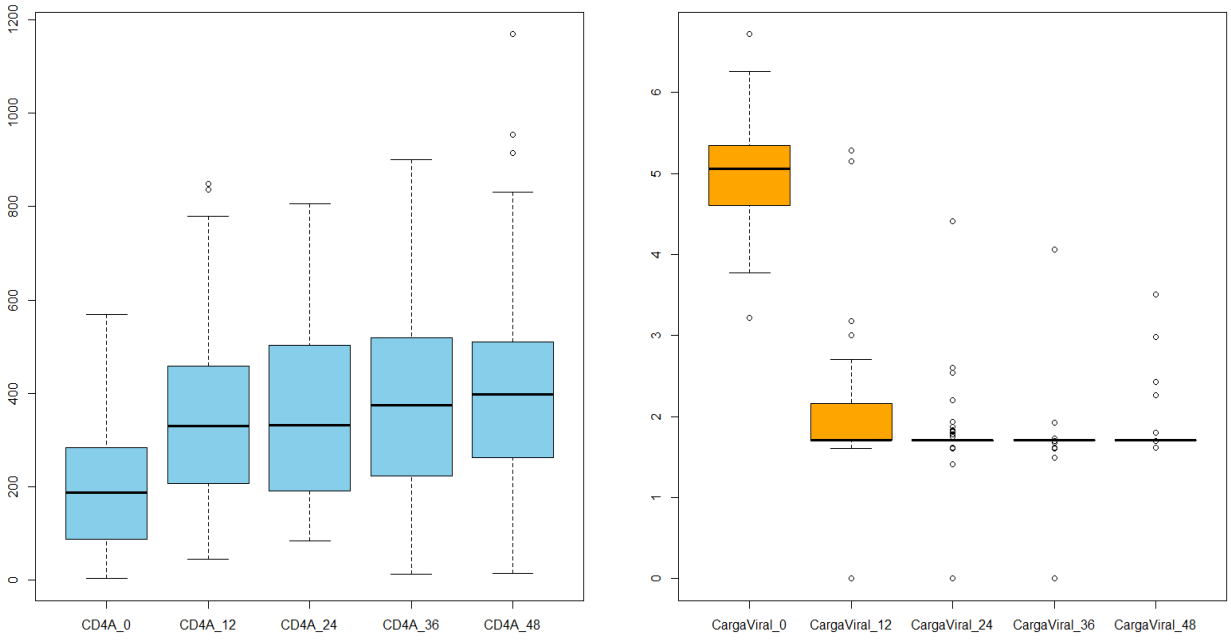

Figura 4.2: *Boxplot* Valor absoluto de los CD4 (Izq.) y Log_{10} carga viral de ARN del VIH, log_{10}(copias/mL) (Der.) en el tiempo.

de observaciones bajo acompañados de datos faltantes, se esperaría que no todas las variables presenten una distribución normal. Al realizar los gráficos *QQplot*, se confirmó visualmente que las variables de Carga viral no presentan una distribución normal (ver anexo A.1).

Por último, al proceder en la creación de los escenarios con 58 observaciones para la base de datos N/2 y 38 observaciones para la base de datos N/3 (en adelante VIH2 y VIH3), se mantuvo la misma proporción de pacientes entre grupos y los porcentajes de datos faltantes se mantuvieron similares, como se aprecia en el cuadro 4.1.

Cuadro 4.1: Características de las bases de datos (BD)

BD	**N° Observaciones**	**N° missings**	**Missings(%)**
VIH	116	452	26.0
VIH2	58	212	24.4
VIH3	38	147	25.8

4.2. Análisis descriptivo de los *missings*

El cuadro 4.2 representa de forma decreciente los porcentajes de datos faltantes de las variables de cada base de datos. Lo importante en este cuadro es observar la similitud de los porcentajes de pérdidas entre las bases de datos y, sobre todo, la tendencia a aumentar a medida que transcurre el tiempo, algo que puede detallarse en la figura 4.3 con las variables CD4A y carga viral. Como se ha mencionado en secciones anteriores, esto se debe a la dimisión de los pacientes en continuar el tratamiento, un factor de pérdida muy frecuente en bases de datos longitudinales [15, 16].

Cuadro 4.2: Porcentajes de *missings* de cada base de datos

Variable	VIH	VIH2	VIH3
CargaViral_48	60.34	58.62	63.16
CD4A_48	60.34	56.9	63.16
CargaViral_36	50.86	51.72	55.26
CD4A_36	50.86	51.72	55.26
CargaViral_24	37.93	36.21	36.84
CD4A_24	34.48	31.03	34.21

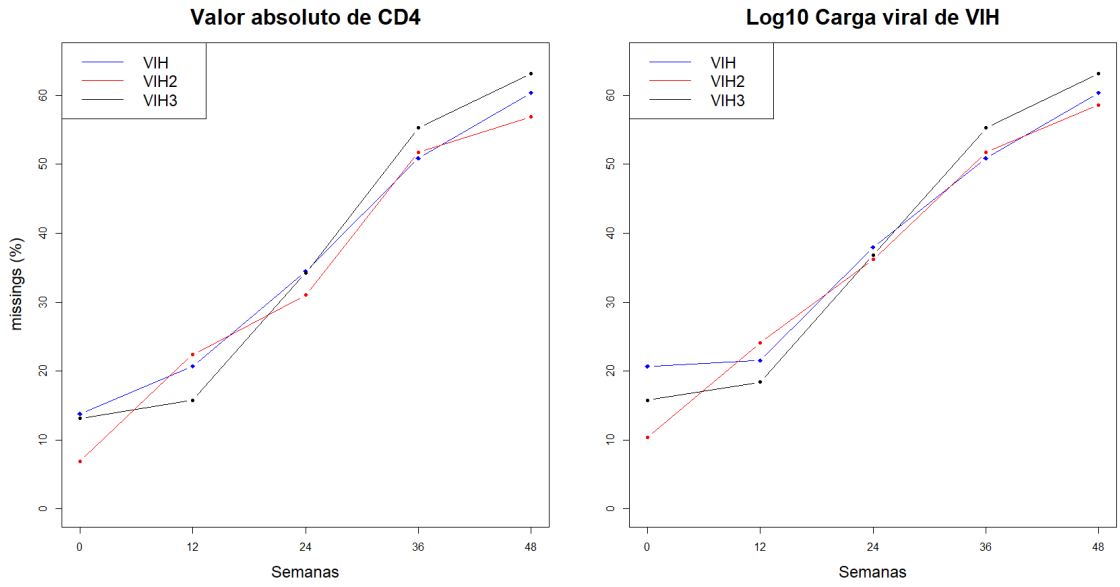

Figura 4.3: Aumento del porcentaje de pérdidas del valor absoluto de CD4 (izq.) y Log_{10}carga viral de ARN del VIH, log_{10}(copias/mL) (der.) en el tiempo.

La información reflejada en el cuadro 4.2 puede representarse gráficamente, tal como se observa en la figura 4.4 para la base de datos más grande (VIH). Dado que los gráficos son similares, los gráficos correspondiente a la base de datos VIH2 y VIH3 pueden encontrarse en los anexos A.2 y A.3.

El color rojo representa los datos faltantes y en azul los observados. En ella, de izquierda a derecha, se observa el volumen de datos faltantes de forma decreciente, comenzando por las variables de la semana 48. De arriba a abajo, representa el número de observaciones/pacientes que presentaron el mismo patrón de pérdida en el tiempo. Hay una columna y fila totalmente azul, es decir, sin ningún dato faltante. La columna representa la variable Grupo mientras que la fila representa el porcentaje de observaciones completas (20.7 % para VIH y VIH2, y 23.7 % para VIH3). Nuevamente, se puede apreciar que la mayoría de los pacientes siguieron su tratamiento abandonando únicamente en las últimas semanas. Al observar el patrón de distribución de los *missings*, el mismo puede ser identificado como mixto dado que se observa el patrón monótono esperado por la dimisión de pacientes, pero también las celdas rojas intercaladas o aleatorias reflejan un patrón general [23].

Al generar los gráficos *fluxplot* para cada base de datos, se puede observar una nube de

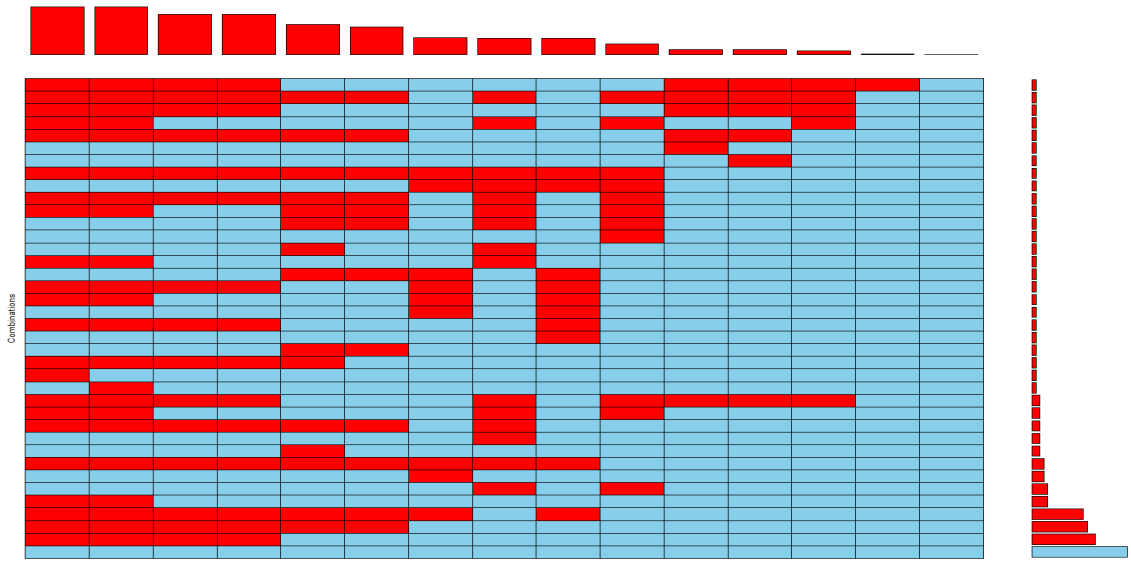

Figura 4.4: Patrón de *missings* por variables y observaciones de la base de datos VIH.

puntos que representan a cada variable. Por ejemplo, como se ha visto en la figura 4.4, la variable Grupo posee todas sus observaciones, por lo que va a tener un $outflux = 1$ e $influx = 0$. Por el contrario, una variable con pocas observaciones como CD4A y Carga viral de la semana 48, tienen valores $outflux = 0.03$ e $influx = 0.51$. Cuando todos los resultados se observan en un *fluxplot* (figura 4.5) se confirma que el patrón principal es monótono, debido a que a medida que el valor *influx* aumenta, el valor *outflux* disminuye, apreciándose la separación de las variables de la recta de referencia [31].

Por último, y dado que los datos no poseen una distribución normal, se procedió a realizar un test no paramétrico para evaluar si el mecanismo de datos faltantes es de tipo MCAR, obteniendo que no hay evidencia suficiente para rechazar la hipótesis nula con un nivel de significancia de 0.05, es decir, el mecanismo de los datos faltantes es de tipo MCAR para la base de datos VIH. Sin embargo, debido a las dimensiones de VIH2 y VIH3, la prueba generaba un error. Esto se arreglaba dejando únicamente las variables temporales, lo que implicaba pérdida de información. Otra alternativa fue utilizar el paquete *PKLMtest* [35], pero por posibles limitaciones computacionales, el mismo no mostró ni resultados ni errores en mi ordenador. Este tipo de prueba es muy básico ya que solo evalúa si el mecanismo es MCAR o no, pero no aporta ningún tipo de información sobre los otros dos mecanismos [4]. Cabe recordar que las técnicas de imputación múltiple generan buenos resultados ante mecanismos MAR y MCAR [4, 23]. Por

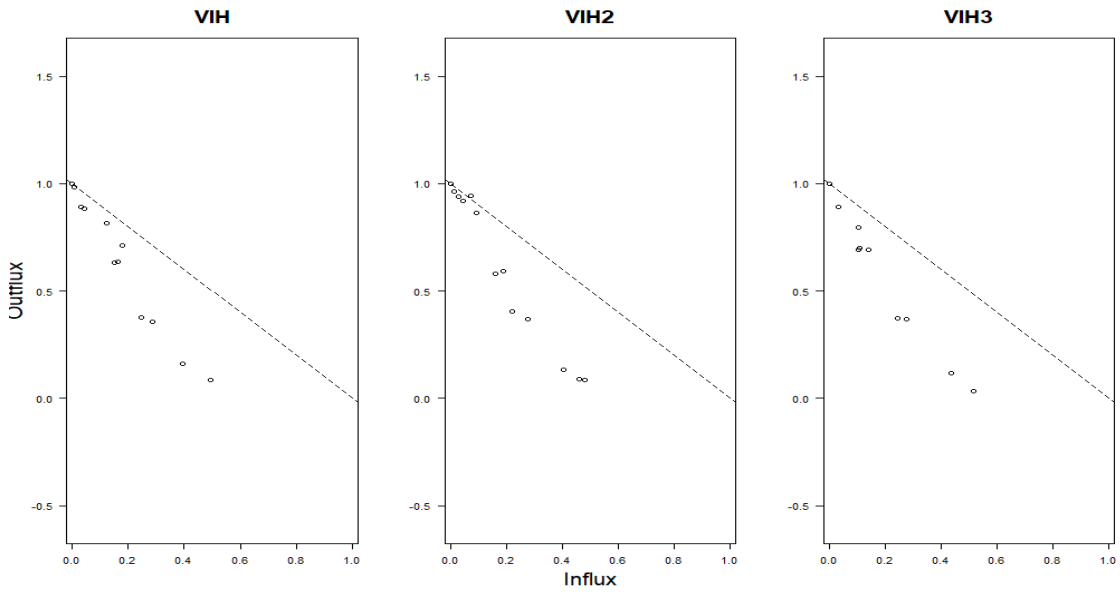

Figura 4.5: *Fluxplots* realizados para las tres bases de datos.

otro lado, conociendo las causas de como se generó la ausencia de los datos [28] y el patrón presente (figura 4.4), se descarta la posibilidad de que el mecanismo sea de tipo MNAR, y se procede con los tratamientos.

4.3. Tratamiento de *missings*

El primer paso en este punto fue ordenar los datos de forma temporal de izquierda a derecha, para que las imputaciones univariantes generadas mantuvieran un orden lógico. También es cierto que en caso de tener un patrón totalmente monótono, el mismo puede ser establecido en el código de la función *mice* utilizando el parámetro *visitSequence*. De esta manera, se organizan automáticamente las variables de acuerdo al porcentaje de datos faltantes [31].

Las imputaciones con mejores resultados fueron aquellas realizadas con un número de imputaciones m = 3, cada una con cinco iteraciones y sin importar el método utilizado. Cabe acotar que cuando se refiere a mejores resultados no es porque sea el mejor modelo; se refiere a la ausencia de eventos o *logged events*, y a la similitud entre las curvas obtenidas al crear los *Densityplots* tal como se describió en la metodología, y como se verá más adelante en esta sección.

Lo resaltante en este punto fue la necesidad de manipular la matriz predictora de la base de datos VIH3 debido al número de eventos *logged events* registrados en todas las alternativas al emplear por primera vez la función *mice*. La razón de esto, fue por el bajo número de observaciones con respecto al N total. Sin importar el número de eventos obtenidos, los mismos podían clasificarse en dos inconvenientes principales, colinearidad y datos insuficientes.

La colinearidad fue ocasionada por la variable Carga Viral de la semana 48. La función *mice* detecta la colinearidad y resuelve el problema removiendo a la variable del modelo, es decir, descarta las funciones de predicción y respuesta de la variable problema en la matriz predictora pasando todos los valores = 0 [27]. En respuesta, una vez obtenida la matriz predictora generada por la función *mice*, se seleccionaron únicamente aquellas variables que funcionan correctamente como predictoras de la variable Carga Viral 48, transformando nuevamente el valor de 0 a 1 (ver anexo A.4).

Por otro lado, la mayoría de los eventos ocurrían porque la matriz predictora por defecto asignaba como variable predictora al factor de riesgo sobre las variables temporales desde la semana 24 a la 48. Dado que el factor de riesgo es una variable categórica de cuatro niveles, las observaciones presentes en cada nivel no son suficientes para que el modelo actúe correctamente. Por lo tanto, se consideró descartar el factor de riesgo como un predictor de las variables temporales de las últimas semanas, modificando los valores correspondientes de la matriz predictora de 1 a 0 [27] (ver anexo A.4).

Una vez realizadas las respectivas imputaciones, se pueden observar las discrepancias entre las mismas con respecto a los datos observados, mediante un *stripplot* y un *densityplot*. Ambos tipos de gráficos presentados a continuación son basados en la alternativa *Midastouch*, mientras que los gráficos de las demás alternativas fueron colocados en la sección Anexos.

En primer lugar, los *stripplot* presentan la distribución de los valores imputados (color rojo), provenientes de cada imputación m, sobre la nube de valores observados (color azul). La primera nube (m = 0) representa únicamente los valores observados. En general se observa un claro solapamiento entre los dos conjuntos de datos, sin discriminar en cuanto al número de la imputación (figura 4.6). La misma observación se presenta ante los gráficos de las bases de datos VIH2 y VIH3, sin discriminar la alternativa utilizada (ver anexos A.5-A.12).

Por otro lado, con los *densityplot* se busca que entre ambos tipos de valores (imputados y observados) exista una concordancia entre sus densidades. Siguiendo el mismo patrón de colores

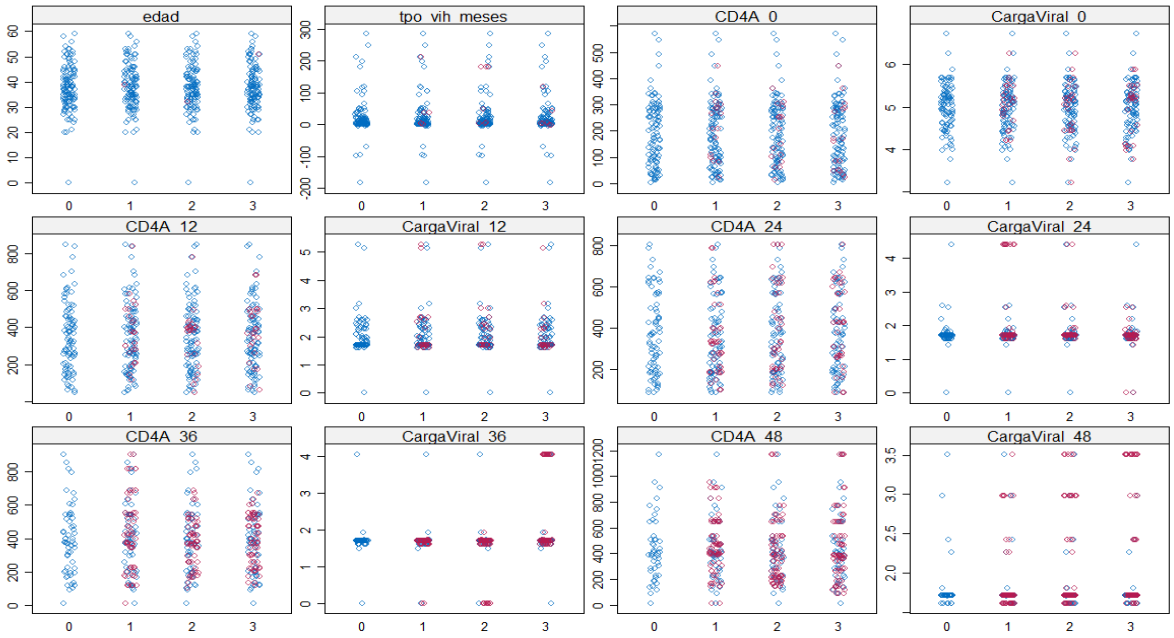

Figura 4.6: *Stripplot* de los valores imputados de la base de datos VIH con m = 3

que el gráfico anterior, se puede observar que en los gráficos de la base de datos VIH hay cierta similitud entre los datos, en unas variables más que otras (figura 4.7). Tal como se mencionó en la metodología, el uso de este tipo de gráfico fue el que permitió conseguir un valor m, acorde para cada base de datos. Es importante recordar que los gráficos de la variable carga viral se está representando la densidad de la transformación log_{10} realizada al inicio.

La diferencias en la forma de las curvas rojas con respecto a la azul, van a ser causadas por los valores observados o donantes escogidos por el modelo. En muchos casos va a estar presente una curva de imputación diferente al resto, dado a que el modelo toma en varias iteraciones al mismo donante, dando como resultado una curva con curtosis grande, tal como se observa en el *densityplot* de la variable CD4A de la semana 12. En otras situaciones, la carencia de diversidad de valores observables o donantes va a ser la causante de que las curvas se comporten diferente a la curva de datos observados, situación que se ve reflejada en variables con alto porcentaje de *missings* como es el caso de la variable Carga Viral 48.

Siguiendo con la idea anterior, entre las tres bases de datos, el porcentaje de datos faltantes es constante (cuadro 4.1), pero el número de observaciones disminuye, por lo que es probable que el sesgo aumente, en comparación al que se pueda encontrar en la base de datos VIH. Parte

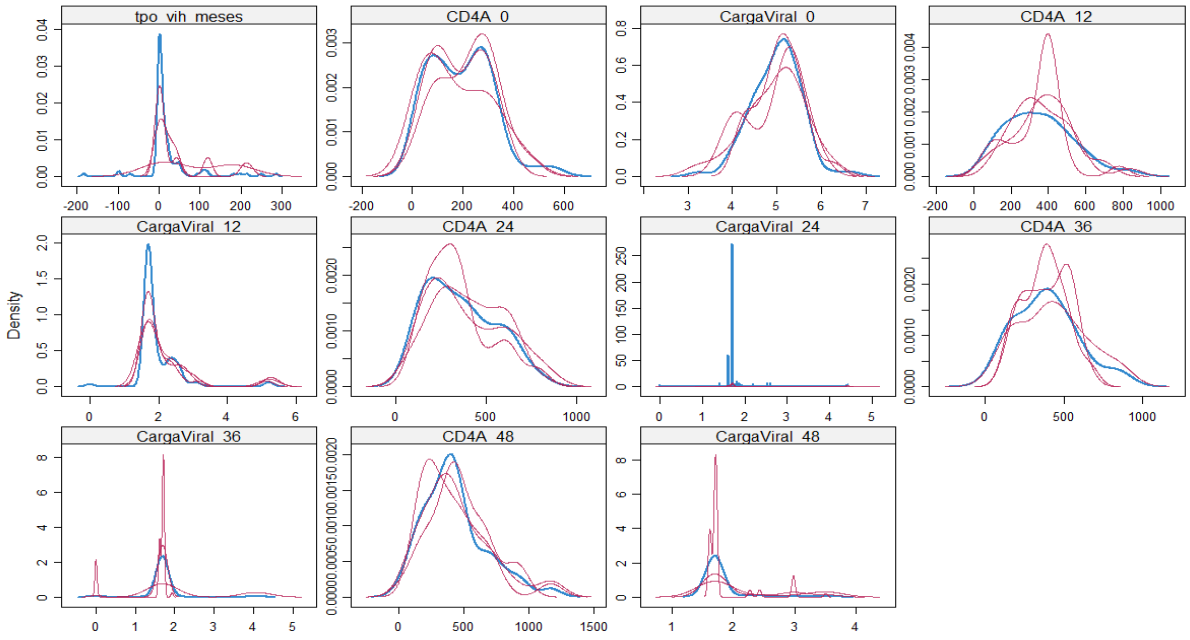

Figura 4.7: *Densityplot* de los valores imputados de la base de datos VIH con m = 3

de esto lo podemos observar al comparar los *densityplots* de las tres bases de datos. Las curvas de la base VIH se muestran muy similares entre sí, con respecto a la curva de valores observados (azul). Sin embargo, la discrepancia va aumentando a medida de que N muestral disminuye, tal como se observa en los gráficos de VIH2 (figura 4.8) y en los gráficos de VIH3 (figura 4.9), dificultando la búsqueda de parámetros adecuados que encaje bien con los datos. Esta situación no es diferente al comparar las bases de datos tratadas con las alternativas *PMM* (ver anexos A.13-A.18).

4.4. Comparación de resultados

Tal como se mencionó en la metodología, se procedió directamente a realizar las comparaciones de los grupos de tratamiento de todas las bases de datos utilizando la prueba de U de Mann-Whitney. En este punto fue necesario recurrir al uso de la función *wilcox_test* para evitar la influencia de los valores de empate o repetidos (tie) y obtener el p valor exacto.

En la tabla 4.3 se observan los valores p obtenidos en el paso anterior, a un nivel de significancia del 95 %. Cabe recordar que aquellos valores p mayores a 0.05 indican que no hay

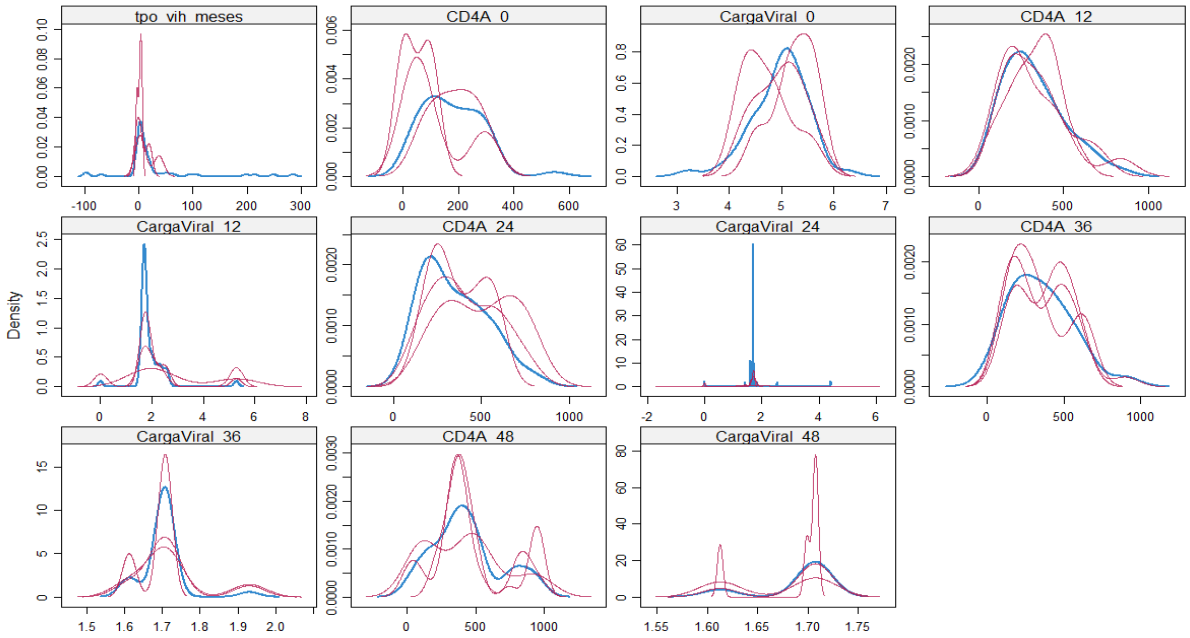

Figura 4.8: *Densityplot* de los valores imputados de la base de datos VIH2 con m = 3

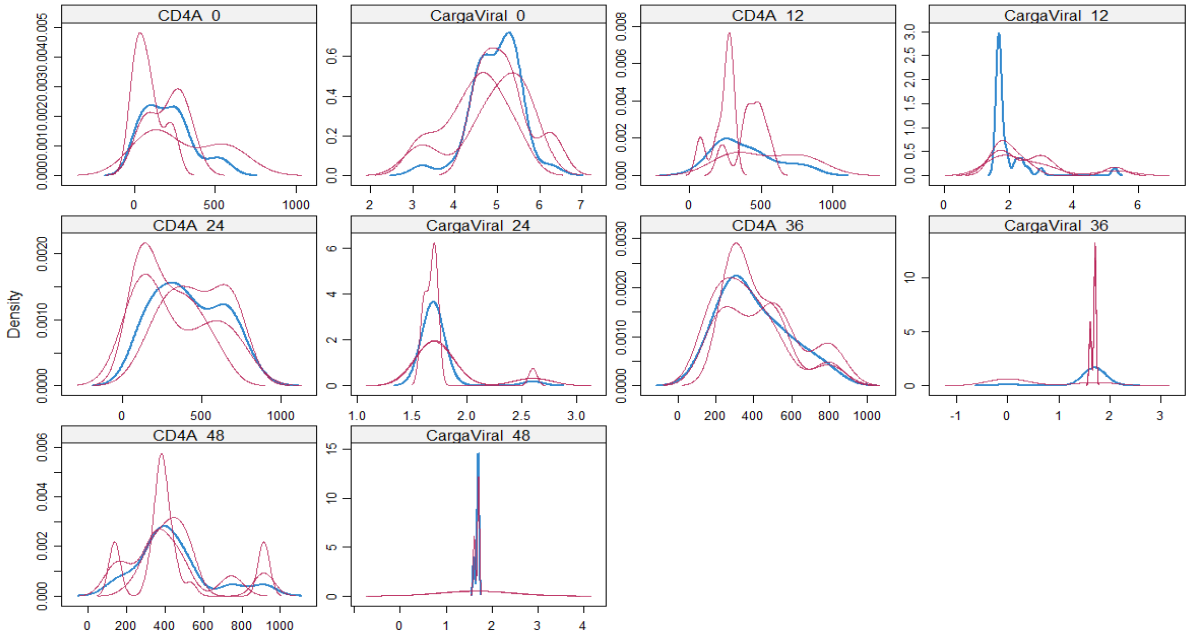

Figura 4.9: *Densityplot* de los valores imputados de la base de datos VIH3 con m = 3

Cuadro 4.3: Tabla de resultados obtenidos a un nivel de significancia del 95 % tras comparar los grupos de tratamiento y control de cada base de datos, para las variables CD4A y Carga viral de la semana 48. En negrita, aquellos p valores menores a 0.05.

		CD4A (semana 48)		Log_{10} Carga viral (semana 48)	
Base de datos	N	P valor	Z	P valor	Z
VIH_defecto	116	0.671	-0.428	0.476	0.701
VIH2_defecto	58	0.911	-0.117	0.895	-0.269
VIH3_defecto	38	**0**	3.461	**0.042**	-2.356
VIH_PMM	116	0.229	1.207	0.646	0.538
VIH2_PMM	58	0.923	0.101	0.079	-1.869
VIH3_PMM	38	**0.002**	3.078	1	0.413
VIH_Midas	116	0.879	0.155	0.262	-1.114
VIH2_Midas	58	0.354	0.936	1	0.066
VIH3_Midas	38	0.315	1.02	1	-0.594

evidencia suficiente para rechazar la hipótesis nula, es decir, no existen diferencias significativas entre los dos grupos comparados. Como se observa en la tabla, la mayoría de los resultados presentaron un p valor mayor a 0.05. La base de datos con menor N (VIH3) imputada únicamente por *PMM* presentó un p valor menor a 0.05 en la comparación de grupos con la variable CD4A. Se obtuvo el mismo resultado en la comparación de ambas variables de la base de datos VIH3 imputada por defecto. En cambio, VIH3 imputado con *Midastouch*, no presentó diferencias significativas.

La mayoría de los resultados obtenidos concuerdan con los descritos por la publicación de origen, es decir, no obtuvieron diferencias significativas entre los tratamientos para ambas variables. Sin embargo, es prudente enfatizar que los estudios realizados originalmente fueron más complejos que los aquí mostrados [28].

Capítulo 5

Discusión

Hasta este punto, se ha descrito brevemente el flujo de trabajo necesario para el tratamiento de datos faltantes en una base de datos longitudinal a diferentes escenarios en cuanto al tamaño de la muestra. En general el flujo se pudo dividir en varios pasos:

- Estudio de la base de datos y sus variables
- Evaluación del patrón de datos faltantes
- Evaluación del mecanismo
- Tratamiento

Comenzar el flujo de trabajo analizando la bases de datos es imprescindible debido a que es importante conocer con qué tipo de variables se está trabajando, así como el porcentaje de datos faltantes presentes en cada una. Lo más resaltante en este punto fue observar que la mayoría de variables eran numéricas, mayormente generadas en el tiempo al realizar mediciones del valor absoluto de CDA y la carga viral del VIH, lo que a su vez es un factor de pérdida de información. Es muy frecuente en bases de datos longitudinales que a medida que transcurre el tiempo, los pacientes suelen abandonar el seguimiento, ocasionando que el porcentaje de *missings* aumente, algo que se pudo corroborar en la figura 4.3. Como consecuencia era de esperarse que los datos faltantes presentaran un patrón principalmente monótono. Sin embargo, en la actualidad no es de suma importancia conocer el tipo de patrón debido a que los métodos modernos como la imputación múltiple o la estimación por máxima verosimilitud se adaptan muy bien a cualquier patrón [23].

Por otro lado, sí es importante conocer qué mecanismo de *missings* está presente en la base de datos, ya que la presencia de datos MNAR provoca sesgos en los tratamientos, aunque en menor escala en los métodos modernos [23, 27], por lo que se deben tomar ciertas consideraciones si se va a tratar este tipo de datos, o realizar tratamientos alternativos no abordados en la primera sección de este trabajo.

Una limitación encontrada para evaluar el mecanismo de *missings* fue conseguir un paquete del software R que permitiera realizar una prueba no paramétrica, ya que la mayoría de los paquetes están enfocados en desarrollar el test de Little, donde uno de sus supuestos es que los datos deben tener distribución normal [11]. Las alternativas no paramétricas encontradas presentaron errores posiblemente por limitaciones computacionales, aunque al momento de abordar este punto durante la elaboración del trabajo, se encontró que los mismos no están en constante actualización e incluso han sido eliminados del repositorio de paquetes CRAN. Dado a que solo se efectuó correctamente la prueba sobre una base de datos, se descartó tomar en cuenta la información obtenida, a pesar de que esta indicaba que el mecanismo era MCAR.

Sin embargo, tomando en cuenta este resultado preliminar y que las causas de los datos faltantes se conocían, se consideró que el mecanismo no era de tipo MNAR. Es importante recordar que los datos MNAR son aquellos en que la probabilidad de perderse está definida por la misma variable [22], lo cual no es el caso debido a que los datos faltantes ocurrían por la deserción del paciente, esto es, por factores ajenos a las variables presentes en la base de datos. Por lo tanto, se pudo proceder con los métodos de imputación múltiple convencionales, los cuales generan mejores resultados ante mecanismos de tipo MAR o MCAR [4, 23, 27].

Como último paso en el flujo de trabajo está realizar el tratamiento de *missings* tomando en cuenta principalmente la naturaleza de los datos y el mecanismo de datos faltantes.

Se realizaron tres alternativas de imputación múltiple, todas enfocadas en la estrategia del emparejamiento predictivo de medias o conocido por su nombre en inglés *Predictive mean matching* o *PMM*. Como se ha mencionado anteriormente, el *PMM* es la estrategia mayormente utilizada en la literatura, a tal nivel de que se encuentra por defecto en muchas funciones de imputación múltiple como lo es el caso de la función *mice*. Dado su versatilidad al momento de imputar los datos, es una de las pocas alternativas que genera buenos resultados ante bases de datos con pocas observaciones, sin distribución normal y sin discriminar el tipo de variable presente [27].

Cuando no se define la estrategia o método en el código de imputación, la función selecciona los métodos de acuerdo al tipo de variables. Por defecto, el *PMM* solo es aplicado sobre variables numéricas [27]. Tomando en cuenta que la imputación se realiza de izquierda a derecha, la primera alternativa, la cual fue denominada *por defecto*, comenzó imputando variables categóricas mediante otras estrategias (*logreg*, *polyreg*), hasta llegar a las variables continuas a las que se aplicó el *PMM*. La segunda alternativa fue aplicando el *PMM* a todas las variables y la tercera alternativa fue igual al primer caso pero sustituyendo a *PMM* por *Midastouch*.

Anteriormente se ha descrito a *Midastouch* como un método mejorado del *PMM*, enfocado en que la probabilidad de que un valor observado sea donante dependerá de la distancia entre el mismo y el dato faltante [26]. Originalmente se buscaba aplicar *Midastouch* a todas las variables, pero la función generaba error con las variables categóricas, a pesar de que en la teoría se indica que si puede ser empleado en este tipo de variables [9, 26]. Por cuestiones de tiempo, fue imposible solucionar el error.

Los resultados de los valores imputados fueron visualmente muy similares, lo cual tiene sentido dado a que se trabajó con los mismos parámetros y con base al método *PMM*. Era de esperarse que los *stripplot* presentaran las nubes de valores imputados dentro de la nube de valores observados, dado a la forma de trabajar de *PMM*, mediante la imputación de valores faltantes basado en los valores observados que actúan como donantes [27].

Por otro lado, las diferencias entre las curvas de densidad imputadas con respecto a la de los datos observados son causadas por el tamaño muestral de la base de datos. Ante el menor número de observaciones, en conjunto con el porcentaje de *missings*, va a ocasionar poca disponibilidad de candidatos adecuados como donantes de los valores a imputar. Como consecuencia, se introduce sesgo en los datos y por lo tanto las curvas de densidad no serán idénticas [9]. Por tal motivo, era más frecuente observar estas diferencias entre las distintas variables, al observar las bases de datos mas pequeñas, sobre todo VIH3 (n = 38).

Finalmente, al realizar un breve análisis estadístico, se obtuvieron resultados similares entre todas las bases de datos, a excepción de las bases de datos VIH3 de las dos primeras alternativas. Si se considera el resultado de la publicación como referencia [28], se puede decir que se obtuvieron los resultados esperados, y que efectivamente el método de imputación empleado (*PMM*) trabaja bien con datos pequeños, hasta cierto punto, e incluso con altos porcentajes de datos faltantes, como los que presentaban las variables comparadas ($> 55\,\%$). Sin embargo, es importante tener en cuenta ciertos aspectos.

El *PMM* es un método robusto que ha tenido bastante aceptación por la comunidad científica debido a su fácil implementación, así como sus estimaciones insesgadas, preservando la distribución original de los datos. Sin embargo, tal como se mencionó en el párrafo anterior, las estimaciones se ven afectadas cuando la cantidad de observaciones son insuficientes para obtener donantes adecuados, ya sea por altos porcentajes de datos faltantes (superior al 50 %), por un tamaño muestral bajo, o por ambas situaciones [18, 27]. Kleinke (2017) encontró que este método generó buenas estimaciones a través de una serie de escenarios, pero enfatizó la pérdida de precisión cuando se trabaja con datos muy sesgados y bases de datos pequeñas [9]. Un claro ejemplo de lo descrito se puede observar en los resultados de la tabla 4.3, ya que se obtuvieron resultados diferentes a los esperados cuando la base de datos poseía el N más pequeño.

Sin embargo, son muchos los casos que por diversos factores (económicos, logísticos, biológicos) es imposible obtener una base de datos con un gran número de observaciones. Si además, a estos factores se le añade la pérdida de información, el resultado serán bases de datos como las que se han puesto a prueba en este trabajo. Ante esta situación, son pocas las herramientas que se disponen para llevar a cabo los análisis del estudio. Por lo tanto, a pesar de que el método *PMM* tiene ciertas limitaciones, es una opción totalmente válida cuando las bases de datos presentan características tan limitantes como las descritas, algo que ha sido posible comprobar al comparar los resultados obtenidos con los presentados en la publicación de origen. Además, los paquetes como *mice* comprenden un conjunto de parámetros adicionales a los utilizados en este estudio, que su uso dependerá en conocer con que variables se esta trabajando, así como un conocimiento más profundo sobre este método [27], por lo que el tratamiento realizado sobre los datos no es el único.

Por último, como se ha ido mencionando a través de las diferentes secciones, la teoría de datos faltantes se encuentra en constante evolución, por lo que el desarrollo de técnicas para el tratamiento de los mismos es cada vez más fácil de implementar e interpretar, obteniendo estimaciones de mejor calidad. Claro es el ejemplo de la alternativa *Midastouch*, el cual en pocas palabras es una actualización del método *PMM*. Efectivamente, al hacer uso de esta alternativa, obtuvimos los resultados esperados sin importar el tamaño muestral.

Basado en los resultados descritos en este trabajo, se hace imperativo seguir explorando estas técnicas con bases de datos que posean características similares, a fin de contar con evidencia científica, más allá de las recurrentes simulaciones.

Capítulo 6

Conclusiones

El tratamiento de datos faltantes abarca una serie de alternativas que están en constante evolución, con la aparición de métodos cada vez más complejos, pero con estimaciones más precisas y menos sesgadas. Sin embargo, aún no existe una metodología estandarizada o única que sea aplicable a cualquier situación. Dado a que en el ámbito de la biología y la salud es frecuente encontrarse con bases de datos con tamaños muestrales pequeños, era necesario conocer y poner a prueba los métodos de imputación múltiple en una base de datos longitudinal real sobre diferentes escenarios, de acuerdo al número de observaciones presentes.

Uno de los aspectos más importantes es conocer los datos con los que se están trabajando, cómo fueron medidos, cuáles fueron los inconvenientes, así como posibles relaciones, a fin de tener una idea preliminar de qué esperar en cuanto al patrón, mecanismo de los *missings*, así como el tratamiento adecuado a emplear. Dado que la idea de este trabajo era explorar el tratamiento de datos faltantes, la metodología descrita intentó abarcar de una forma generalizada todos los pasos previos a realizar antes de llevar a cabo una imputación. Sin embargo, la misma ha de ser adaptada de acuerdo a la base de datos con la que se desee trabajar. El proceso de imputación utilizado generó los resultados esperados, por lo que concuerda con los resultados de simulaciones descritos en la literatura.

Finalmente, el aporte de este trabajo, además de generar un flujo de trabajo general, fue demostrar las capacidades de la estrategia de imputación múltiple *PMM* sobre altos porcentajes de datos faltantes en una base de datos real con pocas observaciones, características que suelen ser frecuentes en diversas investigaciones, pero que suelen tomarse en cuenta cuando se realizan

simulaciones para poner a prueba técnicas de esta índole.

6.1. Líneas de futuro

En las secciones anteriores se han puesto a prueba algunas de las diferentes técnicas de imputación múltiple encontradas en la literatura, que generan buenos resultados al trabajar con una base de datos sin distribución normal, con un tamaño muestral pequeño y con un porcentaje de datos faltantes considerable [9, 27]. Sin embargo, en la literatura también se hace mención a métodos de máxima verosimilitud que incluso otorgan mejores resultados bajo las condiciones de MCAR o MAR comparado con la imputación múltiple [23, 24, 36].

En líneas de investigación futuras es recomendable abordar las técnicas de máxima verosimilitud, como FIML y EM, de la misma forma como se han abordado las de imputación múltiple en este trabajo, teniendo en cuenta que se deben poseer conocimientos previos de máxima verosimilitud y estadística multivariante, específicamente en modelos de ecuaciones estructurales (SEM, por sus siglas en inglés) [23, 36]. Además, se debe tener en consideración que a pesar de que estas técnicas generan buenas estimaciones para cualquier base de datos, también generan errores estándar sesgados cuando los datos no son normales, por lo que se debe recurrir al empleo de correcciones como *bootstraping*, errores estándar robustos [23, 25] o alternativas como la técnica de máxima verosimilitud escalada (SML, por sus siglas en inglés) [36].

Por otro lado, se podría dar continuidad a los resultados encontrados en este trabajo, mediante la reproducción de análisis ya realizados con otra base de datos de las mismas características, e/o indagando en los parámetros aportados por el paquete *mice*.

6.2. Seguimiento de la planificación

El seguimiento de la planificación fue realizado de forma parcial. Tal como se comentó en la sección Análisis de riesgo, varios contratiempos estuvieron presentes durante la elaboración del trabajo. Inicialmente se quería poner a prueba no solo el método de imputación múltiple, sino también las estimaciones de máxima verosimilitud ya que ambos son métodos modernos muy mencionados en la teoría. Sin embargo, las estimaciones de máxima verosimilitud requieren una mayor comprensión de métodos multivariantes, así como el uso de softwares estadísticos

diferentes a R. Por cuestiones de tiempo, era imposible abordar toda esta información y llevarla a cabo con éxito en el trabajo. Por lo tanto, se decidió llevar a cabo diversas alternativas de empleo de la estrategia *PMM* de imputación múltiple. Por otro lado, una vez imputados los datos, se deseaba indagar en comparaciones más complejas que las realizadas, pero por los mismos contratiempos no fue posible.

Capítulo 7

Glosario

EM: *Expectation and maximization algorithm*: Método de Máxima verosimilitud.

FCS: *Fully Conditionally Specification*: Método especial de Imputación Múltiple.

FIML: *Full Information Maximum Likelihood*: Método especial de Máxima Verosimilitud.

GEE: *Generalized Estimating Equation*: Ecuaciones de Estimación Generalizada: Método de tratamiento de datos faltantes que involucra el uso de modelos.

LVCF o LOCF: *Last Value/Observation Carried Forward*: Imputación por última observación

MCAR: *Missing completely at random*: Mecanismo de dato faltante por causas completamente aleatorias.

MAR: *Missing at random*: Mecanismo de dato faltante por causas aleatorias.

MNAR: *Missing not at random*: Mecanismo de dato faltante por causas no aleatorias.

MI o IM: *multiple imputation*: Imputación Múltiple.

MICE: *Multivariate Imputation by Chained Equations*: Algoritmo que sigue los métodos FCS.

ML o MV: *Maximum-likehood*: Máxima Verosimilitud

PMM: *Predictive mean matching*: Técnica univariante de imputación múltiple.

SEM: *Structural equation modeling*: Modelos de ecuaciones estructurales, familia de modelos estadísticos multivariantes.

TFM: Trabajo Final de Máster

Capítulo 8

Bibliografía

[1] Little RJA, Rubin DB. Statistical analysis with missing data. Tercera edición ed. Wiley series in probability and statistics. Hoboken, NJ: Wiley; 2020. Available from: https://learning.oreilly.com/library/view/statistical-analysis-with/9780470526798/?sso_link=yes&sso_link_from=Catalunya.

[2] for Medicinal Products for Human Use (CHMP) C. Guideline missing data confirmatory clinical trials.pdf; 2010. EMA/CPMP/EWP/1776/99. Available from: https://www.ema.europa.eu/en/documents/scientific-guideline/guideline-missing-data-confirmatory-clinical-trials_en.pdf.

[3] Bennett DA. How can I deal with missing data in my study? Australian and New Zealand Journal of Public Health. 2001;25(5):464-9. Available from: https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1467-842X.2001.tb00294.x.

[4] Baraldi AN, Enders CK. An introduction to modern missing data analyses. Journal of School Psychology. 2010;48(1):5-37.

[5] Roberts MB, Sullivan MC, Winchester SB. Examining solutions to missing data in longitudinal nursing research. Journal for specialists in pediatric nursing: JSPN. 2017;22(2).

[6] Graham JW. Missing Data Analysis: Making It Work in the Real World. Annual Review of Psychology. 2009;60(1):549-76. Available from: https://www.annualreviews.org/doi/10.1146/annurev.psych.58.110405.085530.

[7] Barnes SA, Lindborg SR, Seaman Jr JW. Multiple imputation techniques in small sample clinical trials. Statistics in Medicine. 2006;25(2):233-45. Available from: https://onlinelibrary.wiley.com/doi/abs/10.1002/sim.2231.

[8] Mallol Roselló PJ. Importancia del tratamiento de datos perdidos. Aplicación en estudios longitudinales pequeños; 2017. Accepted: 2017-06-21 Publisher: Universitat Oberta de Catalunya. Available from: http://openaccess.uoc.edu/webapps/o2/handle/10609/64105.

[9] Kleinke K. Multiple imputation by predictive mean matching when sample size is small. Methodology. 2018.

[10] Mirzaei A, Carter SR, Patanwala AE, Schneider CR. Missing data in surveys: Key concepts, approaches, and applications. Research in Social and Administrative Pharmacy. 2022;18(2):2308-16.

[11] Little RJA. A Test of Missing Completely at Random for Multivariate Data with Missing Values. Journal of the American Statistical Association. 1988;83(404):1198-202. Publisher: [American Statistical Association, Taylor & Francis, Ltd.]. Available from: https://www.jstor.org/stable/2290157.

[12] R Core Team. R: A Language and Environment for Statistical Computing. Vienna, Austria; 2021. Available from: https://www.R-project.org/.

[13] Reina Ortiz M, Sharma V. Diseños de Investigación. Revista Médica Vozandes Volumen 23, Número 2, 2012. Revista Médica Voz Andes. 2012;23:95.

[14] Delgado Rodríguez M, Llorca Díaz J. Estudios longitudinales: concepto y particularidades. Revista española de salud pública. 2004;78:141-8.

[15] Pérez Andrés C, Martín Moreno JM. Sobre los estudios longitudinales en epidemiología. SciELO Public Health; 2004.

[16] Caruana EJ, Roman M, Hernández-Sánchez J, Solli P. Longitudinal studies. Journal of thoracic disease. 2015;7(11):E537.

[17] Graham JW. Missing data theory. In: Missing data. Springer; 2012. p. 346.

[18] Marshall A, Altman DG, Holder RL. Comparison of imputation methods for handling missing covariate data when fitting a Cox proportional hazards model: a resampling study.

BMC Medical Research Methodology. 2010;10(1):112. Available from: https://doi.org/10.1186/1471-2288-10-112.

[19] Powney M, Williamson P, Kirkham J, Kolamunnage-Dona R. A review of the handling of missing longitudinal outcome data in clinical trials. Trials. 2014;15(1):237. Available from: https://doi.org/10.1186/1745-6215-15-237.

[20] Wood AM, White IR, Thompson SG. Are missing outcome data adequately handled? A review of published randomized controlled trials in major medical journals. Clinical Trials. 2004;1(4):368-76. Available from: http://journals.sagepub.com/doi/10.1191/1740774504cn032oa.

[21] Bell ML, Fiero M, Horton NJ, Hsu CH. Handling missing data in RCTs; a review of the top medical journals. BMC Medical Research Methodology. 2014;14(1):118. Available from: https://doi.org/10.1186/1471-2288-14-118.

[22] Rubin DB. Inference and missing data. Biometrika. 1976;63(3):581-92. Available from: https://doi.org/10.1093/biomet/63.3.581.

[23] Enders CK. Applied missing data analysis. Guilford press; 2010.

[24] García JG, Albaladejo JP, Fernández JAM. Métodos de inferencia estadística con datos faltantes: estudio de simulación sobre los efectos en las estimaciones. Estadística española. 2006;48(162):241-70.

[25] Savalei V. Small sample statistics for incomplete nonnormal data: Extensions of complete data formulae and a Monte Carlo comparison. Structural Equation Modeling. 2010;17(2):241-64.

[26] Gaffert P, Meinfelder F, Bosch V. Towards an MI-proper predictive mean matching. English URL: https://www uni-bamberg de/fileadmin/uni/fakultaeten/sowi_lehrstuehle/statistik/Personen/Dateien_Florian/properPMM pdf. 2016.

[27] Van Buuren S. Flexible imputation of missing data. CRC press; 2018.

[28] Echeverria P, Negredo E, Carosi G, Gálvez J, Gómez J, Ocampo A, et al. Similar antiviral efficacy and tolerability between efavirenz and lopinavir/ritonavir, administered with

abacavir/lamivudine (Kivexa®), in antiretroviral-naïve patients: A 48-week, multicentre, randomized study (Lake Study). Antiviral research. 2010;85(2):403-8.

[29] Tierney N, Cook D, McBain M, Fay C. naniar: Data Structures, Summaries, and Visualisations for Missing Data; 2021. R package version 0.6.1. Available from: https://CRAN.R-project.org/package=naniar.

[30] Kowarik A, Templ M. Imputation with the R Package VIM. Journal of Statistical Software. 2016;74(7):1-16.

[31] Van Buuren S, Groothuis-Oudshoorn K. mice: Multivariate imputation by chained equations in R. Journal of statistical software. 2011;45(3):1-67.

[32] Jamshidian M, Jalal S, Jansen C. MissMech: An R Package for Testing Homoscedasticity, Multivariate Normality, and Missing Completely at Random (MCAR). Journal of Statistical Software. 2014;56(6):1-31. Available from: http://www.jstatsoft.org/v56/i06/.

[33] Castro Cacabelos M. Imputación de datos faltantes en un modelo de tiempo de fallo acelerado. Universidad de Coruña; 2014. Available from: http://eio.usc.es/pub/mte/descargas/ProyectosFinMaster/Proyecto_940.pdf.

[34] Hothorn T, Hornik K, van de Wiel MA, Zeileis A. Implementing a class of permutation tests: The coin package. Journal of Statistical Software. 2008;28(8):1-23.

[35] Spohn ML, Michel L, Naef J. PKLMtest: Classification Based MCAR Test; 2021. R package version 1.0.1. Available from: https://CRAN.R-project.org/package=PKLMtest.

[36] Shin T, Davison ML, Long JD. Maximum likelihood versus multiple imputation for missing data in small longitudinal samples with nonnormality. Psychological methods. 2017;22(3):426.

Apéndice A

Anexos

A.1. Análisis de *missings*

A.1.1. Gráficos QQplot

A.1.2. Patrón de *missings*

A.1.3. matriz predictora

A.2. *Stripplots*

A.2.1. Alternativa 1: Por defecto

A.2.2. Alternativa 2: *PMM*

A.2.3. Alternativa 3: *Midastouch*

A.3. *Densityplots*

A.3.1. Alternativa 1: Por defecto

A.3.2. Alternativa 2: *PMM*

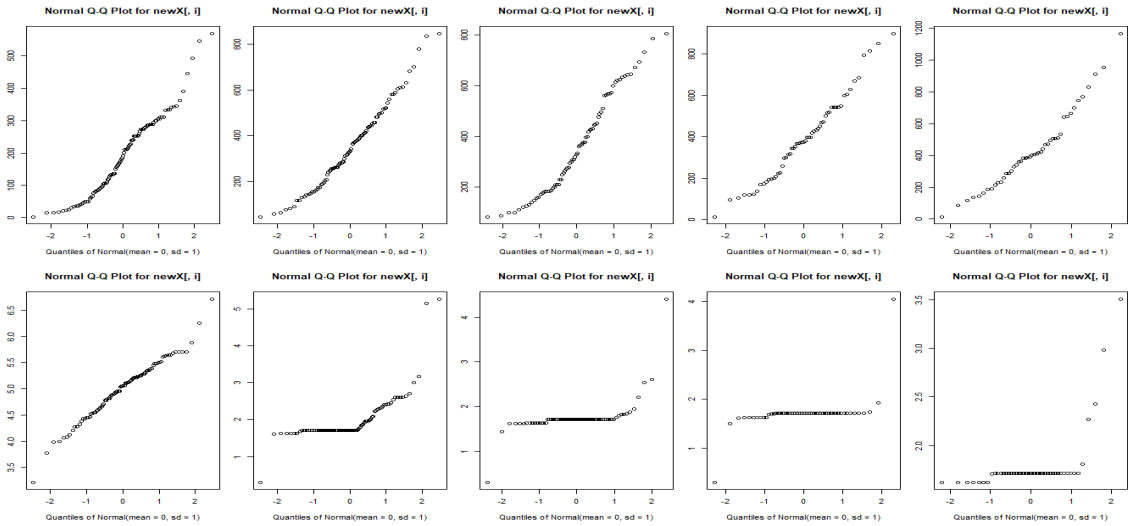

Figura A.1: Gráficos QQplot (Arriba) Variable CD4A. (Abajo) log_{10} de Carga viral

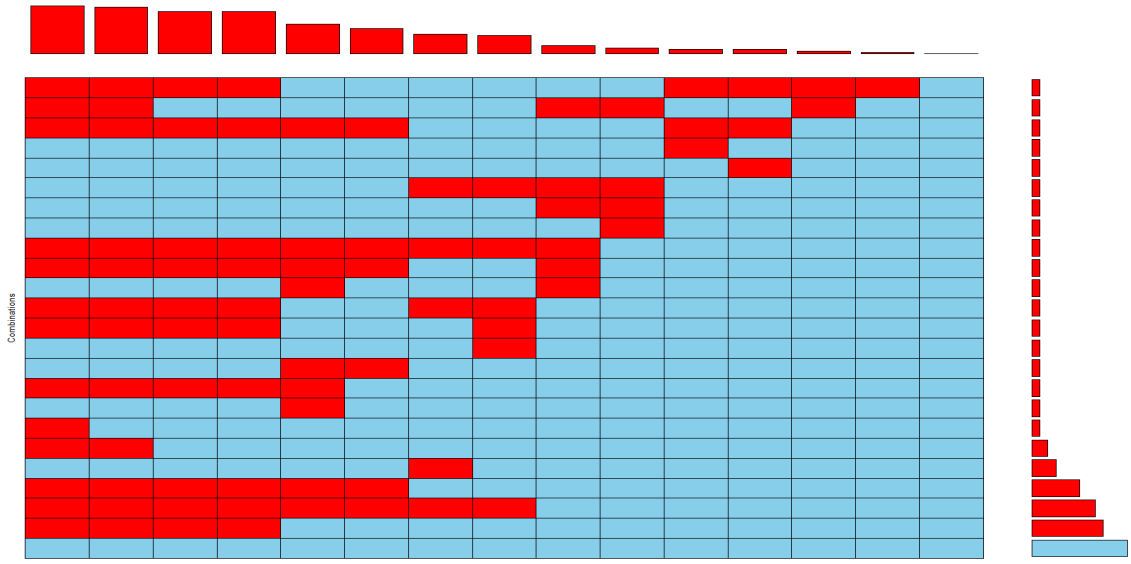

Figura A.2: Patrón de *missings* por variables y observaciones de la base de datos VIH2

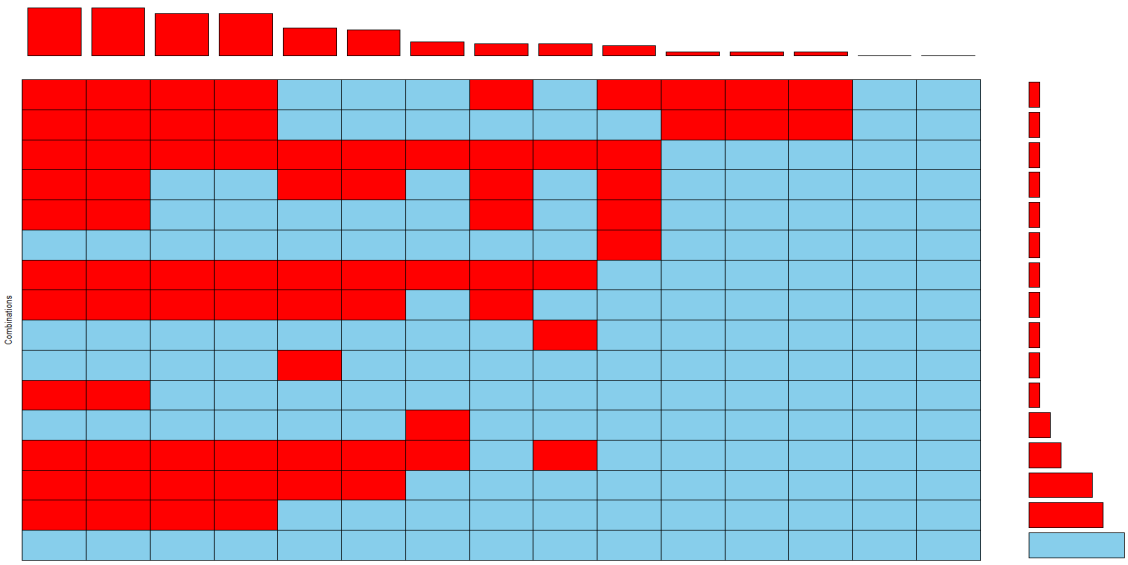

Figura A.3: Patrón de *missings* por variables y observaciones de la base de datos VIH3

	Grupo	sexo	edad	t_vih_mes	factor_r_t	CD4A_0	CViral_0	CD4A_12	CViral_12	CD4A_24	CViral_24	CD4A_36	CViral_36	CD4A_48	CViral_48
Grupo	0	1	1	1	1	1	1	1	1	1	1	1	1	1	0
sexo	1	0	1	1	1	1	1	1	1	1	1	1	1	1	0
edad	1	1	0	1	1	1	1	1	1	1	1	1	1	1	0
t_vih_mes	1	1	1	0	1	1	1	1	1	1	1	1	1	1	0
factor_r_t	1	1	1	1	0	1	1	1	1	1	1	1	1	1	0
CD4A_0	1	1	1	1	0	0	1	1	1	1	1	1	1	1	0
CViral_0	1	1	1	1	0	1	0	1	1	1	1	1	1	1	0
CD4A_12	1	1	1	1	0	1	1	0	1	1	1	1	1	1	0
CViral_12	1	1	1	1	0	1	1	1	0	1	1	1	1	1	0
CD4A_24	1	1	1	1	0	1	1	1	1	0	1	1	1	1	0
CViral_24	1	1	1	1	0	1	1	1	1	1	0	1	1	1	0
CD4A_36	1	1	1	1	0	1	1	1	1	1	1	0	1	1	0
CViral_36	1	1	1	1	0	1	1	1	1	1	1	1	0	1	0
CD4A_48	1	1	1	1	0	1	1	1	1	1	1	1	1	0	0
CViral_48	1	1	1	1	0	0	1	0	1	0	0	0	1	0	0

Figura A.4: matriz predictora de la función *mice* utilizada con la base de datos VIH3. En rojo los cambios de valor 1 a 0. En azul los cambios de valor 0 a 1.

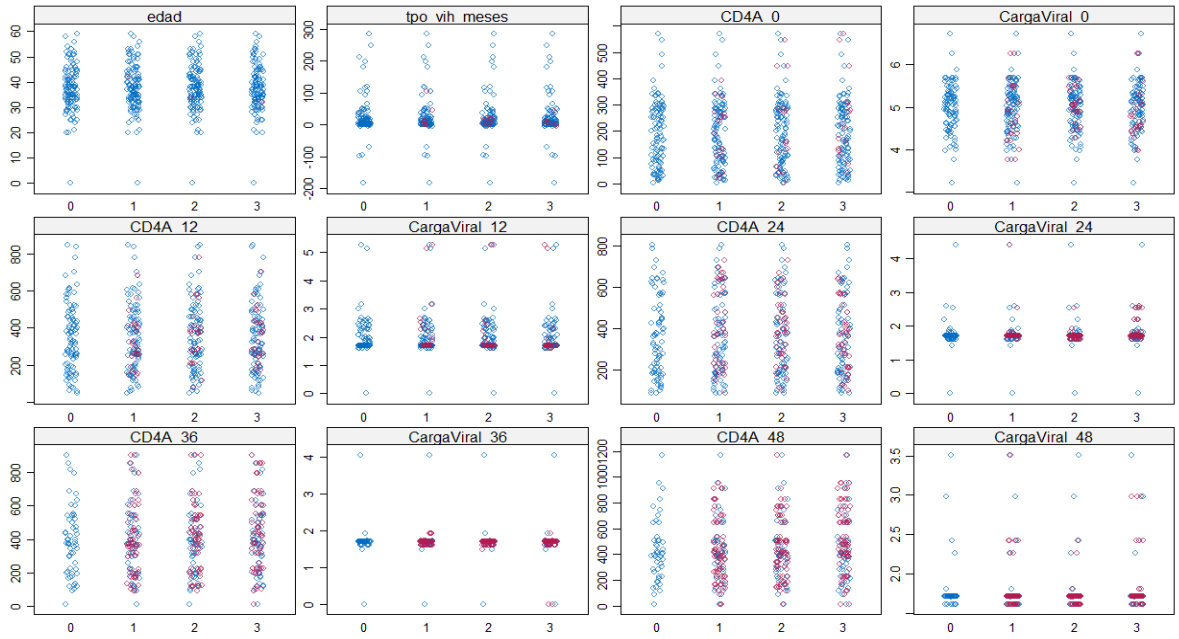

Figura A.5: *Stripplot* de los valores imputados de la base de datos VIH con m = 3

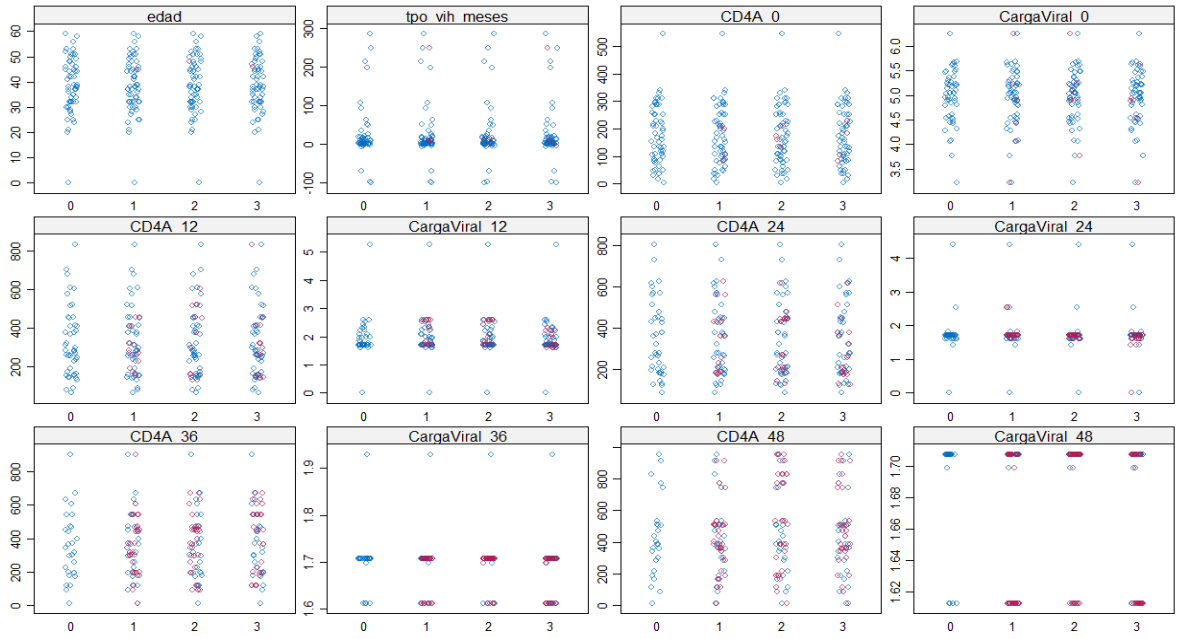

Figura A.6: *Stripplot* de los valores imputados de la base de datos VIH2 con m = 3

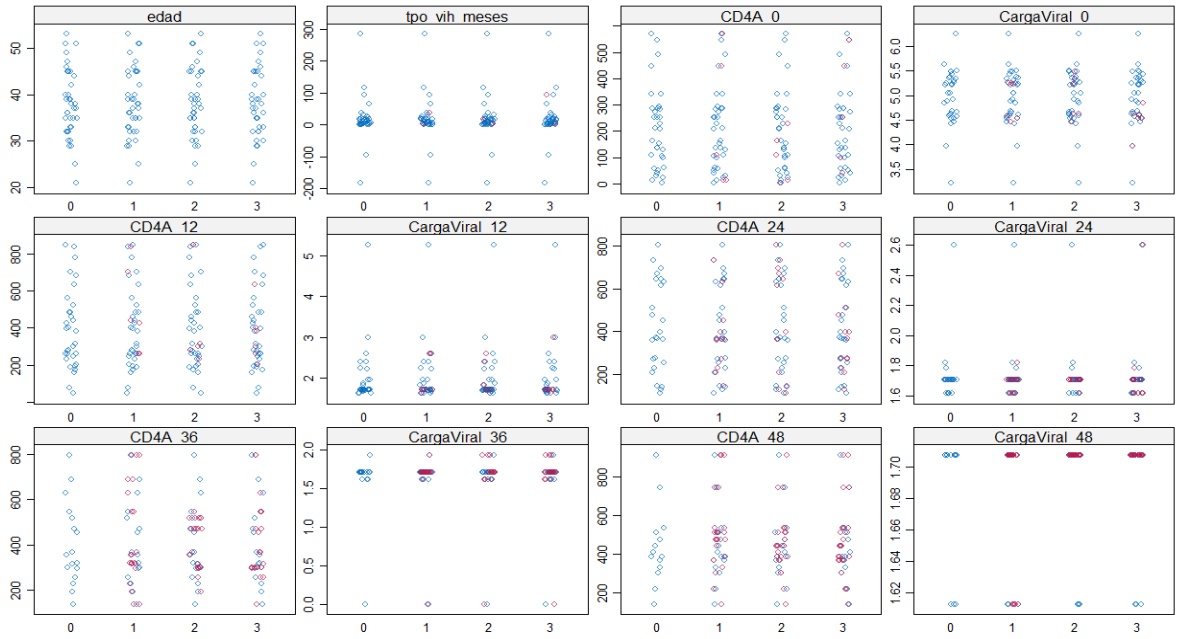

Figura A.7: *Stripplot* de los valores imputados de la base de datos VIH3 con m = 3

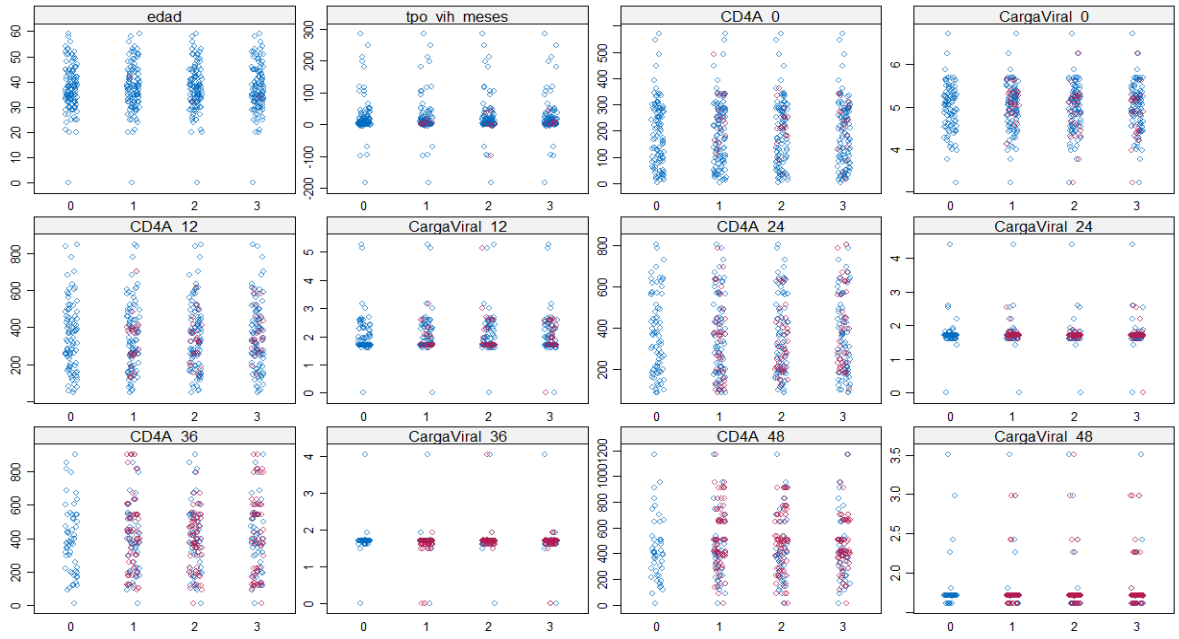

Figura A.8: *Stripplot* de los valores imputados de la base de datos VIH con m = 3

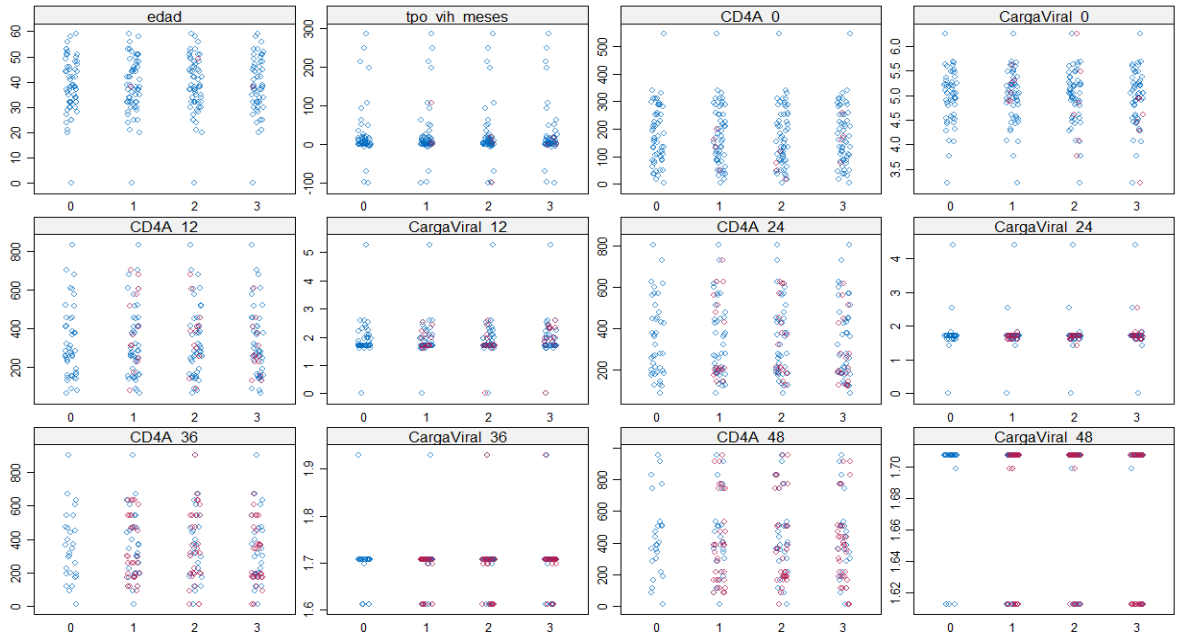

Figura A.9: *Stripplot* de los valores imputados de la base de datos VIH2 con m = 3

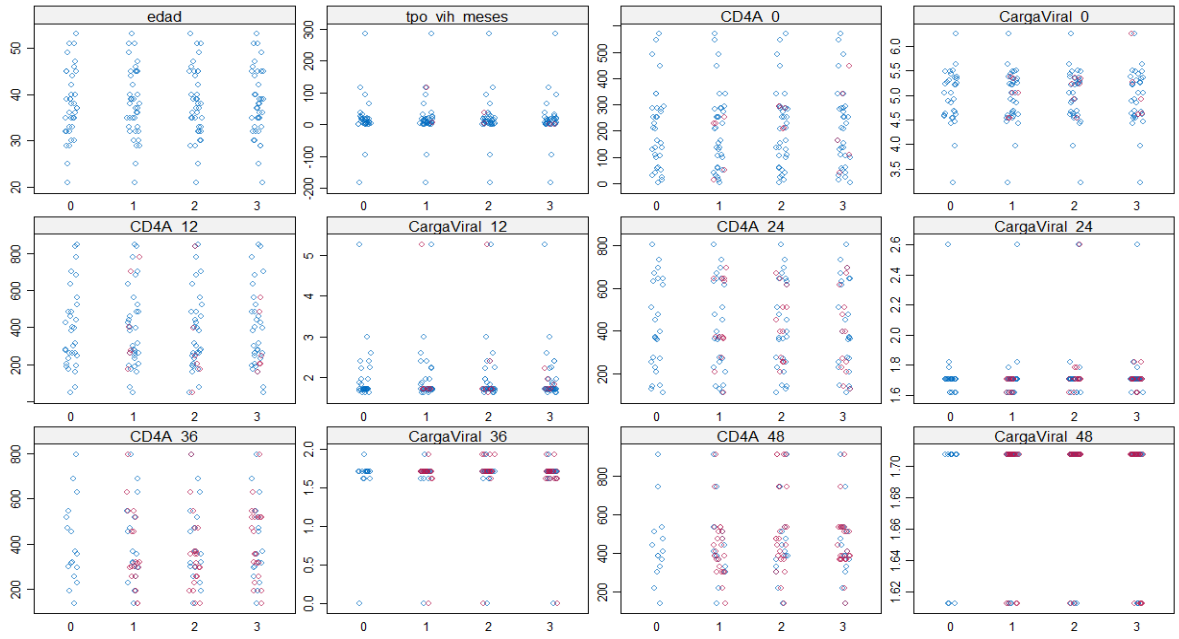

Figura A.10: *Stripplot* de los valores imputados de la base de datos VIH3 con m = 3

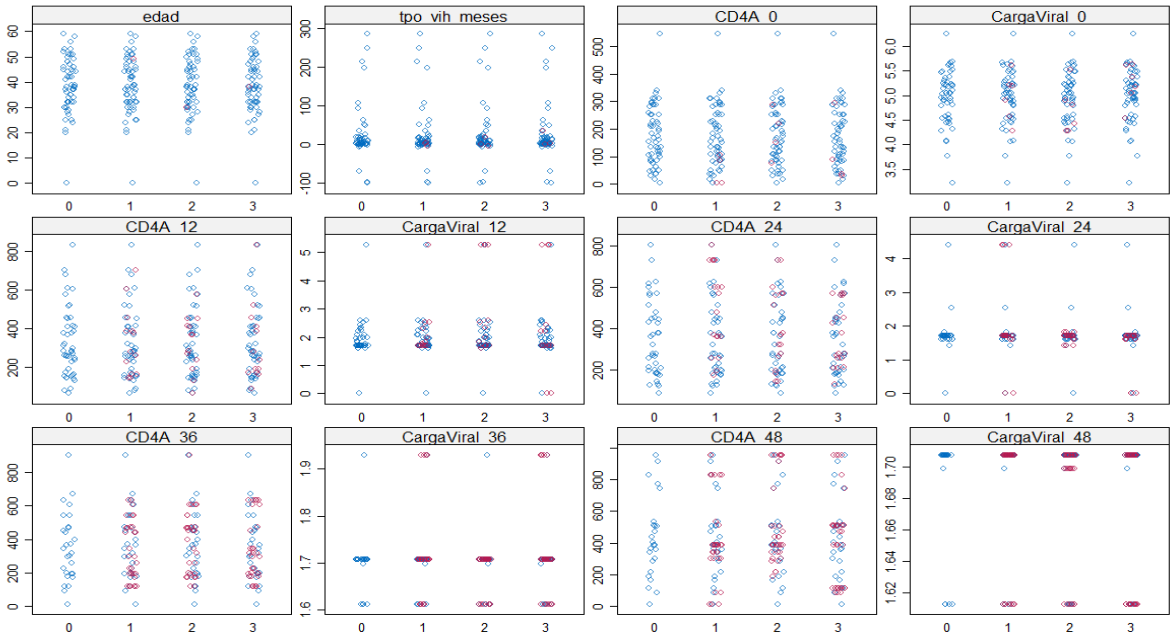

Figura A.11: *Stripplot* de los valores imputados de la base de datos VIH2 con m = 3

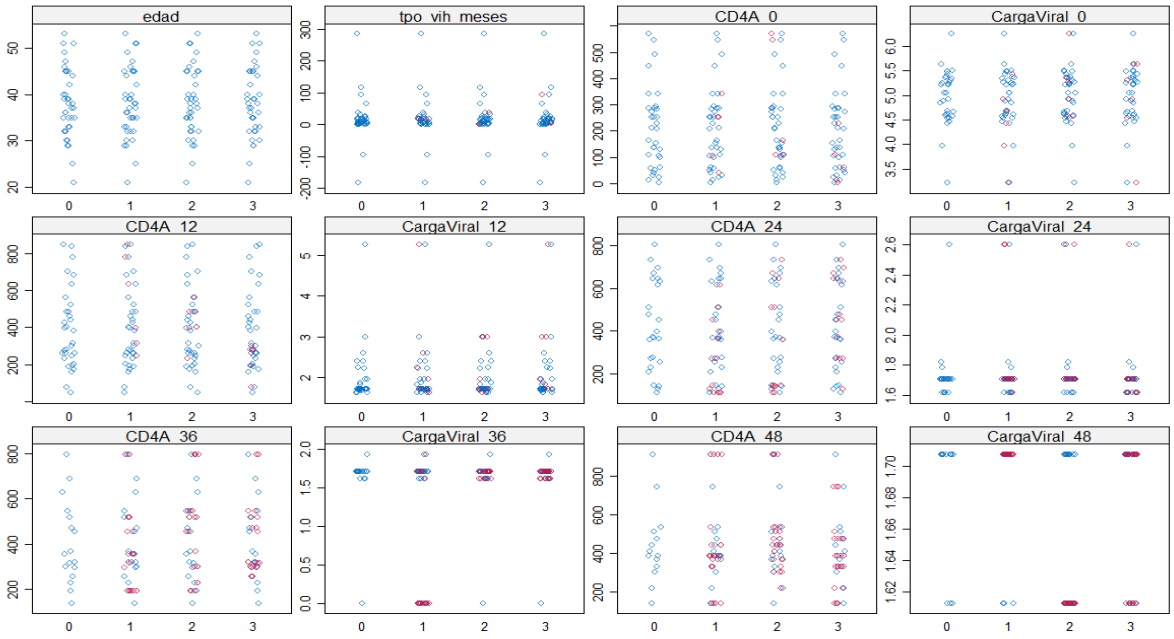

Figura A.12: *Stripplot* de los valores imputados de la base de datos VIH3 con m = 3

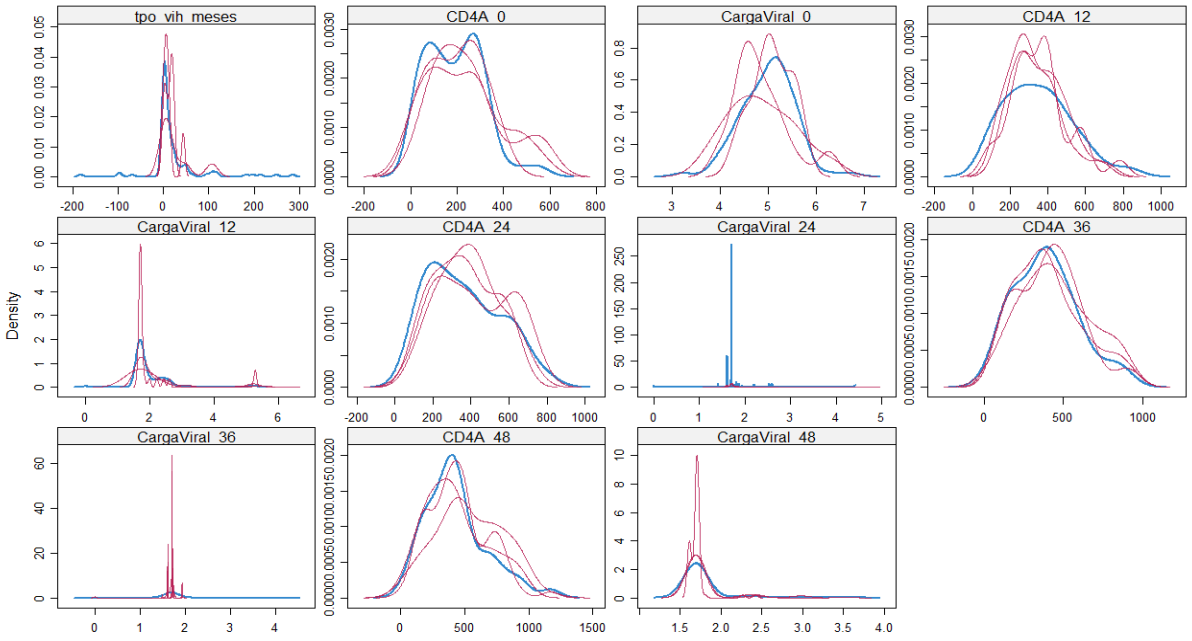

Figura A.13: *Densityplot* de los valores imputados de la base de datos VIH con m = 3

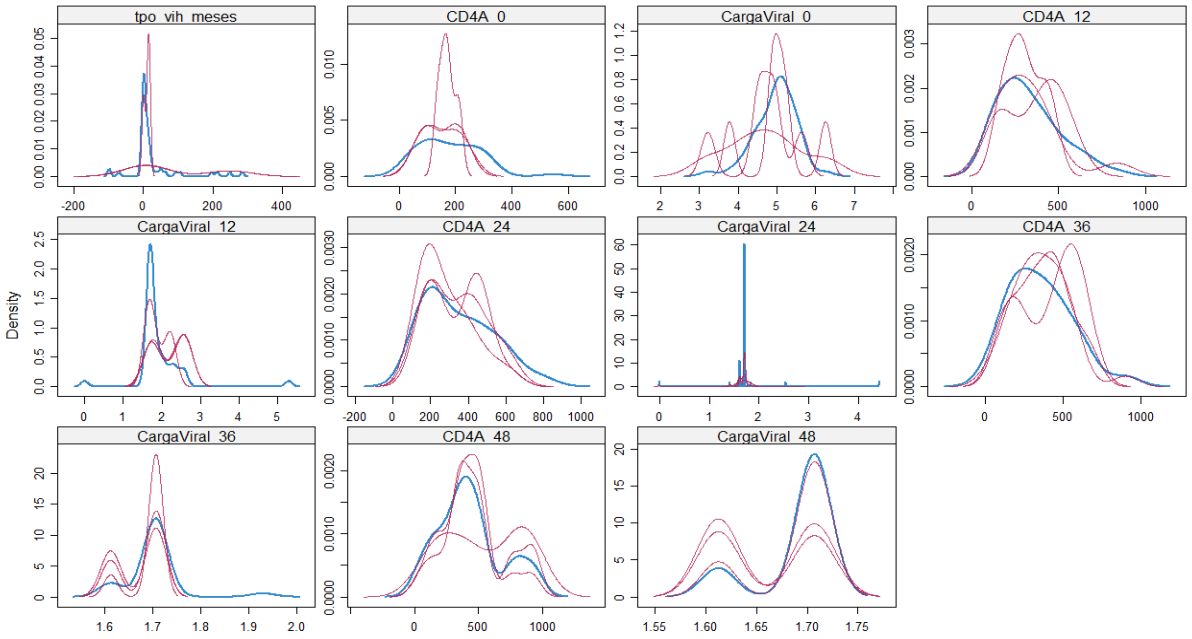

Figura A.14: *Densityplot* de los valores imputados de la base de datos VIH2 con m = 3

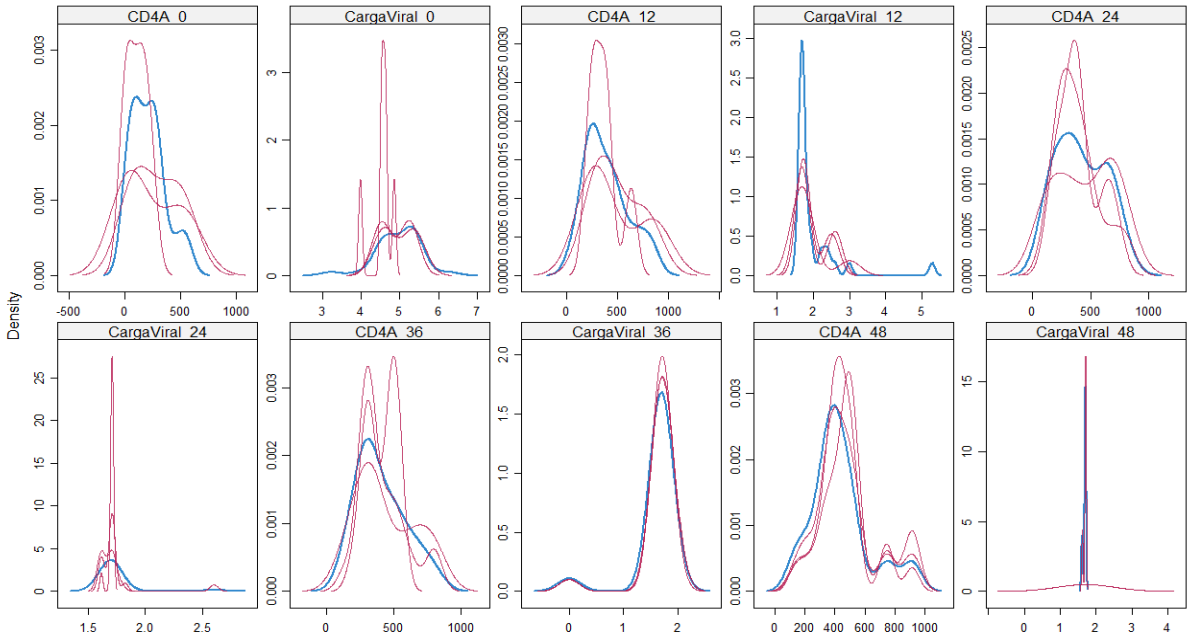

Figura A.15: *Densityplot* de los valores imputados de la base de datos VIH3 con m = 3

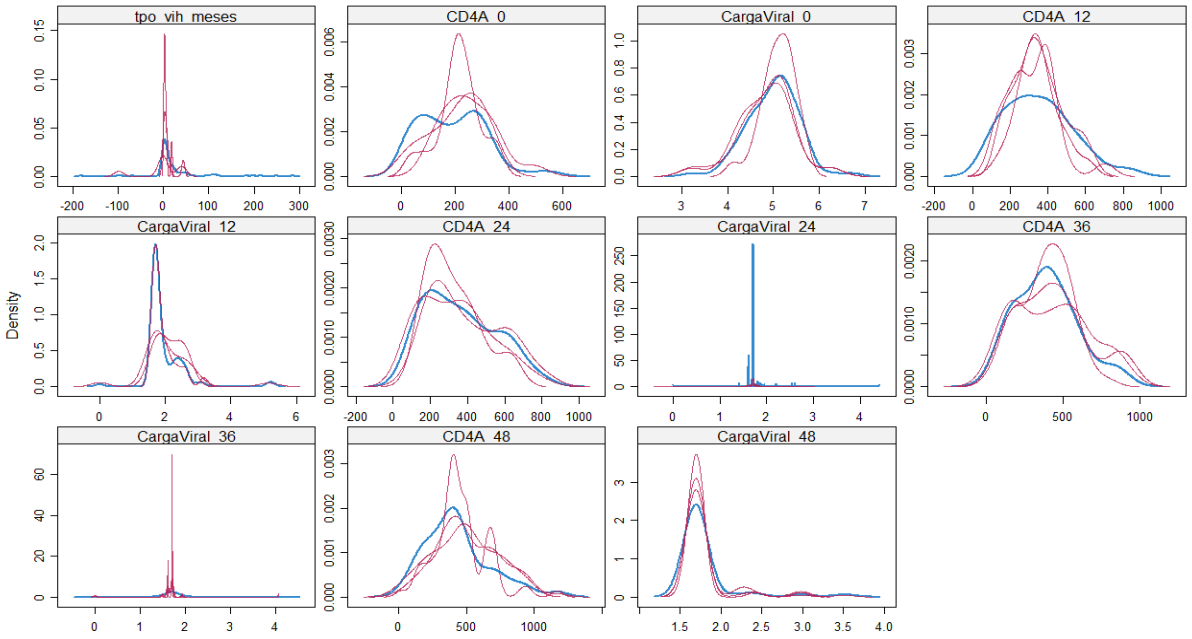

Figura A.16: *Densityplot* de los valores imputados de la base de datos VIH con m = 3

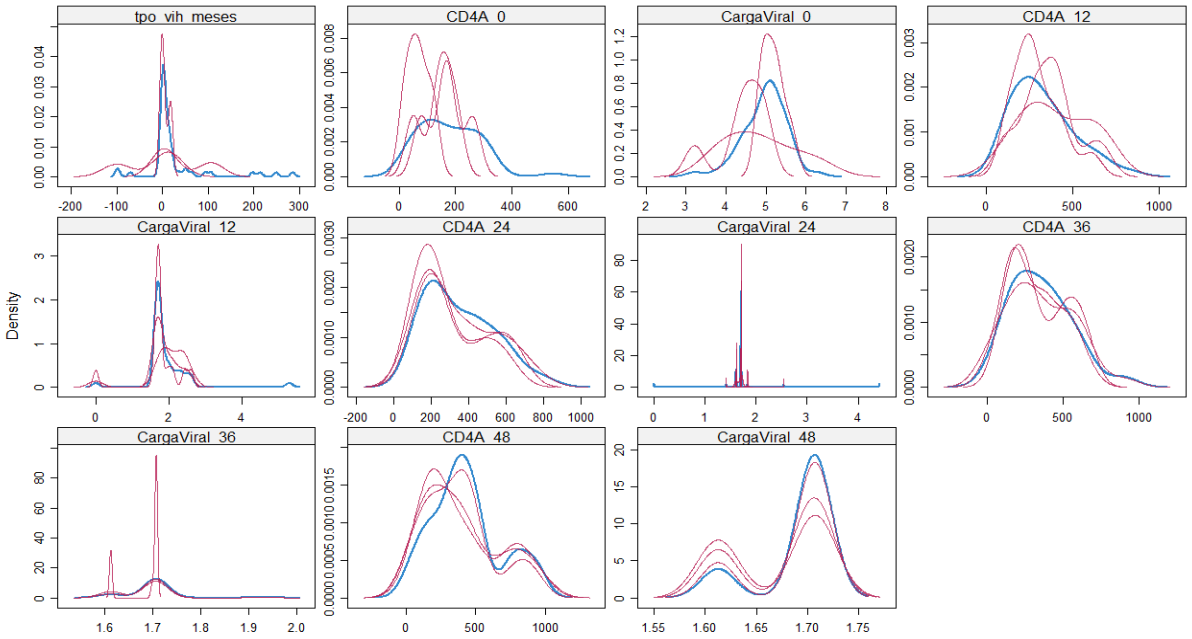

Figura A.17: *Densityplot* de los valores imputados de la base de datos VIH2 con m = 3

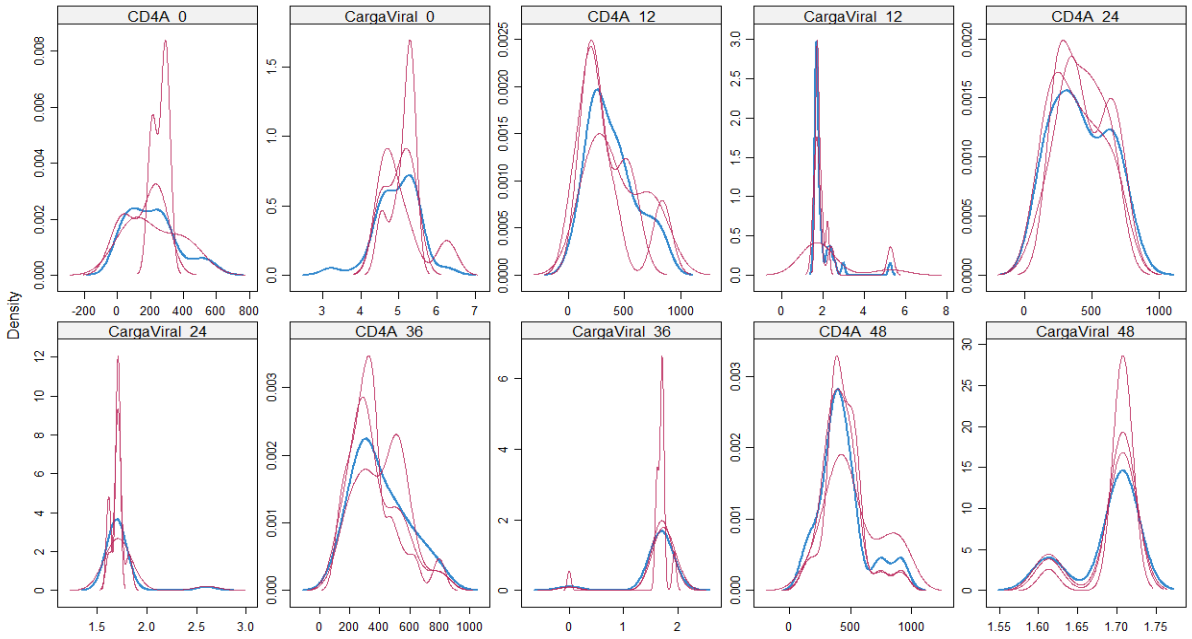

Figura A.18: *Densityplot* de los valores imputados de la base de datos VIH3 con m = 3

Apéndice B

Código R

```
## ----Librerías-----------------------------------------------------------

library(dplyr)
library(knitr)
library(naniar)
library(visdat)
library(ggplot2)
library(VIM)
library(MissMech)
library(mice)
library(tidyverse)
library(EnvStats)
library(coin)

## ----Opciones------------------------------------------------------------

# Definimos los parámetros para los chunks del código.
options(max.print="75")
opts_chunk$set(echo=TRUE,
               cache=FALSE,
```

```
                prompt=FALSE,
                comment=NA,
                message=FALSE,
                warning=FALSE,
                fig.align='center')
opts_knit$set(width=75)

## ----Carga de datos-------------------------------------------------------

vih <- data.frame(read.csv("lake.csv", sep = ";", dec = ","))

# Dimensiones de la base de datos original
dim(vih)

## ----Obtención de variables-----------------------------------------------

# Seleccionamos las variables de interés
VIH <- select(vih, Grupo,sexo,edad,tpo_vih_meses,factor_riesgo_total,
                starts_with("CD4A"), starts_with("CargaV"))

# Transformamos en factores las variables categóricas
cols <- c("Grupo","sexo", "factor_riesgo_total")
VIH[cols] <- lapply(VIH[cols], factor)

# Asignación de niveles de cada factor
levels(VIH$Grupo) <-  c("EFV", "LVP/r")
levels(VIH$sexo)  <- c("Masculino", "Femenino")
levels(VIH$factor_riesgo_total) <- c("ADVP","Heterosexual","Homosexual","Otros")

# Transformación variables CargaViral_ a log10(x+1)
VIH[11:15] <- VIH %>% select(starts_with("CargaV")) %>% +1 %>%  log10()

# Transformación de un outlier encontrado en un valor NA
```

```
VIH[60,"CD4A_12"] <- NA

## ----Estructura VIH-------------------------------------------------------

# Estructura de la base de datos y la clase de cada variable
str(VIH)

# Gráfico estructura de datos y missings
vis_dat(VIH) + ylab("Observaciones") + guides(fill = guide_legend(title = "Tipo")) +
  scale_fill_hue(labels = c("Factor", "Entero", "Numérico")) +
  theme(axis.text = element_text(size = 15),
        axis.title.y = element_text(size=rel(1.5)),
        legend.title=element_text(size=15),
        legend.text=element_text(size=15))

## ----Resumen VIH----------------------------------------------------------

# Resumen estadístico de la base de datos
summary(VIH)

# Proporción entre grupo y sexo
prop.table(table(VIH$Grupo,VIH$sexo))

#Proporción por factor de riesgo
prop.table(table(VIH$factor_riesgo_total))

## ----boxplot VIH----------------------------------------------------------

par(mfrow = c(1,2))
VIH %>% select(starts_with("CD4")) %>% boxplot(col = "skyblue")
VIH %>% select(starts_with("CargaViral")) %>% boxplot(col = "orange", las = 0)

## ----qqplots--------------------------------------------------------------
```

```
par(mfrow = c(2,5))
VIH[-c(1:5)] %>% apply(2,qqPlot)

## ----Base de datos-----------------------------------------------------

# Generar las bases de datos N/2 y N/3

# Se fija la semilla
set.seed(545)

# La función slice_sample requiere el número de filas por grupo
# Se obtiene la mitad de las filas originales
VIH2 <- ungroup(VIH %>% group_by(Grupo) %>% slice_sample(n = nrow(VIH)/4))

# Se obtiene un tercio de las filas de la base original
VIH3 <- ungroup(VIH %>% group_by(Grupo) %>% slice_sample(n = nrow(VIH)/6))

## ----dimensiones Bases de datos----------------------------------------

D <- c("VIH","VIH2","VIH3")
N <-c(dim(VIH)[1],dim(VIH2)[1], dim(VIH3)[1])
Nv <- c(dim(VIH)[2],dim(VIH2)[2], dim(VIH3)[2])
M <- c(n_miss(VIH),n_miss(VIH2),n_miss(VIH3))
Mp <- round(c(prop_miss(VIH),prop_miss(VIH2),prop_miss(VIH3)),3)
VIH_DF <- data.frame("DF" = D , "N_obs" = N, "N_var" = Nv, "N_miss" = M,
                     "P_miss" = Mp*100)
VIH_latex <- knitr::kable(VIH_DF)

## ----missing BDs-------------------------------------------------------

# Se obtiene el número de missings por variable
mv_1 <- data.frame(miss_var_summary(VIH))
```

```
mv_2 <-data.frame(miss_var_summary(VIH2))
mv_3 <-data.frame(miss_var_summary(VIH3))

# Se unen las bases de datos
mv_12 <- merge(mv_1,mv_2,by = "variable")
mv_total <- merge(mv_12,mv_3,by = "variable")

# Se selecciona las variables de interés y se ordena de forma decreciente
mv_var <- mv_total %>% select(variable, starts_with("pct"))
colnames(mv_var) <- c("Variable","VIH","VIH2","VIH3")
mv_var <- arrange(mv_var,desc(VIH))
mv_var[-1] <- round(mv_var[-1],2)

# Se guarda en formato latex
mv_t_latex <- knitr::kable(mv_var)

## ----Comparación missings-----------------------------------------------

mv_var <- arrange(mv_var,Variable)

# Porcentajes CD4A
mv_var_cd4a <- mv_var[6:10,]
mv_cd4 <- mv_var_cd4a %>% mutate(week = seq(0,48,12))

# Porcentajes Carga viral
mv_var_cv <- mv_var[1:5,]
mv_cv <- mv_var_cv %>% mutate(week = seq(0,48,12))

par(mfrow = c(1,2))
# Gráfico aumento de missings: variable CD4A en cada base de datos
plot(mv_cd4$week,mv_cd4$VIH, type = "b", pch = 18, col = "blue", xlab = "Semanas",
     ylab = "missings (%)", xlim = c(0,48), ylim = c(0,65),
     main = "Valor absoluto de CD4", xaxt = "n",
```

```
     cex.main = 2, cex.lab = 1.5, cex.axis = 1)
lines(mv_cd4$week,mv_cd4$VIH2, type = "b", pch = 20, col = "red")
lines(mv_cd4$week,mv_cd4$VIH3, type = "b", pch = 20, col = "black")
legend("topleft", legend=c("VIH", "VIH2", "VIH3"),
       col=c("blue","red","black"), lty = 1, cex=1.5)
axis(1,at = seq(0,48,12))

# Gráfico aumento de missings: variable carga viral en cada base de datos
plot(mv_cv$week,mv_cv$VIH, type = "b", pch = 18, col = "blue", xlab = "Semanas",
     ylab = "", xlim = c(0,48), ylim = c(0,65), main = "Log10 Carga viral de VIH",
     xaxt = "n", cex.main = 2, cex.lab = 1.5, cex.axis = 1)
lines(mv_cv$week,mv_cv$VIH2, type = "b", pch = 20, col = "red")
lines(mv_cv$week,mv_cv$VIH3, type = "b", pch = 20, col = "black")
legend("topleft", legend=c("VIH", "VIH2", "VIH3"),
       col=c("blue","red","black"), lty = 1, cex=1.5)
axis(1,at = seq(0,48,12))

## ----patrón VIM-------------------------------------------------------

#Gráficos patrón utilizando el paquete "VIM"
aggr(VIH, sortVars = T, prop = T, sortCombs = T,  cex.axis = 0.75, combined = T,
    axes = F)
aggr(VIH2, sortVars = T, prop = T, sortCombs = T,  cex.axis = 0.75, combined = T,
    axes = F)
aggr(VIH3, sortVars = T, prop = T, sortCombs = T,  cex.axis = 0.75, combined = T,
    axes = F)

## ----influx outflux---------------------------------------------------

# Tabla de valores influx y outflux de cada variable
flux(VIH)[,1:3]
flux(VIH2)[,1:3]
flux(VIH3)[,1:3]
```

```
## ---- Gráfico fluxplot-----------------------------------------------------

par(mfrow = c(1,3))
fluxplot(VIH, main = "VIH", labels = FALSE, cex.main = 2, cex.lab = 2, xlab = "")
fluxplot(VIH2, main = "VIH2", labels = FALSE, cex.main = 2, cex.lab = 2, ylab = "")
fluxplot(VIH3, main = "VIH3", labels = FALSE, cex.main = 2, cex.lab = 2, ylab = "",
        xlab = "")

## ----TestMCAR--------------------------------------------------------------

# Test MCAR puesto en prueba para la base de datos VIH.
VIH %>% select("edad",tpo_vih_meses,starts_with("C"))  %>% TestMCARNormality()

## ----función orden_IM------------------------------------------------------

# Función para ordenar variables temporales
orden_IM <- function(x){
  select(x,Grupo,sexo,edad,tpo_vih_meses,factor_riesgo_total,
              ends_with("_0"),ends_with("_12"),
              ends_with("_24"),ends_with("_36"),ends_with("_48"))
}

## ----orden BD--------------------------------------------------------------

# Aplicación de la función
VIH <- orden_IM(VIH)
VIH2 <- orden_IM(VIH2)
VIH3 <- orden_IM(VIH3)

## ----mice sin método-------------------------------------------------------

# Imputación mútiple, utilizando MICE
```

```
VIH_SM <- mice(VIH, m=3, maxit = 5, print = FALSE, seed = 457)
VIH_SM2 <- mice(VIH2, m=3, maxit = 5, print = FALSE, seed = 457)
VIH_SM3 <- mice(VIH3, m=3, maxit = 5,  print = FALSE, seed = 457)

# Listado de eventos registrados (logged event)
VIH_SM3$loggedEvents

# Tratamiento del logged event
prep_SM <- VIH_SM3$pred  # Se obtiene la matriz de predictora

# Se modifican los predictores
prep_SM[6:14,"factor_riesgo_total"] <- 0
prep_SM["CargaViral_48",c(1:4,7,9,13)] <- 1

# Se utiliza la matriz de predictora modificada
VIH_SM3 <- mice(VIH3, m=3, maxit = 5, print = FALSE, pred = prep_SM,seed = 457)

# Se puede observar que métodos fueron asignados por la función mice
VIH_SM$method

## ----Gráficos_IM-----------------------------------------------------------

densityplot(VIH_SM)
stripplot(VIH_SM)
densityplot(VIH_SM2)
stripplot(VIH_SM2)
densityplot(VIH_SM3)
stripplot(VIH_SM3)

## ----Añadir imputados------------------------------------------------------

VIH_SIM <- complete(VIH_SM)
VIH_SIM2 <- complete(VIH_SM2)
```

```
VIH_SIM3 <- complete(VIH_SM3)

## ----mice pmm------------------------------------------------------------

# Imputación con el método PMM, utilizando MICE
VIH_pmm <- mice(VIH, m=3, maxit = 5, method = "pmm", print = FALSE, seed = 457)
VIH_pmm2 <- mice(VIH2, m=3, maxit = 5, method = "pmm", print = FALSE, seed = 457)
VIH_pmm3 <- mice(VIH3, m=3, maxit = 5, method = "pmm", print = FALSE, seed = 457)

# Listado de eventos registrados (logged event)
VIH_pmm3$loggedEvents

# Tratamiento del logged event
prep3 <- VIH_pmm3$pred
prep3[6:14,"factor_riesgo_total"] <- 0
prep3["CargaViral_48", c(1:4,7,9,13)] <- 1

# Se utiliza la matriz predictora modificada
VIH_pmm3 <- mice(VIH3, m=3, maxit = 5, method = "pmm", print = FALSE,
                 pred = prep3,seed = 457)

## ----Gráficos_IM_pmm-----------------------------------------------------

densityplot(VIH_pmm)
stripplot(VIH_pmm)
densityplot(VIH_pmm2)
stripplot(VIH_pmm2)
densityplot(VIH_pmm3)
stripplot(VIH_pmm3)

## ----Añadir imputados pmm------------------------------------------------

VIH_IM <- complete(VIH_pmm)
```

```
VIH_IM2 <- complete(VIH_pmm2)
VIH_IM3 <- complete(VIH_pmm3)

## ----mice midas-------------------------------------------------------------

method <- c("","logreg","midastouch","midastouch","polyreg","midastouch",
            "midastouch", "midastouch","midastouch","midastouch","midastouch",
            "midastouch","midastouch", "midastouch","midastouch")

# Imputación con el método PMM, utilizando MICE

VIH_midas <- mice(VIH, m = 3, method = method, print = FALSE, seed = 457)
VIH_midas2 <- mice(VIH2, m = 3, method = method, print = FALSE, seed = 457)
VIH_midas3 <- mice(VIH3, m = 3, method = method, print = FALSE, seed = 457)

# Listado de eventos registrados (logged event)
VIH_midas3$loggedEvents

# Tratamiento del logged event
premi3 <- VIH_midas3$pred
premi3[6:15,"factor_riesgo_total"] <- 0
premi3["CargaViral_48", c(1:4,7,9,13)] <- 1

# Se utiliza la matriz predictora modificada
VIH_midas3 <- mice(VIH3, m = 3,method = method, print = FALSE, pred = premi3,
                   seed = 457)

## ----Gráficos_IM_midas------------------------------------------------------

densityplot(VIH_midas)
stripplot(VIH_midas)
densityplot(VIH_midas2)
stripplot(VIH_midas2)
```

```
densityplot(VIH_midas3)
stripplot(VIH_midas3)

## ----Añadir imputados midas-----------------------------------------------

VIH_IMM <- complete(VIH_midas)
VIH_IMM2 <- complete(VIH_midas2)
VIH_IMM3 <- complete(VIH_midas3)

## ----función test---------------------------------------------------------

test <- function(variable,grupo){
  # Test para obtener el valor p exacto
  x <- wilcox_test(variable ~ grupo,conf.int = TRUE, distribution = "exact")
  # Valor p
  p <- pvalue(x)
  # Estadístico/valor Z
  z <- statistic(x)
  # Extraemos ambos valores redondeados
  c <- round(c(p,z),3)
  return(c)
}

## ----Variable CD4A_48-----------------------------------------------------

# Alternativa 1 Por defecto
U_SIM_CD <- test(VIH_SIM$CD4A_48,VIH_SIM$Grupo)
U_SIM2_CD <- test(VIH_SIM2$CD4A_48,VIH_SIM2$Grupo)
U_SIM3_CD <- test(VIH_SIM3$CD4A_48,VIH_SIM3$Grupo)

# Alternativa 2 PMM
U_IM_CD <- test(VIH_IM$CD4A_48,VIH_IM$Grupo)
U_IM2_CD <- test(VIH_IM2$CD4A_48,VIH_IM2$Grupo)
```

```
U_IM3_CD <- test(VIH_IM3$CD4A_48,VIH_IM3$Grupo)

# Alternativa 3 Midastouch
U_IMM_CD <- test(VIH_IMM$CD4A_48,VIH_IMM$Grupo)
U_IMM2_CD <- test(VIH_IMM2$CD4A_48,VIH_IMM2$Grupo)
U_IMM3_CD <- test(VIH_IMM3$CD4A_48,VIH_IMM3$Grupo)

## ----Variable CargaViral_48-----------------------------------------------

# Alternativa 1 Por defecto
U_SIM_CV <- test(VIH_SIM$CargaViral_48,VIH_SIM$Grupo)
U_SIM2_CV <- test(VIH_SIM2$CargaViral_48,VIH_SIM2$Grupo)
U_SIM3_CV <- test(VIH_SIM3$CargaViral_48,VIH_SIM3$Grupo)

# Alternativa 2 PMM
U_IM_CV <- test(VIH_IM$CargaViral_48,VIH_IM$Grupo)
U_IM2_CV <- test(VIH_IM2$CargaViral_48,VIH_IM2$Grupo)
U_IM3_CV <- test(VIH_IM3$CargaViral_48,VIH_IM3$Grupo)

# Alternativa 3 Midastouch
U_IMM_CV <- test(VIH_IMM$CargaViral_48,VIH_IMM$Grupo)
U_IMM2_CV <- test(VIH_IMM2$CargaViral_48,VIH_IMM2$Grupo)
U_IMM3_CV <- test(VIH_IMM3$CargaViral_48,VIH_IMM3$Grupo)

## ----tabla resultados-----------------------------------------------------

res_cd <- data.frame(t(
            data.frame("VIH_defecto" = U_SIM_CD, "VIH2_defecto" = U_SIM2_CD,
                       "VIH3_defecto" = U_SIM3_CD, "VIH_PMM" = U_IM_CD,
                       "VIH2_PMM" = U_IM2_CD, "VIH3_PMM" = U_IM3_CD,
                       "VIH_Midas" = U_IMM_CD, "VIH2_Midas" = U_IMM2_CD,
                       "VIH3_Midas" = U_IMM3_CD)))
```

```
res_cv <- data.frame(t(
            data.frame("VIH_defecto" = U_SIM_CV, "VIH2_defecto" = U_SIM2_CV,
                       "VIH3_defecto" = U_SIM3_CV, "VIH_PMM" = U_IM_CV,
                       "VIH2_PMM" = U_IM2_CV, "VIH3_PMM" = U_IM3_CV,
                       "VIH_Midas" = U_IMM_CV, "VIH2_Midas" = U_IMM2_CV,
                       "VIH3_Midas" = U_IMM3_CV)))

res <- cbind(res_cd,res_cv)
colnames(res) <- c("P_valor(CD4A)","Z(CD4A)","P_valor(log10 CargaViral)",
                   "Z(Log10 CargaViral)")
kable(res)
```