Universitat Oberta de Catalunya

uoc.edu

Estrategias para el abordaje de la problemática de los datos faltantes en ensayos clínicos longitudinales

Beatriz Pardo Montenegro
Máster de Bioinformática y Bioestadística
Área 2. Subárea 2: Análisis de datos

Nuria Pérez Álvarez
Carles Ventura Royo

12/2021

FICHA DEL TRABAJO FINAL

Título del trabajo:	*Estrategias para el abordaje de la problemática de los datos faltantes en ensayos clínicos longitudinales*
Nombre del autor:	*Beatriz Pardo Montenegro*
Nombre del consultor/a:	*Nuria Pérez Álvarez*
Nombre del PRA:	*Carles Ventura Royo*
Fecha de entrega (mm/aaaa):	12/2021
Titulación:	Máster de Bioinformática y Bioestadística
Área del Trabajo Final:	*TFM-Bioinformática y Bioestadística Área 2*
Idioma del trabajo:	*Castellano*
Palabras clave	*Datos faltantes, imputación, análisis supervivencia, ensayos longitudinales*

Resumen del Trabajo (máximo 250 palabras): *Con la finalidad, contexto de aplicación, metodología, resultados i conclusiones del trabajo.*

Los ensayos clínicos longitudinales se hacen con medidas repetidas de algunas variables a lo largo de un tiempo. Es un problema frecuente encontrarnos con que alguna de estas medidas falta. Desarrollar estrategias adecuadas en el tratamiento de datos faltantes supone un gran reto. Una de las alternativas recomendadas es la imputación múltiple. Los análisis de supervivencia analizan el tiempo que tarda en ocurrir un evento de interés, si éste no ocurre en el tiempo de seguimiento del ensayo, se denomina censura. Existen numerosas alternativas para la realización de análisis de supervivencia con datos censurados, el método convencional más utilizado es Kaplan Meier. En los últimos años se han desarrollado algoritmos de machine learning, uno de ellos es random survival forest. En este TFM, tras realizar labores de data management, se realiza imputación de datos faltantes de la base de datos del ensayo Lake con 2 opciones de imputación disponibles en R, en las librerías mice y randomForestSRC. Se comparan los resultados de la función de supervivencia de Kaplan Meier y del algoritmo random survival forest aplicados al resultado de ambas imputaciones. El evento estudiado es fracaso virológico en pacientes diagnosticados de VIH tratados con tratamiento experimental vs control. Los resultados del índice C son muy similares. Por Kaplan Meier, se concluye que el tratamiento experimental tiene menos fracasos virológicos que el tratamiento estándar, pero las diferencias son significativas sólo en la base de datos imputada con la librería randomForestSRC.

Abstract (in English, 250 words or less):

Longitudinal clinical trials are carried out by means of repeated measurements of certain variables in a period of time. A recurring issue is finding that some of these measurements are missing. Developing suitable strategies to manage missing data presents a challenge. One of the recommended alternatives is multiple imputation. Survival analyses examine how long it takes for an event to occur; if this event does not occur in the follow-up time it is referred to as censored. There are numerous alternatives to perform survival analyses with censored data; the most common conventional method is Kaplan Meier. In recent years machine learning algorithms have been developed, one of them being random survival forest. In this TFM, after carrying out data management tasks, the imputation of missing data from the Lake trial database is performed with two available methods in R: package mice and randomForestSRC. The results of the survival function of Kaplan Meier and the random survival forest algorithm are compared, applied to the result of both imputations. The event studied is virological failure in patients diagnosed with HIV and treated with experimental treatment vs control. The C-index results are very similar. With the Kaplan method, it is concluded that the experimental treatment has fewer virological failures than the standard treatment, but the differences are only significant in the database imputed with randomForestSRC package.

Índice

Lista de figuras

1. Introducción

1.1 Contexto y justificación del Trabajo

Los ensayos clínicos longitudinales se hacen con medidas repetidas de algunas variables a lo largo de un tiempo. Es un problema muy frecuente encontrarnos con que alguna de estas medidas falta. Se suele considerar que, si estos datos faltantes suponen menos de un 5% de los datos, excluirlos no supondrá un sesgo en los resultados del ensayo.

Pero en muchos de los ensayos clínicos que se llevan a cabo, el porcentaje de estos datos faltantes es elevado, con lo que ser capaz de tratarlos adecuadamente, evitando el impacto en los resultados de los estudios y evitando que estos datos faltantes reduzcan la representatividad de la muestra obtenida, es una temática que supone un gran reto para la comunidad científica.

En los análisis de supervivencia se analiza el tiempo que tarda en ocurrir un evento de interés. Cuando este evento no ocurre, se denomina censura, puede pasar que no ocurra porque el evento se convierte en inobservable o bien porque no se experimenta en el tiempo que se fijó de observación. Actualmente se están desarrollando muchos algoritmos de aprendizaje automático para tratar con datos censurados, la mayoría de estos algoritmos no pueden utilizarse con datos faltantes, por ello para poder utilizar este tipo de herramientas también es fundamental tratar adecuadamente los datos faltantes. [5]

Dentro de las áreas disponibles para la realización del TFM, éste pertenece a la subárea 2 cuya temática es de análisis de datos. Dentro de esta subárea, la línea elegida es la de generación automática de un informe de resultados para un ensayo clínico con medidas repetidas en el tiempo.

Los ensayos clínicos son la base de la investigación clínica que sirve para la identificación de estrategias diagnósticas y terapéuticas que deriven en mayor eficacia y seguridad. Un ensayo clínico es un estudio que se lleva a cabo en personas para evaluar la eficacia y la seguridad de un tratamiento, a través de ellos se busca descubrir nuevas formas de diagnosticar, prevenir, tratar y entender las enfermedades que afectan a los seres humanos. [6]

Para que un nuevo medicamento pueda introducirse en el mercado, debe haber demostrado su eficacia y su seguridad en los ensayos clínicos. Éstos siempre se realizan siguiendo un protocolo de investigación estrictamente controlado y no pueden empezar sin haber sido aprobados por un comité ético independiente de los investigadores y los promotores del estudio.

Al comenzar el diseño de un ensayo clínico se definen las variables que se necesitan recoger para luego poder estudiar los objetivos que se recojan en el protocolo. Todos los registros de estas variables serán los que formen la base de datos del ensayo.

Un correcto manejo de las bases de datos, así como conocer los posibles problemas que pueden surgir es un tema de gran interés científico y es el motivo de que esta temática haya sido la elegida para este TFM.

Este TFM abordará dos problemáticas que pueden surgir en los datos de los ensayos clínicos, la primera de ellas: los datos faltantes. Se explorarán diferentes métodos de imputación disponibles que pretenden subsanar los problemas derivados de un porcentaje elevado de datos faltantes. La segunda problemática es la de la censura. En los análisis de supervivencia, se analiza el tiempo que tarda en ocurrir un evento de interés. Cuando este evento no ocurre, se denomina censura, puede pasar que no ocurra porque el evento se convierte en inobservable o bien porque no se experimenta en el tiempo que se fijó de observación. Actualmente se están desarrollando muchos algoritmos de aprendizaje automático para tratar con datos censurados, así que en este TFM se elegirá uno de ellos y se verá el impacto que tiene en sus resultados haber imputado los datos faltantes con una u otra de las herramientas que hay disponibles para hacerlo en el software estadístico R. También se realizará el análisis de supervivencia por un método convencional, Kaplan Meier, y se analizará si existen diferencias en los resultados de ambas imputaciones.

1.2 Objetivos del Trabajo

Se establecen dos objetivos generales y dentro de cada uno se establecen varios objetivos específicos.

1.- Realizar labores de data management con la base de datos facilitada para el TFM con el software estadístico R. Revisar variables de las que se disponen, limpieza de los datos que no se van a analizar, focalizar en las variables de interés.

1.1.- Aprendizaje y búsqueda de los métodos de análisis.

1.2.- Conocer el diseño y objetivos del ensayo Lake.

1.3.- Realizar un análisis descriptivo de los datos disponibles en la base de datos del TFM. Número de sujetos y evaluación de las variables recogidas.

1.4.- Conocer la distribución de pacientes de la base de datos en ambas ramas de tratamiento, según algunas de las variables demográficas recogidas.

1.5.- Evaluar el porcentaje de datos perdidos y analizar su distribución por variables.

1.6.- Definir evento en el TMF y realizar labores de limpieza de la base de datos focalizando en las variables de interés.

2.- Con el software estadístico R ser capaz de implementar dos de las opciones de imputación disponibles. Ver si se aprecian diferencias en el modelo de Kaplan Meier. Aplicar un mismo algoritmo de aprendizaje automático para realizar análisis de supervivencia con datos censurados, a cada una de las dos bases de datos que se obtendrán tras las dos imputaciones realizadas con herramientas diferentes y ser capaz de hacer una correcta evaluación de los resultados.

2.1.- Aplicar imputación de datos con la librería mice.

2.2.- Aplicar imputación de datos con la librería randomForestSRC.

2.3.- Realizar modelo de Kaplan Meier con ambas bases de datos obtenidas de las dos imputaciones y ver si hay diferencias.

2.4.- Cálculo de la función de supervivencia y del log rank para ambos modelos de Kaplan Meier obtenidos tras las dos imputaciones.

2.5.- Aplicar un mismo algoritmo de aprendizaje automático, random survival forest, para realizar análisis de supervivencia con datos censurados, a las dos bases de datos que se obtendrán tras las dos imputaciones realizadas.

2.6.- Evaluación de la capacidad de discriminación del algoritmo mediante índice C y ver si existen diferencias imputando los datos con una u otra herramienta.

1.3 Enfoque y método seguido

En este TFM se parte de la base de datos de un ensayo de VIH, el ensayo Lake, en él se compara la eficacia de dos tratamientos, es un ensayo longitudinal, ya que se recogen las mismas variables a lo largo del tiempo. El principal problema que tiene esta base de datos es que hay un 41% de datos faltantes. Tras leer sobre análisis de datos, sobre el manejo de datos faltantes y sobre análisis de supervivencia, el enfoque de este TFM es comparar dos funciones de imputación múltiple disponibles en el software estadístico R. Dicha comparación se realiza con los resultados obtenidos tras implementar dos métodos de análisis de datos para el análisis de supervivencia, uno convencional no paramétrico, como es Kaplan Meier y un algoritmo de machine learning utilizado para análisis de supervivencia, el random survival forest.

El método seguido para realizarlo fue el siguiente, en primer lugar, se hizo un análisis descriptivo de las variables disponibles y de los datos faltantes. Tras ello, se hicieron labores de data management seleccionando las variables de interés. Se definió el evento que se estudiaría en los análisis de supervivencia, el evento es fracaso virológico. Se imputó la base de datos tras la eliminación de algunas variables con dos funciones de imputación múltiple, disponibles en las librerías mice y randomForestSRC. Se realizó el método de análisis de

supervivencia de Kaplan Meier y el algoritmo de supervivencia de random survival forest al resultado de ambas imputaciones y se compararon los resultados.

1.4 Planificación del Trabajo

Para la realización del plan de trabajo del TFM se utiliza el programa GanttProject 3.1.3102. Se realiza un diagrama de Gantt para definir el plan de trabajo. Un diagrama de Gantt es una herramienta muy popular dentro de la gestión de proyectos, es utilizada para exhibir diversas actividades y objetivos claros con el tiempo previsto para su terminación, así como el tiempo exacto de duración real. Su ventaja principal es que utiliza un enfoque visual para mostrar claramente, las tareas programadas asignándoles a cada una de dichas tareas el tiempo necesario para su ejecución. [7]

Lo primero que debe de realizarse es la identificación de tareas, asignándole a cada una un tiempo de realización. Habrá tareas que puedan realizarse paralelamente unas a las otras, sin embargo, hay tareas que para iniciarse tienen que haber sido realizadas otras previamente. Es muy importante tener claro en el momento de la planificación qué tareas deben de realizarse para poder empezar otras, ya que hay algunas que tienen holgura (tiempo que puede prolongarse una actividad sin que retrase el proyecto) y otras tareas si se retrasan producen el retraso del proyecto.

El camino crítico marca la duración de un proyecto. Viene determinado por aquellas actividades que tienen que terminarse para que comience la siguiente actividad, es decir aquellas que no tienen flexibilidad dentro del calendario. La suma de todas estas actividades marcará la duración de un proyecto.

Se definen cada una de las tareas necesarias para la consecución de cada uno de los objetivos generales y específicos. Se asigna un tiempo de duración para la ejecución de cada tarea y se establecen qué tareas se deben completar antes de del inicio de otras, antecesoras, y aquellas que se pueden realizar paralelamente.

Se establece la fecha de inicio del proyecto. El 15 de septiembre de 2021, fecha en la que el aula del TFM estuvo disponible en la web de la UOC. Se configuran los días de fin de semana como sábado y domingo. Se establece que todas las tareas se ejecutarán tanto de lunes a viernes como durante el fin de semana. Se elige un calendario de días festivos personalizado estableciendo fechas que de antemano se conoce que no se podrán realizar tareas del TFM para que sean excluidas en la planificación del TFM.

Los principales hitos en la planificación de este TFM son la entrega de cada una de las PECs que se estable en el aula de la asignatura. Para lograr cada uno de estos hitos se tendrán en cuenta las tareas necesarias para cada entrega. Como se comentó anteriormente, una de las principales ventajas de los diagramas de Gantt es que son un método muy visual para la planificación de los proyectos así que se asignará a cada hito, en este caso, la entrega de cada PEC, un color y las tareas que necesitan terminarse antes de cada entrega estarán en ese color. Hay tareas que su ejecución durará más de una PEC, a estas tareas se les asignará el color de la PEC en la que deben

finalizar. En el programa utilizado para la temporalización de este plan de trabajo, GanttProject, cada hito aparecerá como un rombo.

Como la realización de cualquier proyecto, la realización del TFM está expuesta a diferentes riesgos. Analizarlos de antemano y pensar posibles soluciones para mitigarlos en el caso de que ocurran es una parte fundamental de cualquier planificación. Para evaluar más profundamente los riesgos de un proyecto, en este caso, la realización del TFM, se debe evaluar el impacto, es decir, el efecto que tendrán en el caso de producirse y la probabilidad de que ocurran. En función de ambas cosas se asignará la seriedad del riesgo y según esto, la frecuencia con la que debe ser revisado el riesgo.

Durante la ejecución del TFM se fueron entregando algunos informes del desarrollo del trabajo, en cada uno de ellos, se revisó y adaptó el Gantt a y las tareas a los problemas que fueron surgiendo.

A continuación, se recogen las tareas y el diagrama de Gantt del TFM.

Hitos y tareas	Fecha de inicio	Fecha de fin
PEC0 - Definición de los contenidos del trabajo	**HITO 1**	
Lectura protocolo ensayo Lake	15/09/2021	16/09/2021
Lectura sobre VIH y fármacos antirretrovirales	16/09/2021	16/09/2021
Lectura trabajo Marcella Marinelli	17/09/2021	18/09/2021
Búsqueda y lectura de artículos científicos sobre imputación	18/09/2021	21/10/2021
Búsqueda y lectura de artículos científicos sobre algoritmos de machine learning	18/09/2021	18/11/2021
Redacción PEC0	17/09/2021	22/09/2021
PEC1 - Plan de trabajo	**HITO 2**	
Elección de los objetivos del TFM	23/09/2021	25/09/2021
Definición de las tareas del TFM	26/09/2021	27/09/2021
Realización diagrama Gant	26/09/2021	30/09/2021
Análisis de riesgos	01/10/2021	01/10/2021
Redacción PEC1	25/09/2021	04/10/2021
PEC2- Desarrollo del trabajo - Fase 1	**HITO 3**	
Aprendizaje y búsqueda de los métodos de análisis	05/10/2021	15/11/2021
Análisis descriptivo de la base de datos	05/10/2021	18/10/2021
Definicón de evento	18/10/2021	20/10/2021
Data management. Creación base de datos con variables interés	21/10/2021	25/10/2021
Imputación mediante MICE	31/10/2021	04/11/2021
Redacción PEC2	06/10/2021	08/11/2021
PEC3- Desarrollo del trabajo - Fase 2	**HITO 4**	
Imputación mediante RandomForestSRC	09/11/2021	12/11/2021
Aplicación modelo Kaplan Meier a ambas bases de datos	13/11/2021	14/11/2021
Comparación resultados del modelo Kaplan Meier	15/11/2021	16/11/2021
Aplicación algoritmo machine learning sobre dataset MICE	16/11/2021	19/11/2021
Aplicación algoritmo machine learning sobre dataset RandomForestSRC	20/11/2021	24/11/2021
Evaluación de resultados mediante índice C	25/11/2021	08/12/2021
Redacción PEC3	10/11/2021	09/12/2021
PEC4- Cierre de la memoria	**HITO 5**	
Redacción y cierre de la memoria	01/10/2021	23/12/2021
PEC5a- Elaboración de la presentación	**HITO 6**	
Redacción PEC5. Presentación TFM	26/12/2021	30/12/2021
PEC5b- Defensa pública	21/01/2022	21/01/2022
Defensa pública de forma síncrona	13/01/2022	21/01/2022

Ilustración 1: Tareas de la planificación del TFM 1

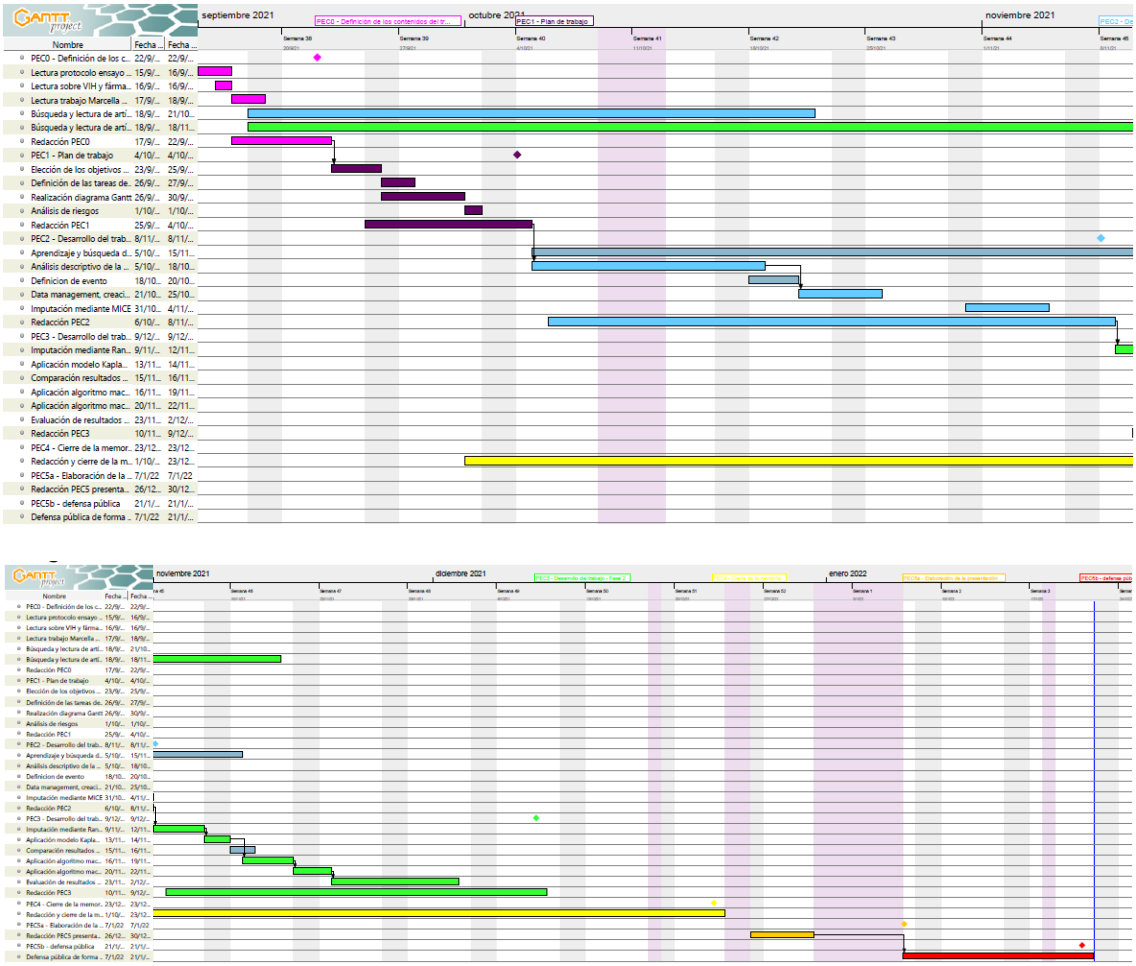

Ilustración 2: Diagrama de Gantt

1.5 Breve sumario de productos obtenidos

Los productos obtenidos de la realización de este TFM son:

- Memoria del trabajo. En dicha memoria se presenta el resultado del trabajo, justificando su interés, sus objetivos, la metodología seguida para conseguirlo, así como sus resultados.
- Planificación del trabajo. Diagrama de Gantt que incluye todas las tareas necesarias para la realización del TFM así como una visualización del diagrama.
- Anexo código R del trabajo. Se recoge todo el código de R necesario para la realización de este TFM.
- Presentación y defensa virtual del proyecto. La presentación incluye un archivo en powerpoint y un video donde está grabada dicha presentación. La defensa se realiza de forma síncrona por videollamada.

1.6 Breve descripción de los otros capítulos de la memoria

Además de este capítulo de introducción, la memoria del TFM recoge los siguientes capítulos:

- Capítulo 2: Marco conceptual. En él se recoge una revisión de conceptos necesarios para poder realizar el TFM, se habla de ensayos clínicos, del VIH, del ensayo Lake, de la problemática de los datos faltantes y de los análisis de supervivencia.

- Capítulo 3: Análisis de la base de datos. Se hace un análisis descriptivo de la base de datos de la que se parte y se analiza la distribución de los datos faltantes.

- Capítulo 4: Transformación de la base de datos. Se describe proceso de limpieza de la base de datos, la definición de evento y se describe la base de datos resultante de las labores de data management realizadas.

- Capítulo 5: Imputación de datos faltantes. Se habla de las dos opciones de imputación elegidas.

- Capítulo 6: Análisis de supervivencia Kaplan Meier. Se realizan las curvas de supervivencia de las dos bases de datos resultantes de ambas imputaciones. Se calcula el log rank para ambas y se analizan los resultados.

- Capítulo 7: Análisis de supervivencia Random Survival Forest. Se aplica el algoritmo de machine learning para el análisis de supervivencia de random forest a las dos bases de datos después de haber hecho la imputación y se comparan los resultados con el índice C.

- Capítulo 8: Conclusiones. Se describen las conclusiones obtenidas, el aprendizaje, la consecución de los objetivos, el ajuste a la temporización prevista y nuevas líneas a explorar si hubiese más tiempo.

- Capítulo 9: Glosario. Listado de acrónimos utilizados en el TFM.

- Capítulo 10: Bibliografía.

- Capítulo 11: Anexos.

2. Marco conceptual

2.1 Ensayos clínicos

Los ensayos clínicos son investigaciones en las que se incluyen personas para analizar la seguridad y la eficacia de productos, técnicas, medicamentos u otras novedades médicas o sanitarias que permitan la mejora de la salud de la población. [12]

Antes de llegar a las investigaciones con humanos en ensayo clínico hay que pasar por fases previas que garanticen la seguridad de lo que va a analizarse en el ensayo con personas. Por ello, de manera previa se hacen ensayos in vitro (en laboratorio, sobre células, tejidos u órganos) e in vivo sobre animales no humanos. [12]

Los ensayos clínicos tienen varias fases: [13] [14]

- Fase I: en esta fase se analiza principalmente la seguridad de la intervención que se esté estudiando. Se realiza sobre una población sana muy pequeña, en general menos de 100 personas. Se prueban la seguridad, los efectos secundarios, dosis, vía de administración y el momento adecuado de administrar el tratamiento nuevo.

- Fase II: en esta fase se analiza la eficacia de la intervención y se recoge más datos sobre la seguridad. En la fase 2 del ensayo clínico los pacientes reciben la dosis más alta del tratamiento que no causó efectos secundarios perjudiciales en la fase 1 del ensayo clínico. Se realiza con un grupo reducido pacientes que padezcan la enfermedad, entre 100 y 300.

- Fase III: En esta fase se analizan la eficacia y seguridad de la intervención en las condiciones similares a las que se encontraran cuando se comercialice. Se realizan con una muestra de pacientes más amplia que en la fase anterior (entre 300 y 3000) y representativa de la población general a la que vaya destinada la intervención. Además, la intervención estudiada se compara con el tratamiento estándar utilizado habitualmente en la práctica clínica. Si no existiera un tratamiento habitual se compararía con placebo u otras terapias. Estos estudios constituyen el soporte para conseguir la autorización y comercialización por parte de las agencias reguladoras.

- Fase IV: Estos estudios, también denominados estudios post-comercialización, analizan los efectos a largo plazo del fármaco comercializados. También se pueden estudiar nuevas indicaciones de la intervención, nuevas formulaciones, formas de dosificación etc. Por ejemplo, para la comercialización de un fármaco nuevo se necesita pasar por todas las fases, desde los estudios preclínicos hasta la fase III,

se estima que son necesarios 10 años para completar todo este proceso.

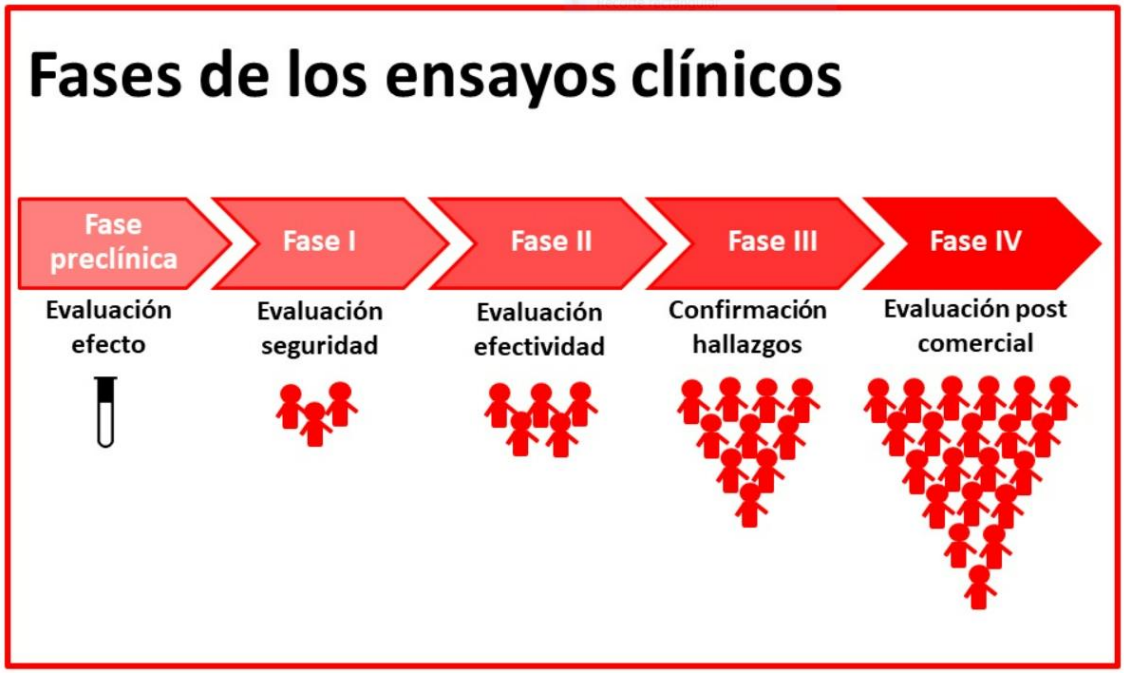

Ilustración 3: Fases de los ensayos clínicos (Fuente web Instituto Salud Carlos III)

Los estudios clínicos evalúan una presunta relación causal entre un factor y un efecto, respuesta o resultado. Según un criterio analítico, podemos dividir a los estudios en: [13]

- Experimentales: los investigadores controlan las condiciones bajo las cuales se realizará la investigación, seleccionan el tipo de paciente, que intervención se va a realizar y durante cuánto tiempo. Además, se realiza un seguimiento de los pacientes durante un tiempo determinado para evaluar el efecto de la intervención. Dentro de estos estudios se encuentran los ensayos clínicos aleatorizados y los ensayos comunitarios.

- Observacionales el equipo investigador no controla el factor de estudio; se limitan a observar, medir y analizar. Entre estos estudios se encuentran los estudios de cohortes, casos controles, transversales, ecológicos o estudios de caso o series de caso.

Los ensayos clínicos aleatorizados (ECAs) son considerados los estudios más sólidos. Son estudios controlados donde los participantes se asignan al azar a un grupo donde reciben el tratamiento o intervención que se quiere estudiar o a un grupo de comparación o control, donde reciben el tratamiento o intervención que se utiliza habitualmente o un placebo. [13]

Los ensayos clínicos se caracterizan por los siguientes componentes básicos:
• Temporalidad: concurrente y prospectivo, es decir se siguen a los participantes desde que reciben la intervención hacia el futuro.
• Aleatorización: asignación al azar de la intervención.

• Variable dependiente (respuesta) debe ser única, simple, fácil de medir, consistente, relevante para la práctica y que permita evaluar objetivamente el resultado del estudio.
• Sujetos de estudio: personas
• Evaluación objetiva del efecto de la intervención.
• Grupo de comparación (control): para contrastar hipótesis.

Hay que tener en cuenta una serie de parámetros para realizar correctamente un ECA y disminuir los sesgos que pueden llevar a obtener conclusiones incorrectas. Entre los parámetros más importantes destacan la selección de la población, aleatorización, enmascaramiento y seguimiento. [15]

En los estudios se debe realizar un seguimiento de los pacientes, que sea lo suficientemente prolongado, para poder identificar cambios que se produzcan tras la exposición a la intervención, cambios en los factores pronósticos o eventos de interés, como efectos secundarios. El seguimiento debe ser el mismo en los dos grupos con idénticas intervenciones en ambos grupos, evitando así los posibles sesgos que se pudieran producir por tener un mayor control en unos de los grupos. [15]

Si clasificamos a los ensayos según su dimensión temporal, tenemos a los ensayos longitudinales que se caracterizan por tener más de dos mediciones a lo largo de un seguimiento, implican mediciones repetidas en los sujetos a lo largo del tiempo. Se debe de garantizar que todas las mediciones se realicen en el momento oportuno y con técnicas normalizadas. [16]

Los problemas más frecuentes encontrados en los ensayos longitudinales son que la larga duración de algunos estudios obliga a prestar una atención especial al cambio de personal, al deterioro de los equipos, al cambio de tecnologías. Otro problema importante es que existe una mayor probabilidad de abandono durante el seguimiento. En tercer lugar, otro problema es la existencia de datos perdidos. Si se requiere que un participante tenga todas las mediciones hechas, puede producir un problema similar al de los abandonos durante el seguimiento.[16]

2.2 Análisis de supervivencia

Los datos proporcionados por los estudios clínicos se expresan en múltiples ocasiones en términos de supervivencia. Esta medida no queda limitada a los términos de vida o muerte, el término supervivencia se debe a que en las primeras aplicaciones de este método de análisis se utilizaba como evento la muerte de un paciente, pero en realidad sirven para analizar todas aquellas situaciones en las que se mide el tiempo que transcurre hasta que sucede un evento de interés, como puede ser, tiempo de recurrencia, tiempo que dura la eficacia de una intervención, tiempo de un aprendizaje determinado, etc. Por tanto, la supervivencia es una medida de tiempo a una respuesta, fallo, muerte, recaída o desarrollo de una determinada enfermedad o evento. [19]

El análisis de supervivencia es una parte de la estadística donde el objetivo es analizar y modelar los datos donde el resultado es el tiempo hasta la ocurrencia de un evento de interés. Uno de los principales desafíos en este contexto es la presencia de instancias cuyos resultados de eventos se vuelven inobservables después de un cierto punto de tiempo o cuando algunas instancias no experimentan ningún evento durante el período de seguimiento. Tal fenómeno se llama censura. Esta censura se puede manejar de manera eficaz utilizando técnicas de análisis de supervivencia. Los algoritmos de aprendizaje están adaptados para manejar de manera efectiva los datos de supervivencia. [5]

El objetivo principal es obtener una estimación lo mejor posible del momento en que ocurre un evento de interés. Uno de los principales retos es la existencia de instancias censuradas, es decir, el evento de interés no se observa debido a la limitación de tiempo del período de estudio o pérdidas de seguimiento durante el período de observación. [5]

El análisis de supervivencia proporciona varios mecanismos para manejar los problemas de datos censurados que surgen al modelar los datos de tiempo transcurrido hasta el evento. Se han desarrollado nuevos algoritmos computacionales para manejar con eficacia un desafío tan complejo como es el de modelar este tipo de datos. Estos métodos de aprendizaje automático para resolver problemas de análisis de supervivencia complementan a los tradicionales (o convencionales) métodos de estadística. [5]

Existen una serie de técnicas estadísticas, análisis de la supervivencia, apropiadas para estudios en los que cada paciente es seguido durante un determinado período y en los que se recoge el intervalo que transcurre entre el hecho inicial y el hecho final, o hasta que acaba el seguimiento si no ocurre el hecho final. Además, entre estas técnicas, se disponen de pruebas para comparar curvas de supervivencia, y modelos más complejos basados en la regresión que permiten valorar el efecto de un conjunto de valores pronósticos. [20]

El análisis de la supervivencia es una técnica muy apropiada para analizar respuestas binarias, aparición o no aparición del evento, en estudios longitudinales o de seguimiento que se caractericen por:

- Duración variable del seguimiento: los estudios de seguimiento tienen fechas muy bien definidas de inicio y de cierre, pero los sujetos se incorporan al estudio en momentos diferentes.
- Observaciones incompletas: en la fecha de cierre del estudio aún no se ha producido el evento en ciertos sujetos, sujetos censurados. Además, puede haber pérdidas de seguimientos. Estas observaciones incompletas dan lugar a lo que se llama “datos censurados”, y el análisis de supervivencia se caracteriza por incluir la información que aportan estos datos. [20]

En este tipo de análisis se manejan fundamentalmente dos variables, el tiempo de seguimiento, que es el tiempo que transcurre entre la fecha de entrada en el estudio hasta la fecha registrada en la última observación. Y la aparición o no

aparición del evento que se estudia. Si el evento se produce, el tiempo de supervivencia es el tiempo transcurrido desde el inicio hasta que se produce el evento, se trata de un tiempo completo, o no censurado. Si el evento no se produce, se mide tiempos incompletos o censurados, es el tiempo desde el inicio hasta la última medición. Los tiempos censurados pueden tener varios orígenes, pueden ser sujetos que cuando se termina un estudio, en el momento del cierre del estudio no han presentado el evento. O bien pueden ser sujetos perdidos, sujetos que, por alguna causa, no han completado el tiempo de seguimiento y que cuando se dejó de medir, no habían presentado el evento. Ambos producen tiempos incompletos censurados por la derecha que son analizados de manera idéntica, pero son muy diferentes. El tiempo de supervivencia de los sujetos que no han presentado el evento es desconocido. [20]

Los datos de supervivencia aportan dos tipos de probabilidades diferentes: supervivencia y riesgo. La probabilidad de supervivencia o función de supervivencia o S(t) es la probabilidad de que un individuo sobreviva (no presente el evento) desde la fecha de entrada en el estudio hasta un momento determinado en el tiempo. Estos valores van a describir la supervivencia global de toda nuestra población.

$$S(t) = P(T > t) = 1 - F(t)$$

Ilustración 4: Función de supervivencia

La función de riesgo o h(t) es la probabilidad de que a un individuo que está siendo observado, en el tiempo t, le suceda el evento en ese momento. Es el cociente entre la función de densidad y la función de supervivencia. [22]

$$h(t) = \frac{f(t)}{S(t)}$$

Ilustración 5: Función de riesgo

Hay que destacar las diferencias entre ambas. Mientras que la función de supervivencia se centra sobre todo en la no ocurrencia del evento, la función de riesgo se centra en la ocurrencia del evento. [20]

Los métodos para el análisis de datos de supervivencia son válidos sólo si los pacientes censurados, aquellos en los que no se ha presentado el evento al finalizar el seguimiento, han sido censurados por un motivo no informativo. Un paciente perdido puede presentar el evento en cualquier momento una vez lo hemos perdido. Para los análisis de supervivencia asumimos que los pacientes perdidos no tienen ninguna característica que lo haga diferente de los que permanecen en el estudio y no se han perdido y, por tanto, es comparable a los que se quedan en el estudio. [20]

Actualmente, existen una gran variedad de métodos para realizar análisis de supervivencia.

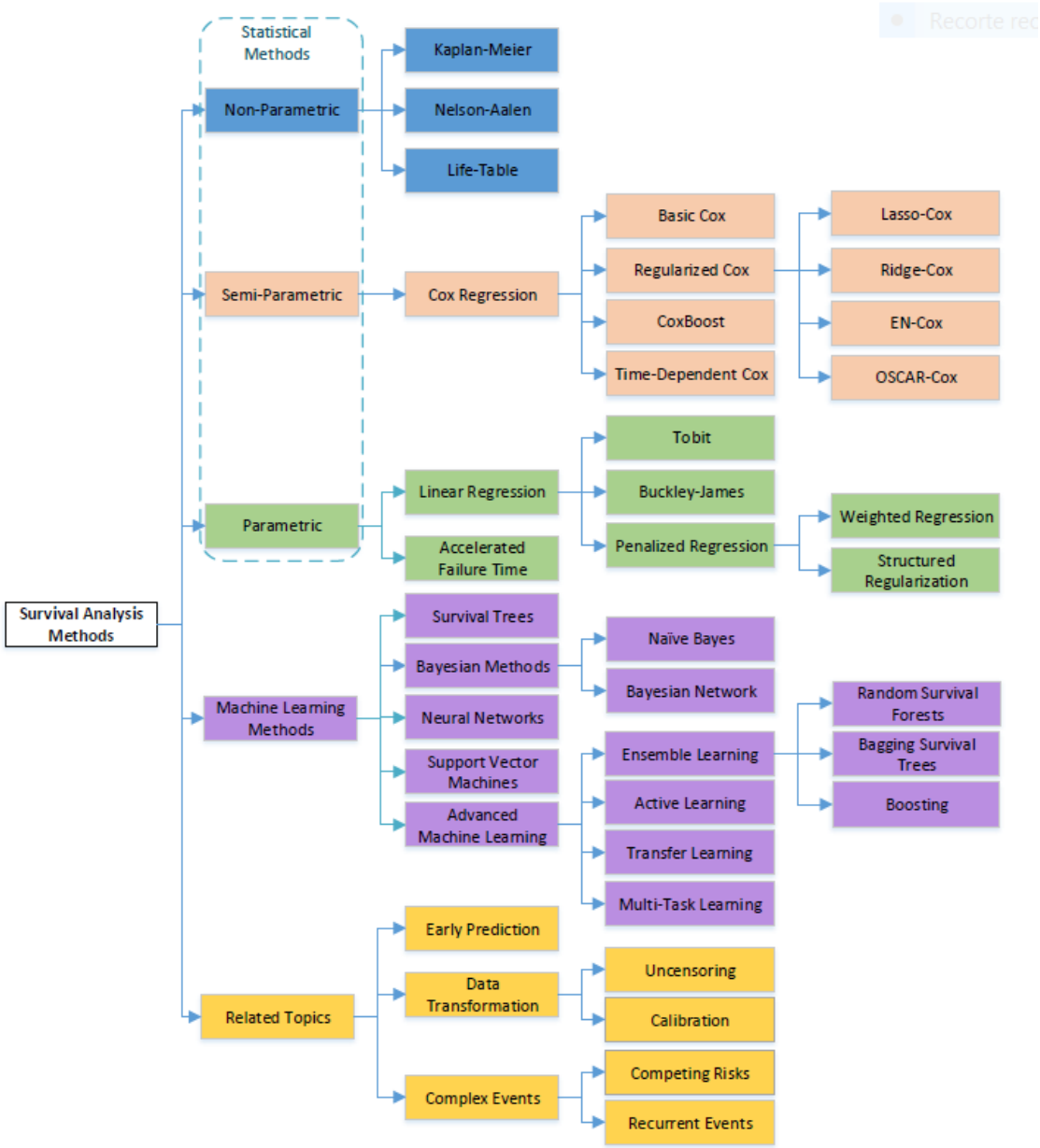

Ilustración 6: Métodos de análisis de supervivencia (Fuente [5])

En el método estadístico no paramétrico de Kaplan Meier, se calcula la supervivencia cada vez que un sujeto sufre el evento de estudio generando la probabilidad de cada momento. Las probabilidades de supervivencia se calculan a partir del número de pacientes en riesgo justo en el momento de producirse el evento y nos da la probabilidad de que en un determinado momento no se haya producido el evento. Al inicio, la probabilidad de que no haya ocurrido el evento es del 100%. La incidencia acumulada mide la probabilidad de presentar el evento antes de un determinado momento. Al inicio, la incidencia acumulada es 0 y va aumentado a medida que transcurre el

tiempo. Con este método también obtenemos valores del riesgo, la tasa de riesgo mide el riesgo de que ocurra el evento habiendo llegado sin el evento a un determinado tiempo.

Habitualmente con las curvas de Kaplan Meier se comparan dos tratamientos, se parte de la hipótesis nula de que el evento tiene la misma probabilidad de producirse en ambos brazos de tratamientos durante todo el seguimiento y se calcula el test log rank, si el valor del p valor log rank es menor de 0,05 se considera que se puede descartar la hipótesis de la igualdad de probabilidad de que ocurra el evento entre ambos tratamientos. Es un test de significancia pura, no mide la magnitud de las diferencias entre grupos.

En los últimos años, las técnicas de aprendizaje automático, machine learning, han avanzado mucho y sus ventajas para modelar y predecir se están incorporando en muchas de las áreas del análisis de datos. El aprendizaje automático consiste básicamente en automatizar, mediante distintos algoritmos, la identificación de patrones o tendencias que se esconden en los datos. El objetivo del machine learning es crear un modelo que nos permita resolver una tarea dada. Luego se entrena el modelo usando gran cantidad de datos. El modelo aprende de estos datos y es capaz de hacer predicciones. Según la tarea que se quiera realizar, será más adecuado trabajar con un algoritmo u otro.

En el análisis de supervivencia, el principal desafío de los métodos de aprendizaje automático es la dificultad para tratar adecuadamente la información censurada y la estimación del tiempo del modelo. En los últimos años, se han desarrollado métodos de aprendizaje automático más avanzados. para tratar y predecir a partir de datos censurados. Dentro de los métodos de aprendizaje por conjuntos, nos encontramos con random survival forests (RSF), que es una extensión realizada por Ishwaran et al. 2008 del random forest (RF) desarrollado por Breiman 2001, RF es un método de conjunto propuesto específicamente para hacer predicciones utilizando los modelos estructurados en árbol. [5]

Debido a la presencia de censura en los datos de supervivencia, las métricas de evaluación estándar para la regresión no son adecuadas para medir el rendimiento en el análisis de supervivencia. El rendimiento de la predicción en el análisis de supervivencia debe medirse utilizando métricas de evaluación más especializadas. Una de las más populares es el índice C, desarrollado por Harrell, muy útil en problemas de predicción binaria, el índice C tendrá un significado similar al área bajo la curva AUC. Hay varias formas de calcular el índice C, para evaluar el desempeño durante un período de seguimiento, se define el índice C para un período de seguimiento fijo (0; t) como el promedio ponderado de los valores de AUC en todos los puntos de tiempo de observación posibles, por lo tanto, ct es la probabilidad de que las predicciones sean concordantes con sus resultados para un dato dado durante el período de tiempo (0; t). El objetivo principal del análisis de supervivencia es predecir la ocurrencia de eventos específicos de interés en puntos de tiempo futuros. Mide la capacidad de discriminación, es decir, la capacidad del modelo de distinguir entre los que desarrollarán el evento y los que no lo harán. Mide cómo de

eficaz es el modelo de predicción discriminando a los individuos en los que ocurre el evento frente a los individuos en los que no ocurre el evento. En presencia de datos de supervivencia, el índice C es una generalización del AUC. [5]

2.3 Datos faltantes

Los ensayos clínicos longitudinales se hacen con medidas repetidas de algunas variables a lo largo de un tiempo. Es un problema muy frecuente encontrarnos con que alguna de estas medidas falta.

Los datos faltantes se pueden clasificar en distintos tipos, según el mecanismo de pérdida de datos:

- MCAR (Missing Completely at Random) la causa de este tipo de datos faltantes es totalmente al azar, es por lo tanto independiente de la medicación utilizada en el ensayo y se debe por ejemplo a un cambio de domicilio, a la aparición de otra enfermedad, a un paciente poco cooperativo. Este tipo de mecanismo de generación de datos faltantes fue definido por Little y Rubin (1976), la probabilidad de tener datos faltantes es la misma para los individuos de diferentes grupos de tratamiento y aquellos que tienen diferente severidad en la enfermedad o tratamiento respuesta. [4]. La probabilidad de que una respuesta a una variable sea dato faltante es independiente tanto del valor de esta variable como del valor de otras variables del conjunto de datos. [21]
- MAR (Missing at Random) la causa de este tipo de datos faltantes puede estar relacionado con la medicación, con lo que su existencia puede provocar una falta de medición de algún efecto de la medicación, como podría ser que se deba a efectos adversos. La probabilidad de los datos perdidos se relaciona a variables observadas, pero no a variables no observadas. [4] La probabilidad de que una respuesta sea dato faltante es independiente de los valores de la misma variable, pero es dependiente de los valores de otras variables del conjunto de datos. [21]
- MNAR (Missing not at random), cuando la probabilidad de los datos faltantes depende de datos no observados. El motivo de la pérdida de estos datos se debe precisamente a los datos. Por ejemplo, en ensayos de sustancias de abuso en los cuales la abstinencia es uno de los resultados, es común que el abandono sea mayor en aquellos que han recaído. El problema es que en aquellos que abandonan, la recaída no queda registrada. En este caso, la probabilidad de datos faltantes depende de datos no observados (en este caso la recaída no observada). [4]. La probabilidad de que una respuesta a una variable sea dato faltante es dependiente de los valores de la variable. [21]

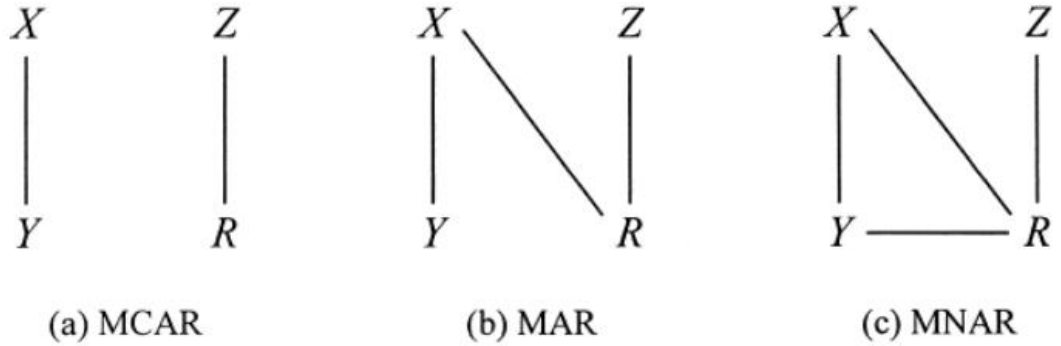

Ilustración 7: Representaciones gráficas patrón datos fatantes (Schafer and Graham [27])

Representaciones gráficas adoptadas por Schafer and Graham de (a) faltan completamente al azar (MCAR), (b) faltan al azar (MAR) y (c) faltan no al azar (MNAR) en un patrón univariante de datos perdidos. X representa variables que se observan por completo, Y representa una variable que falta parcialmente, Z representa el componente de las causas de falta no relacionadas con X e Y, y R representa la falta. [27]

Según el patrón de aparición, los datos faltantes pueden ser intermitentes o monótonos. En los intermitentes hay varias medidas con datos, luego hay datos faltantes en algunas de sus mediciones y vuelve a haber mediciones posteriores. Los monótonos, hay mediciones y a partir de un determinado momento ya no hay registradas más.

En un análisis de las posibles causas de datos faltantes en ensayos clínicos de VIH se vieron que las más comunes son: pérdida de seguimiento, incumplimiento, fallo virológico definido en protocolo, efecto adverso. La naturaleza de los datos faltantes es importante para determinar si la información que falta aportaría información sobre la medicación estudiada o no. [2].

Existen distintas estrategias para tratar con bases de datos con datos faltantes, una de ellas es excluir los datos faltantes y analizar exclusivamente los datos existentes. Podemos eliminar los casos que contienen datos faltantes o bien eliminar sólo los datos faltantes. Este enfoque podría reducir drásticamente el tamaño de la muestra y generar sesgos, sobre todo cuando el número de datos faltantes en los grupos de tratamiento no está equilibrado. [2]

El resto de las estrategias disponibles se basan en sustituir los valores faltantes por un valor, existen muchas opciones de completar estos datos faltantes, una puede ser sustituir los datos faltantes utilizando el último valor observado, independientemente de en qué momento haya sido ese último valor registrado. Algunos autores consideran que este método, en ensayos de VIH con estado avanzado de infección, en el que los pacientes van empeorando, este método puede dar resultados más optimistas que la realidad. La media y la varianza tendrán también sesgo con este método. [2]

Otra opción que existe en la literatura es la que consiste en reemplazar los datos faltantes por valores tomados de otro paciente con características similares, en caso de VIH podemos pensar en el valor del nivel de ARN del

VIH-1 que es el más cercano al valor que tenía el paciente en el momento de abandonar el estudio. [2]

En el caso de estudiar fracaso virológico, otra alternativa posible es sustituir los datos faltantes por causas negativas por el peor valor posible. Otra alternativa en el caso de que el dato faltante se produzca por causas positivas, por ejemplo, curación, es sustituir el dato faltante por el mejor valor posible. Explorar el peor y mejor escenario posible, permiten crear límites inferior y superior para el efecto de la intervención en estudio. [2]

También se pueden completar los datos faltantes con imputación de datos, es decir, estimar el valor perdido, partiendo siempre del supuesto que el comportamiento de los registros con valores faltantes es siempre el mismo que el de los registros con valores existentes. Tenemos imputación simple, que consiste en completar los datos faltantes de cada variable utilizando, por ejemplo, el valor medio de dicha variable. Se reemplaza cada dato faltante por un único valor.

En 1982, Little, introdujo un concepto que hoy constituye una alternativa fundamental en el manejo de datos faltantes, la imputación múltiple. Es una técnica muy robusta y se basa en que cada dato faltante se estima múltiples veces y luego se combinan dichas estimaciones para lograr un único valor.

2.4 VIH

El VIH, virus de la inmunodeficiencia humana, es un virus que ataca el sistema inmunitario del cuerpo. Si no se trata, puede causar SIDA, síndrome de inmunodeficiencia adquirida.

El VIH pertenece a la familia de los lentivirus y se clasifica en dos tipos: VIH-1 y VIH-2, entre ambos hay un 40-50% de homología genética y poseen una organización genómica similar. El VIH-1 es el causante de la pandemia mundial de sida mientras que el VIH-2, aunque también puede producir sida, se considera menos patogénico y transmisible.

Por medio del estudio evolutivo de secuencias se piensa que un retrovitus de simio pasó del chimpancé a la especie humana alrededor de 1900 en África central. El mecanismo de exposición más probable ha sido la caza y el consumo de carne de chimpancé, práctica muy popular en la zona, donde se han descrito infecciones en humanos. Pasó inadvertido hasta que empezó a afectar a los países ricos. La muestra humana más antigua es del año 1959, fue tomada a un marino británico que se cree que contrajo el VIH en la República Democrática del Congo. Los primeros casos de lo que ahora conocemos como SIDA fueron detectados en Nueva York y Los Angeles en 1981, se observó en pacientes jóvenes homosexuales previamente sanos el desarrollo de infecciones oportunista, tras estudiar los primeros casos, se asociaron estas manifestaciones con una inmunodeficiencia celular adquirida no descrita hasta ese momento, pero no fue hasta el 20 de mayo de 1983 cuando se publicó en la revista Science el descubrimiento del Virus de

Inmunodeficiencia Humana (VIH). Por lo tanto, en humanos adquirió consideración pándemica hace 40 años. [3]

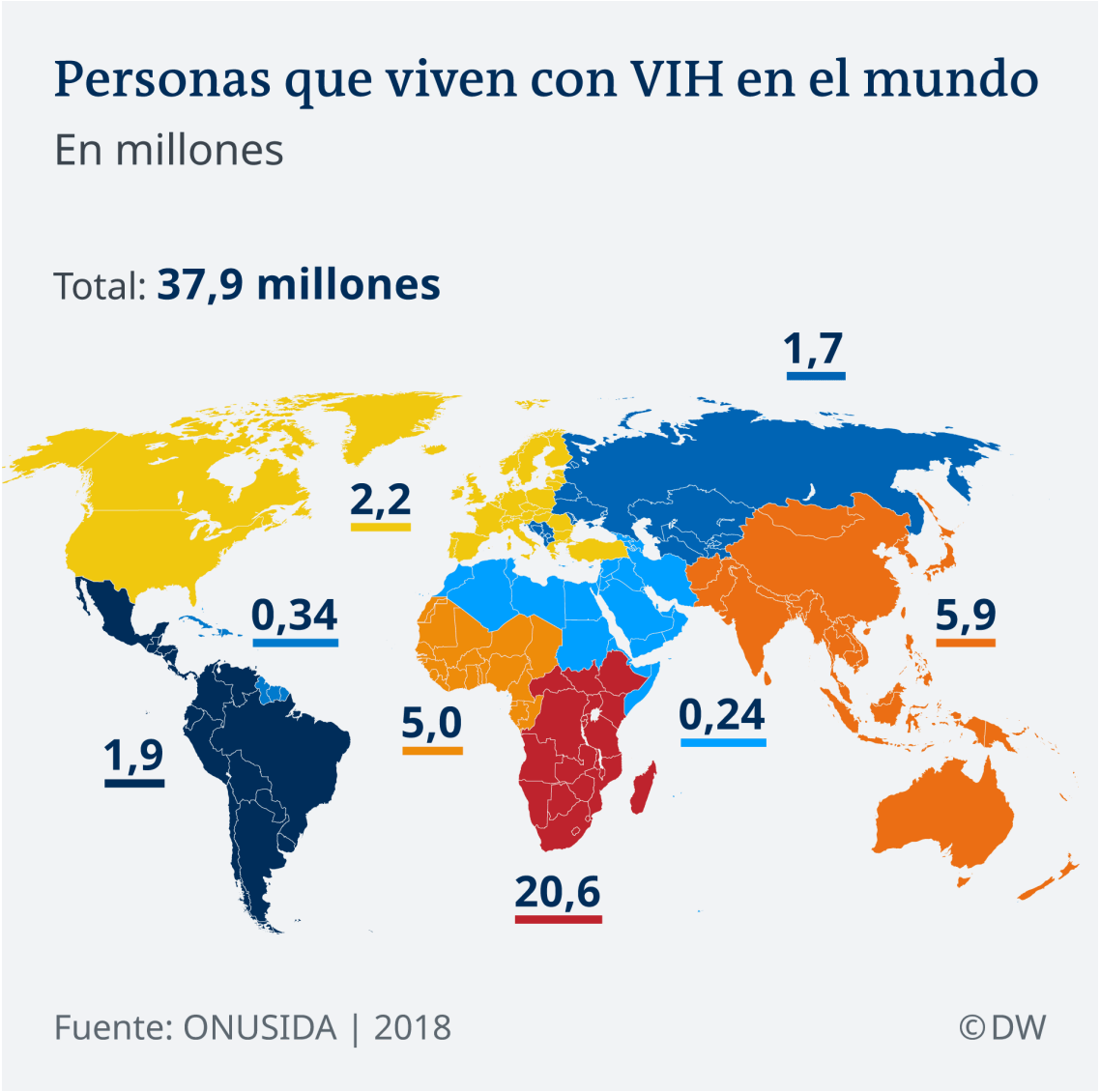

Ilustración 8: Distribución VIH mundo (Fuente: ONUSIDA)

El VIH en cifras: [17] [18]

- 79,3 millones de personas contrajeron la infección por el VIH desde el comienzo de la epidemia. 39 millones de personas han muerto a causa del sida en todo el mundo.

- Se calcula que el VIH afecta a más de 38 millones de personas en todo el mundo. Entre un 66% y un 73% tiene acceso al tratamiento antirretroviral.

- En 2019 se notificaron 1,7 millones de nuevas infecciones: cada 18 segundos se produce una nueva infección.

- Se calcula que un 20% de las personas que tienen VIH no lo saben.

- El número de diagnósticos de casos de sida desciende año tras año en los países en que existe un buen acceso al tratamiento antirretroviral.

- Desde el pico alcanzado en 1997, las nuevas infecciones por el VIH se han reducido en un 52%.

- El pico de infecciones fue en 1997 con 3 millones de infecciones. En 2020, se produjeron 1,5 millones de nuevas infecciones por el VIH.

- Desde 2010, las nuevas infecciones por el VIH descendieron alrededor de un 31%, desde 2,1 millones hasta 1,5 millones en 2020.

- En 2020, alrededor de 680.000 personas murieron de enfermedades relacionadas con el sida en todo el mundo, frente a los 1,9 millones de 2004, que fue el año que se alcanzó el pico de muertes.

- En 2020, de toda la gente que vive con el VIH, el 84% conocían su estado, el 73% tenían acceso al tratamiento y el 66% tenían una carga viral indetectable.

Los hitos en el tratamiento del VIH son los siguientes, en 1987 se aprueba el primer fármaco para el tratamiento del VIH, zidovudina o AZT, los problemas eran sus elevadas dosis y su gran toxicidad. Durante los primeros años de la década de los 90, a pesar de ser aprobados varios antirretrovirales cuyo mecanismo de acción era inhibir la transcriptasa análogos de nucleósido (ITIN) y varios antibióticos para el tratamiento de infecciones oportunistas en individuos inmunodeprimidos, fueron muy duros en la lucha contra el VIH, ya que la incidencia iba en aumento y muchos de los diagnosticados acababan falleciendo de sida.

En 1995 empieza la era del Tratamiento Antirretroviral de Gran Actividad (TARGA) y se produce un giro radical en el tratamiento del VIH ya que existe la posibilidad de tratar el VIH con terapias combinadas y detener la progresión de la enfermedad. Tras esto se vivieron unos primeros años de gran optimismo y de aprobación de muchos fármacos, los primeros con mucha complejidad en las tomas y con gran toxicidad, después hubo algunas mejoras con fármacos inhibidores de la transcriptasa inversa no análogos de nucleósido (ITINN) que disminuyen la complejidad de las tomas pero a pesar de que la replicación del virus de la inmunodeficiencia humana (VIH) puede ser suprimida con los tratamientos actualmente disponibles, la erradicación de la infección por el VIH es todavía un objetivo inalcanzable. Por ello, el tratamiento antirretroviral debe ser establecido de por vida en la mayoría de los infectados por el VIH. [1]

La eficacia del tratamiento antirretroviral de gran actividad (TARGA) ha sido demostrada en varios ensayos clínicos. Aun así, una importante proporción de pacientes no consigue mantener una correcta supresión viral en la práctica clínica diaria. [1]

La adherencia al tratamiento TARGA es crítica para obtener una supresión viral duradera. Por ello, factores que se relacionan con la adherencia como el

elevado número de pastillas o de tomas, la complejidad del régimen antirretroviral, su tolerabilidad y las restricciones alimentarias pueden tener un efecto sobre la replicación viral. [1]

Se ha demostrado que regímenes más sencillos, con escaso número de pastillas, sin restricciones alimentarias y con una sola toma al día son seguros, eficaces, y que mejoran la adherencia. [1]

2.5 Ensayo Lake

La base de datos que se utilizará en el desarrollo de este TFM es la base de datos del ensayo Lake. Con este ensayo clínico se evalúa la equivalencia terapéutica entre las dos ramas de tratamiento. La rama experimental es una asociación terapéutica con una nueva posología, Abacavir 600 mg + lamivudina 300 mg en un sólo comprimido QD + Efavirenz 600 mg QD frente a la rama control: Abacavir 600 mg+ lamivudina 300 mg en un sólo comprimido QD + lopinavir/ ritonavir 400/ 100 mg BID. La combinación abacavir 600 mg + lamivudina 300 mg QD en un sólo comprimido, una vez al día, es una posología novedosa que puede hacer aumentar la adherencia al tratamiento. Los pacientes de la rama A del ensayo deberán tomar 2 comprimidos diarios mientras los de la rama B, que recibirán el tratamiento estándar, deberán tomar 7 comprimidos al día. [1]

Según el protocolo del ensayo Lake, la población de estudio son pacientes con infección por VIH sin experiencia antirretroviral previa. El ensayo clínico se diseñó con un tamaño muestral de 126 pacientes, aleatorizados 1:1, 63 en cada rama. El objetivo principal de este estudio es evaluar la equivalencia terapéutica entre las dos ramas de tratamiento, la variable principal que se estudia es la respuesta virológica durante las 48 semanas de duración del estudio. Evaluar el porcentaje de fracasos virológicos. Mientras que los Objetivos secundarios son evaluar la respuesta inmunológica durante las 48 semanas de duración del estudio, evaluar la tolerabilidad y seguridad de la combinación estudiada durante el periodo de seguimiento, evaluar la repercusión del tratamiento sobre el perfil lipídico, evaluar la adherencia al tratamiento (valorado mediante cuestionario autorreferido y con escalas de satisfacción graduadas) y la calidad de vida de los pacientes (valorado mediante el cuestionario MOS-HIV) y analizar las mutaciones que aparecen en los pacientes que presenten fracaso virológico. [1]

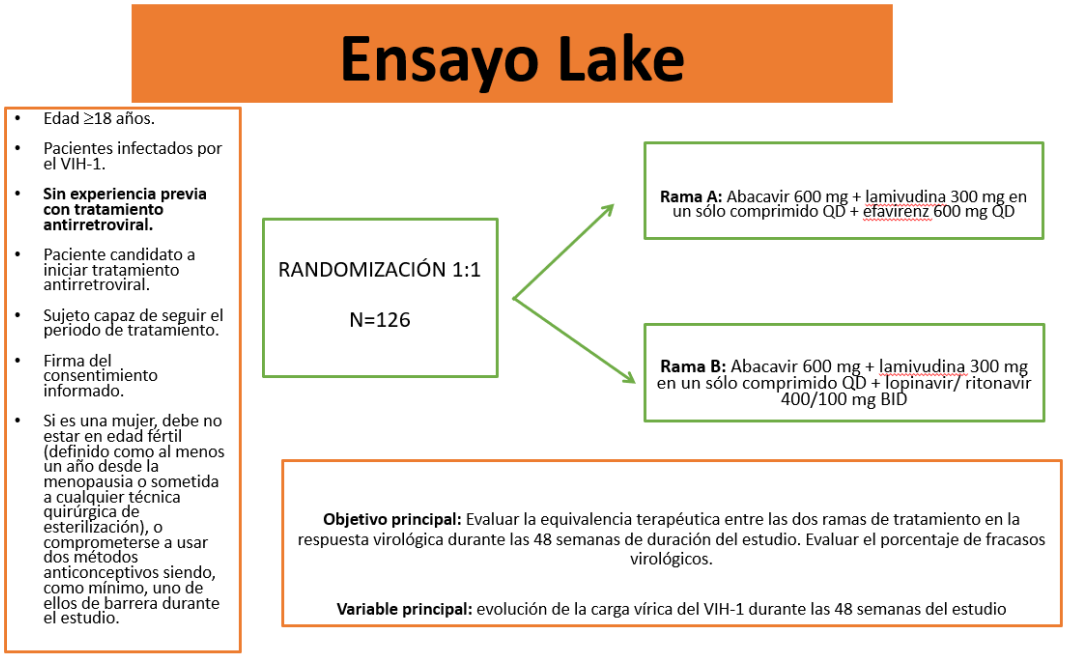

Ilustración 9: Esquema ensayo Lake

El ensayo planifica 7 visitas de los pacientes: screening, inclusión, 4 semanas de seguimiento, 12 semanas de seguimiento, 24 semanas de seguimiento, 36 semanas de seguimiento, 48 semanas de seguimiento.

En el ensayo Lake, el fracaso virológico se define como la aparición de dos cargas virales de más de 400 copias/mL, separadas por un periodo de 4 semanas en aquellos pacientes que han conseguido la indetectabilidad; o bien, imposibilidad de conseguir la indetectabilidad a las 12 semanas de iniciar el tratamiento. [1]

La variable principal de valoración es la proporción de pacientes que sigue con carga viral indetectable al final del periodo de estudio.

La duración del ensayo se programó en 72 semanas aproximadamente (24 semanas de inclusión y 48 semanas de seguimiento).

En el ensayo Lake se mide la carga viral en cinco momentos, al inicio, a las 12 semanas de tratamiento, a las 24, 36 y 48 semanas de tratamiento.

La carga viral mide la cantidad de RNA del virus del VIH presente en la sangre. Se habla del número de copias de VIH en un mililitro de sangre (copias/mL). La carga viral aporta una información relevante tanto durante el diagnóstico como para saber si una determinada medicación está funcionando. El objetivo del tratamiento antirretroviral es mantener la carga viral indetectable. Se considera que una carga viral es indetectable cuando se sitúa por debajo de las 50 copias/mL; no obstante, en la actualidad, el significado de indetectable depende del método de análisis empleado en cada centro sanitario, y puede ser inferior a 20, 37 o 50 copias por mililitro. Por tanto, aunque no se detecten copias del virus en la prueba, puede haber pequeñas cantidades de virus en la sangre de los pacientes con carga viral indetectable. En varios estudios se ha

concluido que cuando los valores de carga viral de VIH se mantienen indetectables, el virus no se transmite.

3. Análisis de la base de datos

3.1 Variables recogidas en la base de datos

La base de datos del ensayo Lake tiene 116 pacientes randomizados 1:1, 58 en cada uno de los dos brazos de tratamiento y recoge 219 variables de cada paciente.

Ilustración 10: Descripción variables base de datos ensayo Lake

Las primeras 22 variables que se recogen son datos identificativos y demográficos de cada paciente: proc, nusuario, npac (número de paciente), nvisita, fecha_nac, sexo, factorriesgo_ADVP, especificar, a19, estadio_VIH_20, fecha_ini_lake, factorriesgo_heterosexual, factorriesgo_homosexual, factorriesgo_hemofilia, factorriesgo_otros, a28, estadio_VIH_31, a32, fecha_vih, Estado, edad, Grupo.

Además, se recogen las mismas 37 variables en cinco momentos diferentes. En la visita basal, a las 12 semanas de tratamiento, a las 24 semanas de tratamiento, a las 36 semanas de tratamiento y a las 48 semanas de tratamiento. Estas variables que se repiten en el tiempo son: week, Fecha, CargaViral, CD4A, CD4P, CD8A, CD8P, Hematocrito, Hemoglobina, Plaquetas, Leucocitos, LinfosTotales, Glucosa_mg, Urea_mg, Creatinina_mumol, Sodio, Potasio, Cloro, Calcio, Bilirrubina_mumol, GPT, GOT, GGT, ProteinasTotales, Albumina, Colesterol_mg, LDL_mg, HDL_mg, Trigliceridos_mg, Amilasa, pH, Bicarbonato, AcidoLactico, AcidoPiruvico, VHC, VHB, Embarazo.

El resto de las variables que recoge la base de datos son variables calculadas a partir de las anteriores: cv50_0, cv50_12, cv50_24, cv50_36, cv50_48, tpo_vih_meses, factor_riesgo_total, diff_cd4_48_0, diff_cd4p_48_0, diff_col_48_0, diff_HDL_48_0, diff_LDL_48_0.

La variable principal será la evolución de la carga vírica del VIH-1 durante las 48 semanas del estudio

3.2 Análisis descriptivo de características demográficas y otros datos basales

De los 116 pacientes incluidos, 58 se incluyeron en la rama experimental del tratamiento: Abacavir 600 mg + lamivudina 300 mg en un sólo comprimido QD + Efavirenz 600 mg QD, mientras que los 58 restantes se incluyeron en la rama control: Abacavir 600 mg+ lamivudina 300 mg en un sólo comprimido QD + lopinavir/ ritonavir 400/ 100 mg BID.

En la base de datos, se recogen distintes variables demográficas, es interesante saber la distribución de los pacientes en función de esas variables en ambas ramas de tratamiento:

Variable	Categoría	Rama Experimental n=58,%rama,%total	Rama control n=58%rama,%total	Total Pacientes, %
Sexo	Hombre	48 (82,76%) (41,38%)	47 (81,03%) (40,52%)	95 (81,90%)
	Mujer	8 (13,79%) (6,90%)	7 (12,07%) (6,03%)	15 (12,93%)
Edad	<30	8 (13,79%) (6,90%)	8 (13,79%) (6,90%)	16 (13,79%)
	30-39	26 (44,83%) (22,41%)	30 (51,72%) (25,86%)	56 (48,28%)
	40-50	16 (27,59%) (13,79%)	15(25,86%) (12,93%)	31 (26,72%)
	>50	7 (12,07%) (6,03%)	5 (8,62%) (4,31%)	12 (10,34%)
Factor de riesgo homosexual	Si	26 (44,83%) (22,41%)	21 (36,21%) (18,10%)	47 (40,52%)
	No	31 (53,83%) (26,72%)	37 (63,79%) (31,90%)	68 (58,62%)
Factor de riesgo heterosexual	Si	16 (27,59%) (13,79%)	24 (41,38%) (20,69%)	40 (34,48%)
	No	41 (70,69%) (35,34%)	34 (58,62%) (29,31%)	75 (64,66%)
Estadío de la enfermedad	A	39 (67,24%) (33,62%)	37 (63,79%) (31,90%)	76 (65,52%)
	B	11 (18,97%) (9,48%)	12 (20,69%) (10,34%)	33 (28,45%)
	C	4 (6,90%) (3,45%)	4 (6,90%) (3,45%)	8 (6,90%)

Ilustración 11: Distribución pacientes según variables demográficas

Del total de los pacientes incluidos en la base de datos del ensayo Lake, un 81,90% de los pacientes son hombres, distribuyéndose homogéneamente (82,75% rama experimental y 81,03% rama control) en ambas ramas de tratamiento.

Por franjas de edad, en ambos brazos de tratamiento, el grupo más numeroso es el de 30-39 años, suponiendo un 48,28 % de los pacientes.

De la población total, un 40,52% de los pacientes cumplían con el factor de riesgo de ser homosexual.

Si nos fijamos en el estadio de la enfermedad, un 6,90% tenía estadio de enfermedad C basal. Siendo el estadio A el más numeroso con un 65,52% de la población.

De las variables importantes para evaluar eficacia en la respuesta, como es la carga viral inicial, antes de empezar el tratamiento del ensayo clínico, el valor medio observado en la población incluida es 228612,8 copias/ml.

En cuanto a las variables que nos permiten evaluar la eficacia inmunológica, cifra de CD4 y CD8, en el momento basal el contaje de CD4 absoluto medio es de 192,557 cel/mcL (191,817 cel/mcL en Rama A y 193,4261 cel/mcL en Rama B). El contaje de CD8 absoluto medio es de 967,596 cel/mcL

3.3 Datos faltantes

En la base de datos del ensayo Lake, el porcentaje de valores faltantes es de un 41% del total de los datos. Hay 116 filas, que se corresponden con 116 pacientes, de cada paciente se recogen 219 variables. En 206 de ellas hay al menos un dato faltante. En todos los pacientes falta mínimo una de sus variables.

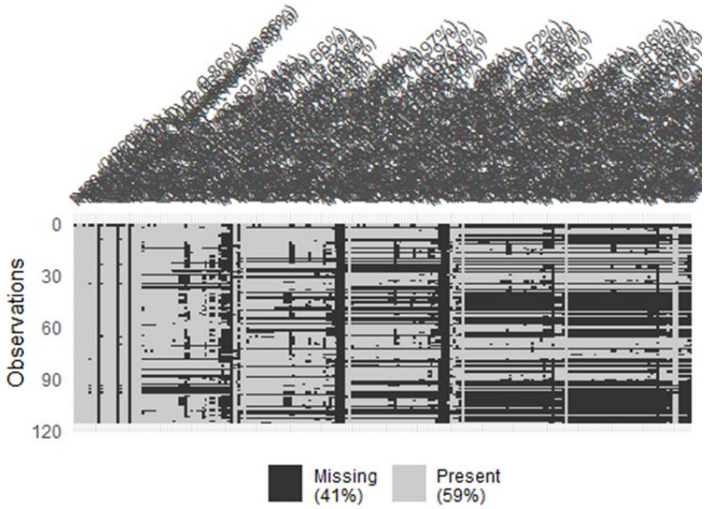

Ilustración 12: Gráfico datos faltantes base de datos ensayo Lake

Se divide la base de datos por grupos de variables. En el primer grupo, están las variables identificativas y demográficas, entre estas hay un 14,9% de datos faltantes. Visualizando el gráfico se puede observar que la presencia de esos datos faltantes es fundamentalmente en 3 de las 22 variables de este grupo.

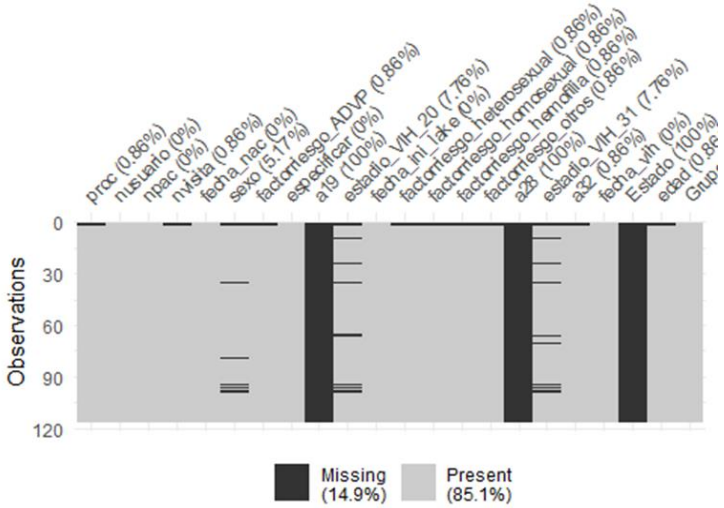

Ilustración 13: Gráfico datos faltantes variables identificativas y demográficas

Como en todos los ensayos longitudinales, existen una serie de medidas repetidas en el tiempo, que son unas variables que se miden en varios momentos a lo largo del ensayo, en la base de datos que se utiliza para el TFM son 37 variables, que se miden en el momento basal, antes de empezar el tratamiento, a las 12, 24, 36 y 48 semanas del inicio del tratamiento.

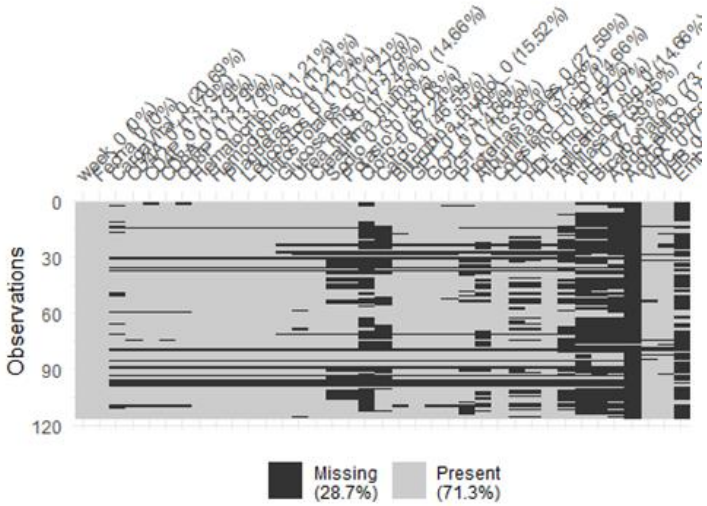

Ilustración 14: Gráfico datos faltantes variables momento basal

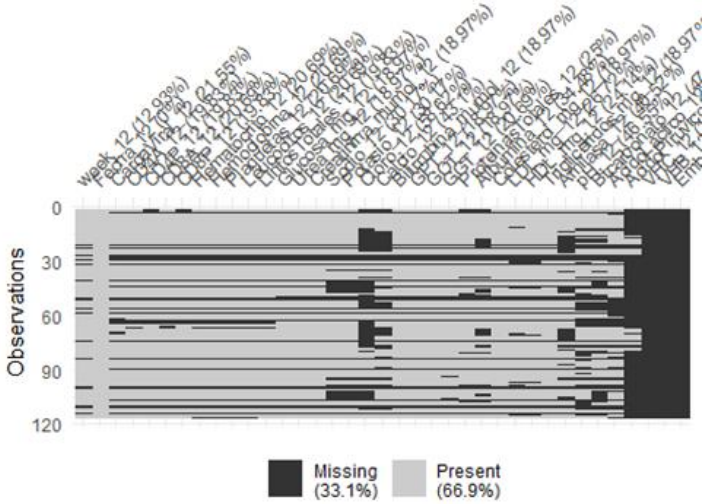

Ilustración 15: Gráfico datos faltantes de variables a las 12 semanas tratamiento

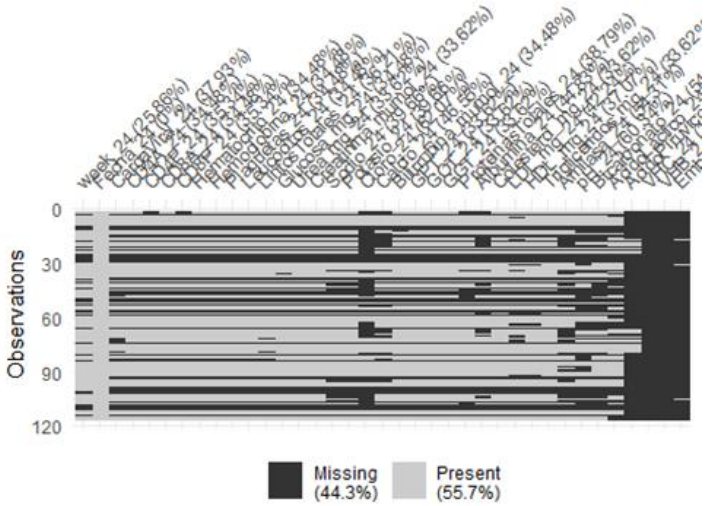

Ilustración 16: Gráfico datos faltantes de variables a las 24 semanas tratamiento

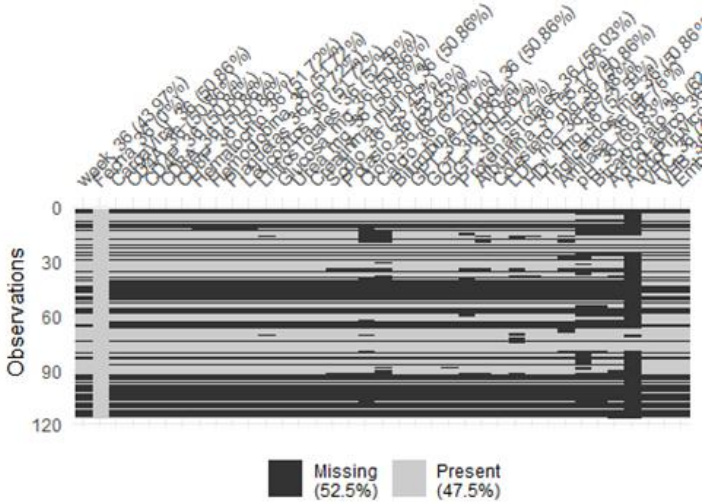

Ilustración 17: Gráfico datos faltantes de variables a las 24 semanas tratamiento

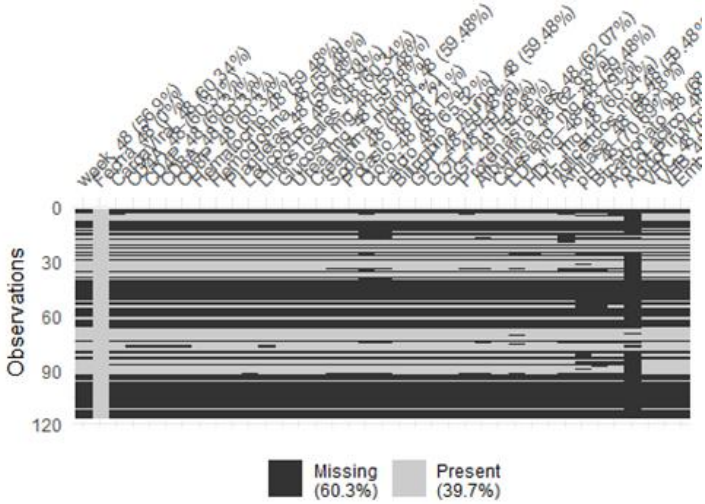

Ilustración 18: Gráfico datos faltantes de variables a las 48 semanas tratamiento

El porcentaje de datos faltantes de ese grupo de variables que se repiten en distintos momentos, podemos observar que aumenta a medida que aumenta el tiempo pasando de un 28,7% antes de iniciar el tratamiento, a un 33,1% en la semana 12, un 44,3% en la semana 24 de tratamiento, un 52,5% en la semana 36 de tratamiento, hasta un 60,3% de datos faltantes en la semana 48 de iniciar el tratamiento.

Hay 8 variables, que todos sus registros son faltantes.

skim_variable	n_missing	complete_rate	mean	count
a19	116	0	NaN	:
a28	116	0	NaN	:
Estado	116	0	NaN	:
VHC_12	116	0	NaN	:
VHB_12	116	0	NaN	:
Embarazo_12	116	0	NaN	:
VHC_24	116	0	NaN	:
VHB_24	116	0	NaN	:

Ilustración 19: Variables con todos los datos faltantes

4. Transformación de la base de datos

4.1 Limpieza de base de datos

Tras el estudio de los datos faltantes, se procede a la limpieza de la base de datos, seleccionando solamente variables que puedan ser interesantes para los análisis que se van a realizar en este TFM.

Se eliminan todas aquellas variables sin registros en ningún paciente y variables que no tengan interés. La base de datos tras la limpieza de datos contiene el mismo número de pacientes que la base de datos original, 116 pacientes con 47 variables. Tras la eliminación de dichas variables, el porcentaje de datos faltantes se reduce a 28,9%.

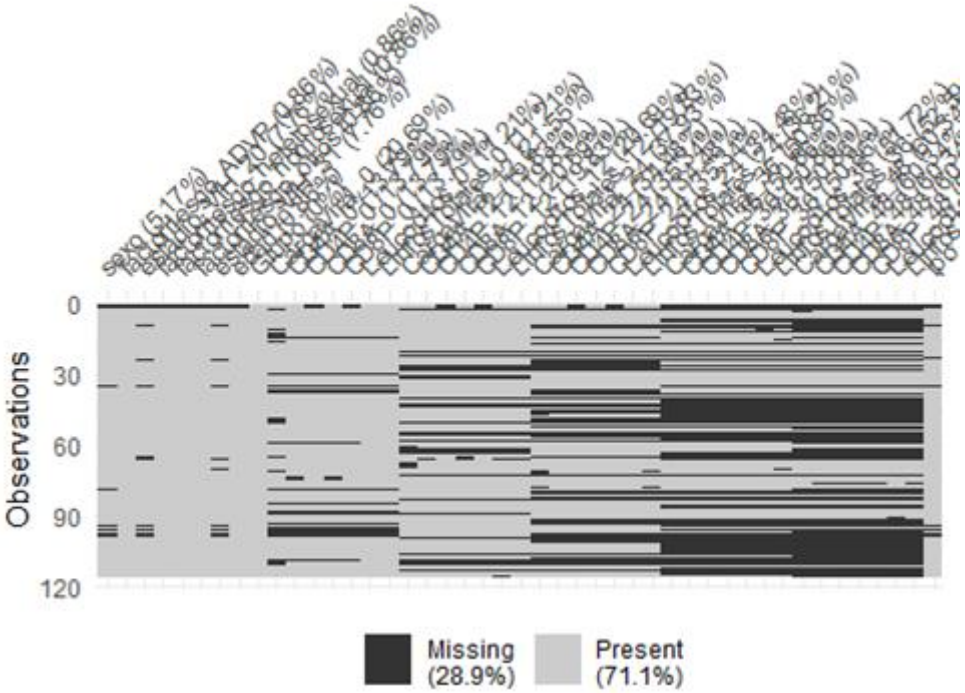

Ilustración 20: Gráfico de datos faltante tras limpieza de datos

4.2 Creación base de datos TFM

En todo análisis de supervivencia, se estudia el tiempo hasta la ocurrencia de un evento. Definir el evento que se va a estudiar es una parte fundamental de dicho análisis.

En este TFM, al igual que en el ensayo Lake, la variable principal que se estudia es la respuesta virológica durante las 48 semanas de duración del estudio. La variable de la base de datos que marca dicha respuesta virológica es la carga viral, mide la cantidad de RNA del virus del VIH presente en la sangre. Se habla del número de copias de VIH en un mililitro de sangre (copias/mL). El evento que se va a estudiar es el fracaso virológico. Para este TFM se define que un paciente tiene fracaso virológico cuando no logre una carga viral igual o menor a 50 copias/ml en la semana 12 de tratamiento y en el caso de haberla conseguido, se considerará fracaso también si no consigue mantenerla hasta la semana 48 de tratamiento.

Se crea la variable indetectable a las 12 semanas, que tendrá un 0 en aquellas pacientes que tengan carga viral menor o igual a 50copias/ml y un 1 si tienen más, es decir, ya han tenido fracaso virológico. Se crea la variable indetectabilidad también en la semana, 24, 36 y 48, tendrán un 0 en dicha variable, aquellas pacientes que tengan una carga viral menor o igual a 50, si por el contrario, en alguna de las mediciones tienen más de 50copias/ml, tendrán un 1, es decir fracaso virológico. Los pacientes que tengan en la variable indetectabilidad un 1, en cualquier de los momentos que se especifica, se considerará que tiene el evento, fracaso virológico. Aquellos que tienen 0, no habrán tenido el evento, serán datos censurados.

5. Imputación de datos faltantes

En este TFM se asume que el mecanismo de generación de datos faltantes sigue el patrón MCAR, Mising Completely at Random, la causa de este tipo de datos faltantes es totalmente al azar, es por lo tanto independiente de la medicación utilizada en el ensayo o bien el patrón MAR, Missing At Random, la probabilidad de que una respuesta sea dato faltante es independiente de los valores de la misma variable, pero es dependiente de los valores de otras variables del conjunto de datos. Para estos dos patrones de datos faltantes, una de las opciones recomendadas es la imputación múltiple, se basa en que cada dato faltante se estima múltiples veces, teniendo en cuenta los valores existentes y luego se combinan dichas estimaciones para lograr un único valor.

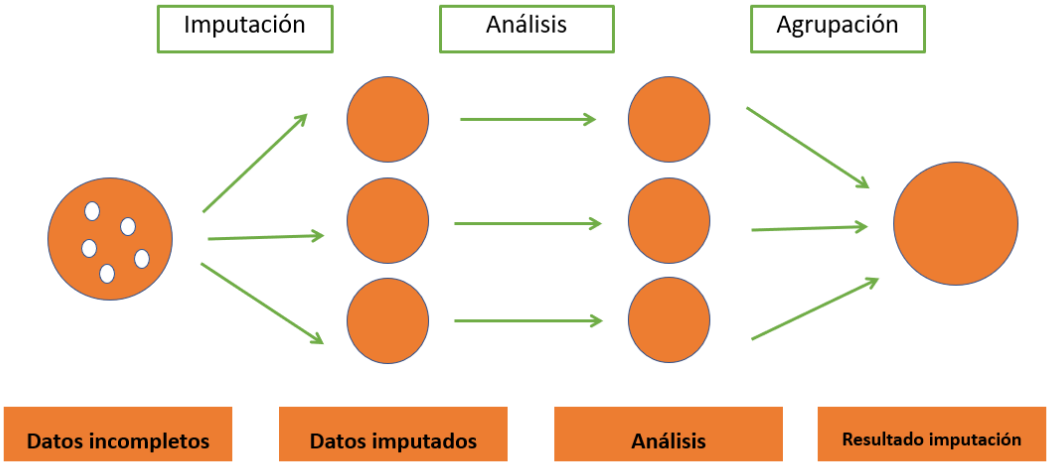

Ilustración 21: Esquema imputación múltiple

5.1 Imputación con MICE

Una de las opciones disponibles dentro de imputación múltiple, probablemente la más utilizada, es la disponible en la librería mice, de sus siglas en inglés, Multiple Imputation by Chained Equations. Utiliza un algoritmo de imputación con ecuaciones encadenadas.

Siguiendo los pasos que se indican en el libro de Stef Van Buuren [32], se realiza una imputación con cero iteraciones. Con ello se podrá extraer y modificar los métodos de imputación para, según el tipo de variable, asignar un método específico a cada variable ordinal y binaria en la imputación definitiva.

Seguidamente se calculan el influx y el outflux mediante la función flux, que también está en el paquete mice. El influx de una variable cuantifica cómo de bien se conectan sus datos faltantes a los datos observados en otras variables. El outflux de una variable cuantifica cómo de bien se conectan sus datos observados a los datos faltantes de otras variables. Por ello el outflux se convierte en un estadístico importante de cara a medir la importancia de una variable como predictora en el proceso de imputación. En este caso, se

excluyeron de la matriz de predictores aquellas variables con un valor menor a 0,4 (que Stef Van Buuren considera ya bajo), y se incluyeron aquellas variables con outflux mayor que 0,7 (se trata de un outflux relativamente alto y que permitía tener en torno a 20 variables predictoras para cada variable a imputar). Además, se añadieron a la matriz de predictores las variables “CargaViral_0” y “CargaViral_12” por su importancia clínica. Finalmente, utilizando la función quickpred del paquete mice se seleccionan los predictores utilizando los estadísticos antes indicados. En esta función se ajustan los parámetros mincor (el o los umbrales mínimos con los que se compara la correlación absoluta en los datos) y minpuc (el o los umbrales mínimos para la proporción de casos utilizables) para que sea exigente de cara a que las variables sean elegidas como predictores.

La imputación definitiva se realizó con: 25 iteraciones, utilizando los métodos y la matriz de predictores antes definidos, con el número 27 como semilla aleatoria, indicando que no escriba la historia de la imputación en la consola, y con el parámetro ridge en 1e-04 debido a la multicolinealidad (relación de dependencia lineal fuerte entre dos o más variables explicativas en una regresión múltiple) de algunas variables del dataset. (1e-03 daría menos errores, pero cabría el riesgo de producir sesgos, y con 1e-05 da muchos errores). De los 5 datasets resultantes de la imputación, se escogió el número 1.

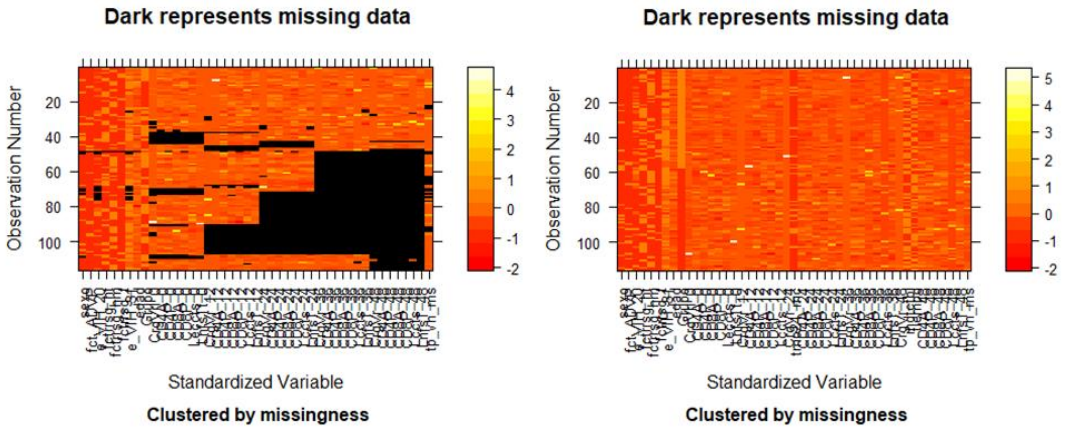

Ilustración 22: Grafico de antes y después imputación MICE

Tras la imputación se hace un gráfico para saber cuántos pacientes y en qué momento han presentado el evento. [23]

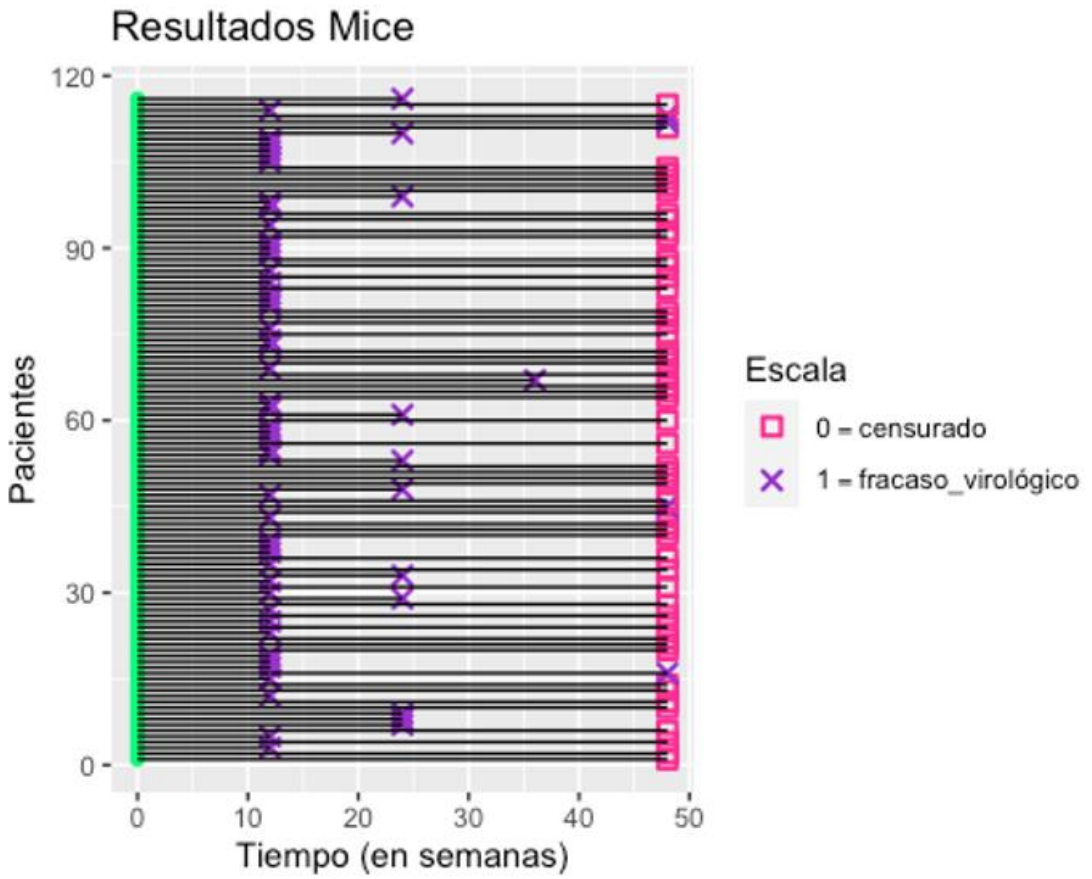

Ilustración 23: Esquema tiempo a fracaso virológico y censura por pacientes tras imputación MICE

Con una cruz se muestran los pacientes que han tenido fracaso virológico y cuantas semanas tras la randomización en el ensayo Lake han tardado en tenerlo. Los pacientes representados con un círculo rosa en la semana 48, representan datos censurados, es decir, aquellos que en las 48 semanas que duraba el seguimiento no han experimentado el evento.

5.2 Imputación con randomForestSRC

Los parámetros de la función impute.rfsrc se definieron con el objetivo de alcanzar la mayor precisión posible, pero no hacer el proceso muy costoso en términos de tiempo de computación. Son los siguientes: el número de árboles se fijó en 500, el tamaño medio del nodo terminal se definió en 1, el valor utilizado para especificar la división aleatoria es 10, el número de iteraciones del algoritmo de imputación se marcó en 5, el número de variables que se muestrean aleatoriamente en cada división se indicó que no se quería utilizar la imputación rápida, se marcó el valor de tolerancia en 0,01, se indicó que se enviase la información del proceso al terminal, se fijó la semilla aletatoria en 27, y la regla de división se definió como aleatoria.

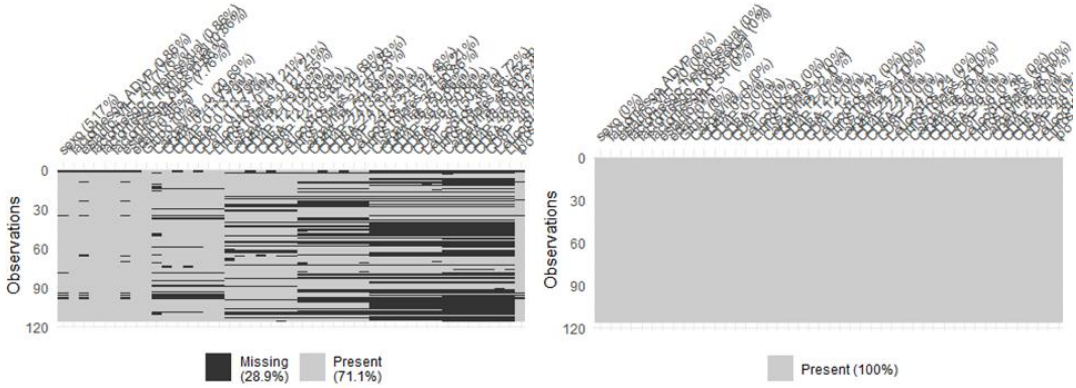

Ilustración 24: Grafico de antes y después imputación randomForestSRC

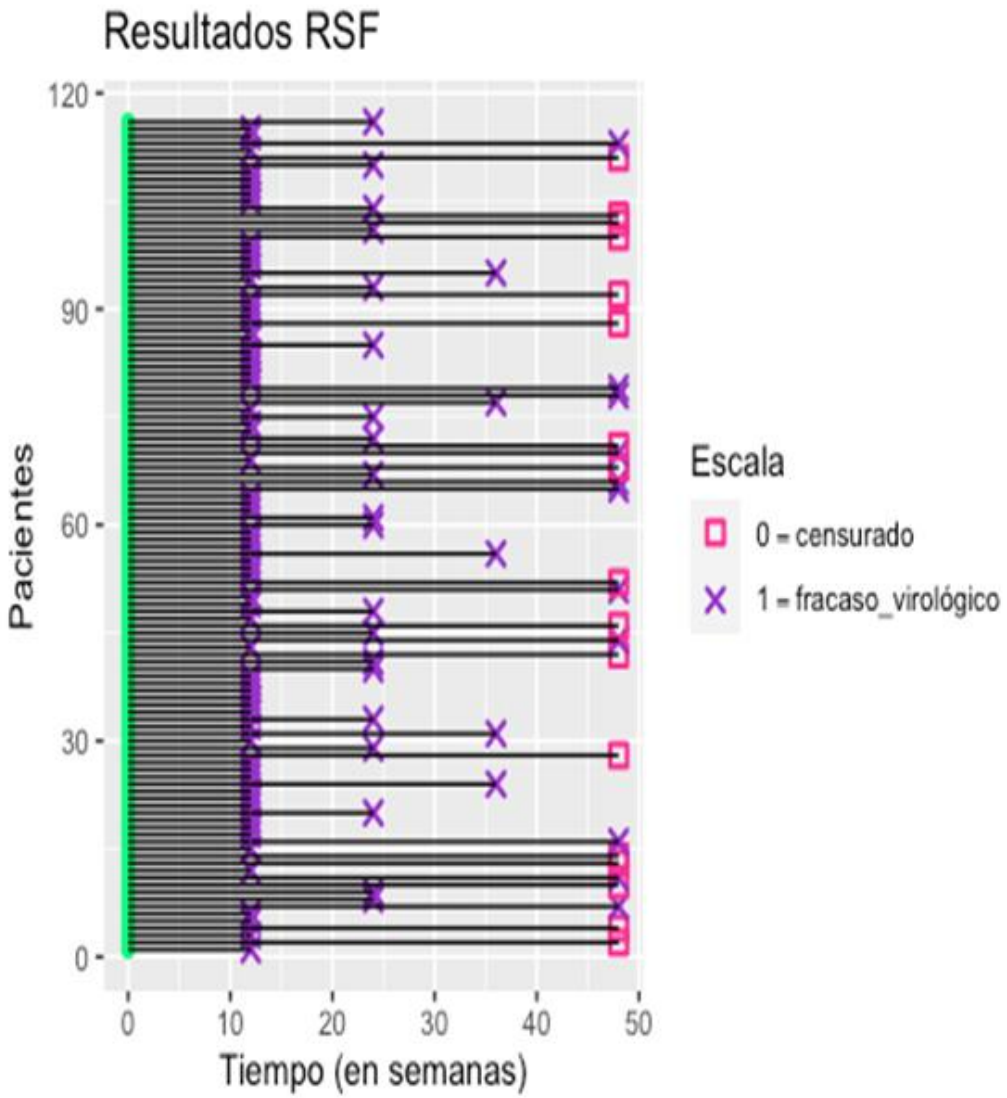

Ilustración 25: Esquema tiempo a fracaso virológico y censura por pacientes tras imputación randomForestSRC

Con una cruz se muestran los pacientes que han tenido fracaso virológico y cuantas semanas tras la randomización en el ensayo Lake han tardado en tenerlo. Los pacientes representados con un círculo rosa en la semana 48, representan datos censurados, es decir, aquellos que en las 48 semanas que duraba el seguimiento no han experimentado el evento.

6. Análisis de supervivencia Kaplan Meier

La variable principal que se estudia en este TFM es la respuesta virológica durante las 48 semanas de duración del estudio. La variable de la base de datos que marca dicha respuesta virológica es la carga viral, mide la cantidad de RNA del virus del VIH presente en la sangre. Se habla del número de copias de VIH en un mililitro de sangre (copias/mL). El evento que se va a estudiar es el fracaso virológico. Para este TFM se define que un paciente tiene fracaso virológico cuando no logre una carga viral igual o menor a 50 copias/ml en la semana 12 de tratamiento y en el caso de haberla conseguido, se considerará fracaso también si no consigue mantenerla hasta la semana 48 de tratamiento.

Para la realización de las curvas de Kaplan Meier se necesitan dos datos de cada paciente, saber si ocurre o no el evento de estudio, en nuestro caso fracaso virológico y cuánto tiempo tarda en ocurrir si ocurre o en el caso de no ocurrencia del evento, tiempo de seguimiento hasta la censura.

El método de Kaplan Meier es un método de análisis de supervivencia no paramétrico, es decir, los datos no se ajustan a ninguna distribución conocida a priori. Dicho método es especialmente útil para tratar con datos censurados, tal y como ocurre en la base de datos que estamos utilizando.

Con la función de supervivencia se puede saber la probabilidad que tiene un paciente en un determinado momento de no haber experimentado el evento que se está estudiando. Por lo tanto, con dicha función se mide la no ocurrencia del evento.

Con la función de riesgo, sin embargo, medimos la ocurrencia del evento, la probabilidad de que a un paciente de los que se sigue en el tiempo, en un determinado momento, se produzca el evento que estamos estudiando.

Habitualmente, con las curvas de supervivencia, se comparan dos grupos de individuos, en el caso de este TFM, se comparan 2 grupos definidos por su tratamiento, los pacientes de la rama A, son aquellos que reciben el tratamiento experimental, mientras que los de la rama B, reciben el tratamiento estándar. Para comparar curvas de supervivencia la prueba no paramétrica más potente y utilizada es la de Mantel-Cox conocida con el nombre de prueba de log-rank. [24]

El log rank es un contraste de hipótesis que compara curvas de supervivencia. La hipótesis nula, H0, es que las funciones de supervivencia de ambos

tratamientos son iguales. La hipótesis alternativa, H1, es que las funciones de supervivencia no son iguales. Ésta prueba compara el número de eventos observados en cada uno de los dos grupos de tratamiento, con el número de eventos esperados en el caso de que los eventos fueran los mismos en todos los grupos (H0). Si el valor del log rank es menor de 0.05 podemos afirmar que la hipótesis nula no se cumple, con lo que ambas funciones de supervivencia son distintas.

6.1 Curvas de supervivencia con base de datos imputada con MICE

Una vez que se tiene la base de datos imputada con mice, se representa la curva de supervivencia de ambos grupos de tratamiento, la rama A, que es la experimental y la rama B que es el grupo control y se calcula la función de supervivencia de cada grupo. Para esto se utiliza la librería survival y survminer de R. [25]

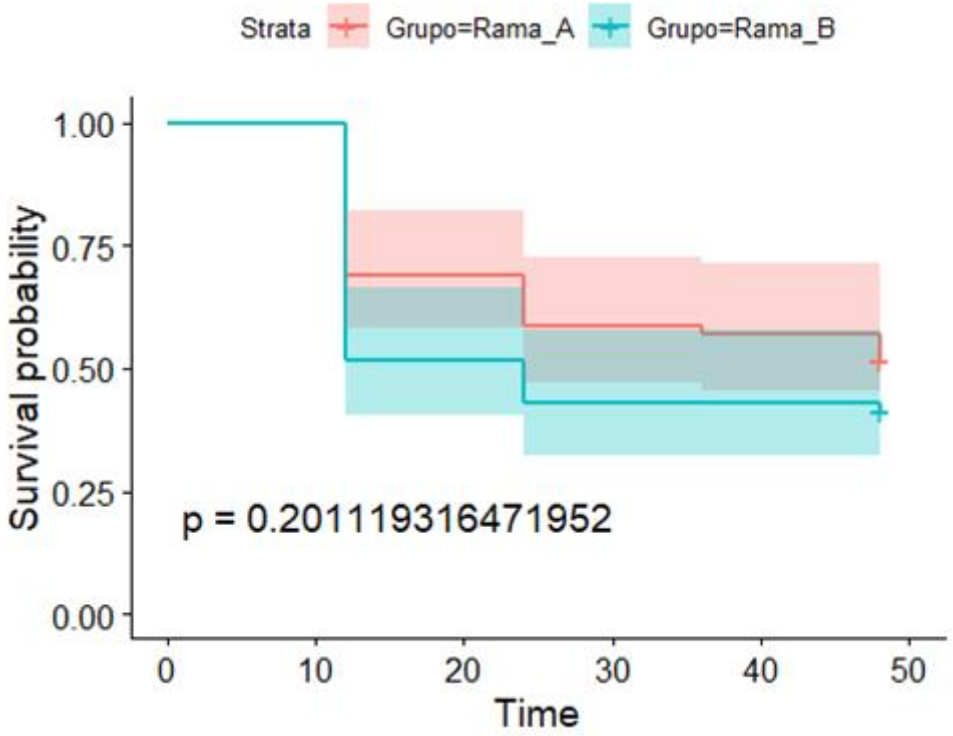

Ilustración 26: Curvas supervivencia Kaplan Meier tras imputación MICE

En la gráfica se puede ver que a las 12 semanas de tratamiento, la probabilidad de no haber tenido el evento, fracaso virológico, en los pacientes con VIH tratados con el tratamiento de la rama A es de aproximadamente 69% mientras que en los pacientes tratados con la rama B, a las 12 semanas, la probabilidad de no haber tenido fracaso virológico es del 51,7%. A las 24, 36 y 48 semanas

de iniciar el tratamiento, la probabilidad no haber tenido fracaso virológico es siempre superior en las pacientes de la rama A que de la rama B.

Se calcula la función de supervivencia por ramas de tratamiento y estos son los resultados. A las 48 semanas de tratamiento, la probabilidad de no haber tenido fracaso virológico en pacientes tratados con la rama experimental es del 51,7% mientras que en pacientes de la rama control, la probabilidad de no haber tenido fracaso virológico es del 41,4%. Es menos probable tener fracaso virológico si el paciente es tratado con la rama experimental.

time	n.risk	n.event	survival	std.err	lower 95% CI	upper 95% CI
12	58	18	0.690	0.0607	0.580	0.820
24	40	6	0.586	0.0647	0.472	0.728
36	34	1	0.569	0.0650	0.455	0.712
48	33	3	0.517	0.0656	0.403	0.663

Ilustración 27: Resultados función supervivencia pacientes rama A tras imputación MICE

time	n.risk	n.event	survival	std.err	lower 95% CI	upper 95% CI
12	58	28	0.517	0.0656	0.403	0.663
24	30	5	0.431	0.0650	0.321	0.579
48	25	1	0.414	0.0647	0.305	0.562

Ilustración 28: Resultados función supervivencia pacientes rama B tras imputación MICE

Para saber si estas diferencias son significativas, se calcula el p-valor asociado al estadístico de contraste, el log rank es 0.2011, que es mayor que 0,05, con lo que no podemos rechazar la hipótesis nula de igualdad de las dos funciones, en este caso las diferencias por lo tanto no son significativas.

Si se observan los resultados de los cálculos de la función de supervivencia tras la imputación con mice para toda la población, las conclusiones son las siguientes, de los 116 sujetos incluidos en el ensayo Lake, la probabilidad de no haber tenido fracaso virológico a las 12 semanas es de un 60,3%. La probabilidad de no haber tenido fracaso virológico tras 24 semanas de tratamiento es del 50,9%. A las 36 semanas la probabilidad de no haber tenido fracaso virológico coincide con la mediana, es del 50% mientras que la probabilidad de no haber tenido fracaso virológico en las 48 semanas que dura el seguimiento del ensayo Lake es del 46,6%.

time	n.risk	n.event	survival	std.err	lower 95% CI	upper 95% CI
12	116	46	0.603	0.0454	0.521	0.699
24	70	11	0.509	0.0464	0.425	0.608
36	59	1	0.500	0.0464	0.417	0.600
48	58	4	0.466	0.0463	0.383	0.566

Ilustración 29: Resultados función supervivencia tras imputación MICE

6.2 Curvas de supervivencia con base de datos imputada con randomForestSRC

Tras la imputación con la función impute.rfsrc de la librería randomForestSRC, se representan las curvas de supervivencia para los pacientes de la rama A y los pacientes de la rama B.

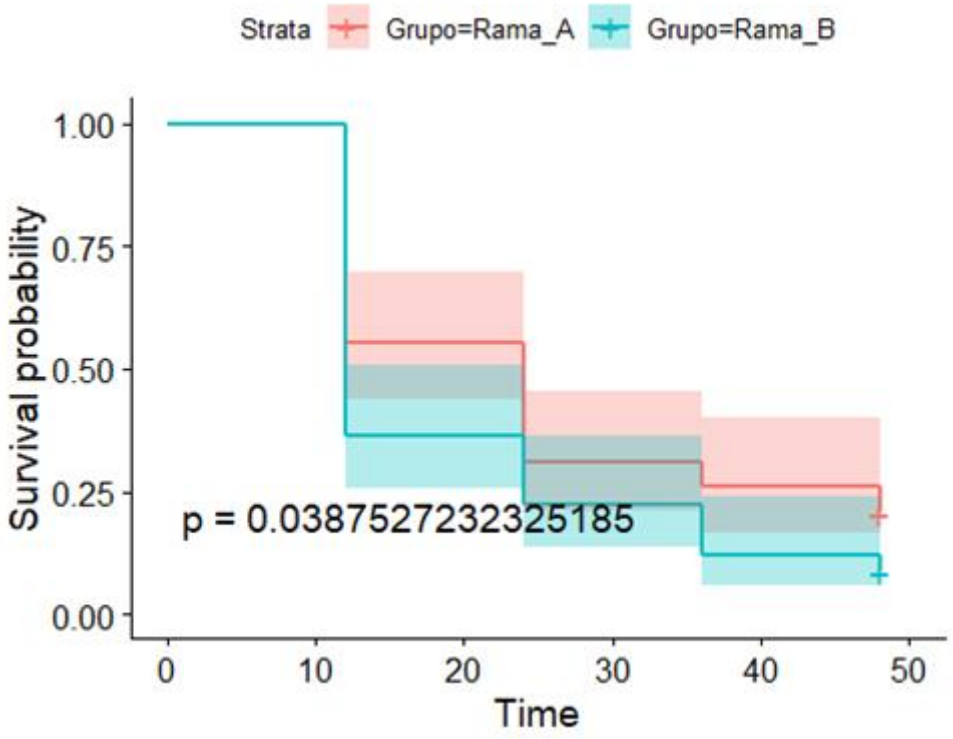

Ilustración 30: Curvas supervivencia Kaplan Meier tras imputación randomForestSRC

Tal y cómo se puede observar en la gráfica, la probabilidad de no tener fracaso virológico en pacientes tratados con la rama A, la experimental, es superior a la probabilidad de no tener el evento en pacientes tratados con la rama B, el tratamiento control.

A las 12 semanas de tratamiento, los pacientes de la rama A tienen un 55,2% de probabilidad de no tener fracaso virológico, mientras que los pacientes de la rama B, en el mismo tiempo, tienen un 36,2% de probabilidad de no haber sufrido dicho fracaso.

En la semana 24, la probabilidad de no evento en pacientes de la rama A es de 31% mientras que en la rama B es del 22,41%. A medida que el tiempo avanza la probabilidad de no tener el evento disminuye. En la semana 36, un 25,9% de la rama A frente al 12,07% de la rama B. En la semana 48, la probabilidad de no haber sufrido fracaso virológico en los pacientes de la rama A es del 20,7% mientras que en los de la rama B es del 8,6%.

Para saber si estas diferencias que se observan, son significativas o no lo son, se calcula el log rank, que en este caso es 0.0387, que es menor que 0,05, con lo que podemos rechazar la hipótesis nula de igualdad de las dos funciones, en este caso, las diferencias sí que son significativas.

time	n.risk	n.event	survival	std.err	lower 95% CI	upper 95% CI
12	58	26	0.552	0.0653	0.437	0.696
24	32	14	0.310	0.0607	0.211	0.455
36	18	3	0.259	0.0575	0.167	0.400
48	15	3	0.207	0.0532	0.125	0.342

Ilustración 31: Resultados función supervivencia pacientes rama A tras imputación randomForestSRC

time	n.risk	n.event	survival	std.err	lower 95% CI	upper 95% CI
12	58	37	0.3621	0.0631	0.2573	0.510
24	21	8	0.2241	0.0548	0.1389	0.362
36	13	6	0.1207	0.0428	0.0603	0.242
48	7	2	0.0862	0.0369	0.0373	0.199

Ilustración 32: Resultados función supervivencia pacientes rama B tras imputación randomForestSRC

Si se analizan los resultados de toda la población, independientemente de la rama de tratamiento, la probabilidad de no haber tenido el evento en la semana 12 de tratamiento es del 45,7% y disminuye hasta llegar al 14,7% en la semana 48 de tratamiento.

time	n.risk	n.event	survival	std.err	lower 95% CI	upper 95% CI
12	116	63	0.457	0.0463	0.3747	0.557
24	53	22	0.267	0.0411	0.1977	0.361
36	31	9	0.190	0.0364	0.1302	0.276
48	22	5	0.147	0.0328	0.0945	0.227

Ilustración 33: Resultados función supervivencia tras imputación randomForestSRC

7. Análisis de supervivencia Random Survival Forest

Desde hace unos años, una alternativa a los análisis de supervivencia convencionales son los realizados mediante machine learning. Dentro de ellos, una opción disponible en R es la que nos da la librería randomForestSRC. El algoritmo de random survival forest permite trabajar con datos censurados por la derecha.

Inicialmente se fija el dataset que se va a utilizar para alimentar al algoritmo. El siguiente paso es, mediante la función “tune” obtener los parámetros “mtry” y “nodesize” para obtener los mejores resultados aplicando el algoritmo rfsrc sobre cada uno de los datasets imputados. Con los valores obtenidos podemos personalizar los parámetros que se utilizarán en la función “rfsrc”. Finalmente, mediante la función “print” se muestran las métricas del algoritmo entrenado para cada dataset. [26]

En la función tune, los argumentos utilizados son los siguientes: inicialmente se especifica la fórmula del modelo que se quiere ajustar, posteriormente se indica el dataset de donde saldrán los datos, luego se fija el valor del punto inicial

para la búsqueda de valores de “mtry”, con nodesizetry se ensayarán los valores de cara a obtener el nodesize óptimo, “ntreeTry” es el argumento que define cuántos árboles se utilizan para el paso de ajuste de los argumentos, “sampsize” es la función que especifica el tamaño solicitado de los datos submuestrados, “nsplit” define el número de divisiones aleatorias utilizadas, “improve” es la mejora relativa del error Out Of Bag que se debe obtener para que la búsqueda continúe, “strikeout” define el número de veces que la búsqueda de valores puede dar mejoras negativas sin que la búsqueda se detenga, “maxIter” marca el número máximo de iteraciones que se permiten para cada búsqueda de bisección de “mtry”, “trace” habilida a la función tune para mostrar el progreso de la búsqueda, y “doBest” hace que en el resultado de la función se muestren los valores óptimos de los parámetros “mtry” y “nodesize”.

En la función rfsrc el primer argumento utilizado es la fórmula del modelo que se quiere ajustar, posteriormente se agrega el valor de “mtry” que se ha calculado previamente. Este argumento define el número de variables aleatoriamente seleccionadas como candidatas para un nodo que se divide. “ntree” define el número de árboles que se calculan en el proceso. “nodesize” define el número medio de casos únicos en un nodo terminal del bosque. Con “data” se marca cuál es el dataset del que se sacarán los datos. Y para finalizar “seed” define cuál es la semilla aleatoria utilizada para que los datos sean reproducibles.

7.1 Algoritmo Random Survival Forest con base de datos imputada con MICE

Se aplica el algoritmo a la base de datos imputada con mice y estos son los resultados. Muestra los parámetros ya comentados.

```
                         Sample size: 116
                    Number of deaths: 62
                     Number of trees: 500
          Forest terminal node size: 2
       Average no. of terminal nodes: 5.228
No. of variables tried at each split: 45
              Total no. of variables: 45
       Resampling used to grow trees: swor
    Resample size used to grow trees: 73
                            Analysis: RSF
                              Family: surv
                      Splitting rule: logrank *random*
       Number of random split points: 10
                           (OOB) CRPS: 0.0036534
   (OOB) Requested performance error: 0.02672266
```

Ilustración 34: Resultados algoritmo Random Survival Forest tras imputación con MICE

El índice C calcula la concordancia con la supervivencia observada y en este caso es del 97,32%, con lo que la capacidad predictiva es muy alta. El índice C

mide la probabilidad de que las predicciones sean concordantes con sus resultados para un dato dado durante el período de tiempo.

```
[1] 0.9732773
```

Ilustración 35: Índice C

7.2 Algoritmo Random Survival Forest con base de datos imputada con randomForestSRC

Se aplica el algoritmo con los mismos parámetros a la base de datos imputada con randomForestSRC y el índice C es de 96,27%, un resultado que muestra también una concordancia muy alta.

```
                         Sample size: 116
                  Number of deaths: 99
                   Number of trees: 500
         Forest terminal node size: 2
      Average no. of terminal nodes: 6.162
No. of variables tried at each split: 45
             Total no. of variables: 45
      Resampling used to grow trees: swor
   Resample size used to grow trees: 73
                           Analysis: RSF
                             Family: surv
                   Splitting rule: logrank *random*
      Number of random split points: 10
                         (OOB) CRPS: 0.003321
  (OOB) Requested performance error: 0.03726661
```

Ilustración 36: Resultados algoritmo Random Survival Forest tras imputación con randomForestSRC

```
[1] 0.9627334
```

Ilustración 37: Índice C

8. Conclusiones

Las conclusiones principales a las que se llega con los resultados obtenidos en este TFM son las siguientes:

- Los pacientes tratados con la rama A, la experimental: Abacavir 600 mg + Lamivudina 300 mg + Efavirenz 600 mg tiene una probabilidad de fracaso virológico igual o menor que aquellas tratadas con la rama B, la rama control: Abacavir 600 mg+ Lamivudina 300 mg + Lopinavir/ Ritonavir 400/100 mg sin embargo por su posología, la facilidad de

adherencia al tratamiento de los pacientes es mucho mejor en la rama A, 2 comprimidos al día que en la rama B, 7 comprimidos al día.

- En la base de datos obtenida de la imputación con mice, las pacientes tratadas con la rama experimental tienen una probabilidad del 51,7% de no tener fracaso virológico a las 48 semanas de tratamiento, mientras que las pacientes de la rama control, tienen una probabilidad de no tener fracaso virológico del 41,4%. Por lo tanto, la probabilidad de tener fracaso virológico a las 48 semanas de tratamiento es mayor en las pacientes tratadas con la rama control, pero estas diferencias no son significativas.

- En la base de datos obtenida de la imputación con randomForestSRC, las pacientes tratadas con la rama experimental tienen una probabilidad del 20,7% de no tener fracaso virológico a las 48 semanas de tratamiento, mientras que las pacientes de la rama control, tienen una probabilidad de no tener fracaso virológico del 8,6%. Por lo tanto, la probabilidad de tener fracaso virológico a las 48 semanas de tratamiento es mayor en las pacientes tratadas con la rama control y las diferencias obtenidas sí son significativas.

- Las predicciones realizadas por el algoritmo random survival forest en las bases de datos con ambas imputaciones consiguen un índice C en torno al 97%, con lo que la probabilidad de concordancia es muy alta.

Además de estas conclusiones, durante la realización de este TFM, se han llegado a otras, algunas de ellas son la base de las líneas de trabajo futuro:

- Existen una gran cantidad de métodos de análisis de supervivencia

- Es fundamental saber el patrón de creación de datos faltantes antes de decidir la estrategia a seguir en su tratamiento.

- Para aplicar los métodos de imputación se ajustan numerosos parámetros, un profundo estudio de dichos parámetros ayudará en optimizar los resultados.

Se han logrado todos los objetivos planteados, tanto los generales como los específicos. En la primera parte del desarrollo del trabajo, se detallaron más y se hizo una modificación, inicialmente se había planteado hacer la comparación de tres métodos de imputación y compararlos aplicando un algoritmo de machine learning pero a medida que se revisó bibliografía sobre análisis de supervivencia, se vio que podría ser todavía más interesante hacer una comparación tanto con un método convencional, como es el estimador de Kaplan Meier como un algoritmo de machine learning por lo que se decidió en lugar de imputar con tres métodos de imputación hacerlo con dos y en su lugar incorporar el análisis de supervivencia con Kaplan Meier.

Se siguió la planificación inicial, pero en cada una de las entregas del desarrollo del trabajo se revisó y se hicieron los ajustes oportunos. Fue muy útil haber hecho un análisis de los riesgos y establecer como posible riesgo, quedarse bloqueado en alguno de las tareas previstas, esto ocurrió con la primera de las imputaciones previstas, que era Hmisc, el plan de contingencia era cuando ya no se viese avance, saltar dicha tarea y realizar una alternativa, así se hizo y gracias a eso el TFM no resultó comprometido. La metodología prevista fue la adecuada, los plazos de cada una de las entregas de las PECs hacen que el ritmo del TFM sea exigente. Marcar como hitos en la planificación dichas entregas fue un acierto. Una dificultad a la hora de planificar el TFM es el hecho de que, una vez identificadas las tareas a realizar, asignarles un tiempo dentro de la planificación no resultó sencillo ya que al ser la primera vez que se realizan, es complicado cuantificar la duración de estas. Ayudó mucho tener claro los hitos, que eran las entregas de las PECs, con lo que los pequeños descuadres en la planificación se iban compensando unos con otros. También fue de gran ayuda incorporar un margen para imprevistos antes de la segunda entrega del desarrollo del trabajo ya que fue necesario utilizarlo finalmente para poder completar todos los objetivos planeados.

Las líneas de trabajo futuro que no se han podido explorar en este TFM y han quedado pendiente son:

- Estudiar los mecanismos de pérdida de datos faltantes de la base de datos utilizada. Se partió del supuesto que no había datos faltantes que sigan el patrón MNAR, pero sería muy interesante estudiarlo con las opciones disponibles en R.

- Realizar la imputación con otras librerías disponibles: Amelia, Mi, DMwR, Hmisc.

- Realizar regresión de Cox para estudiar la relación de algunas de las variables con los resultados del análisis de supervivencia.

- Aplicar algún otro algoritmo disponible para análisis de supervivencia.

- Realizar un tutorial de imputación.

9. Glosario

Listado de acrónimos utilizados durante la memoria por orden de aparición. Los términos de interés han sido definidos en la memoria.

TFM: trabajo final de máster
UOC: universitat oberta de catalunya
MICE: multiple imputation by chained equations
VIH: virus de inmunodeficiencia humana
PEC: prueba de evaluación continuada
ECA: ensayo clínico aleatorizado
RSF: random survival forest
RF: random forest
AUC: area under the curve
MCAR: missing completely at random
MAR: missing at random
MNAR: missing not at random
SIDA: síndrome de inmunodeficiencia adquirida
TARGA: tratamiento antirretroviral de gran actividad
ARN: ácido ribonucleico
ITIN: inhibidores de la transcriptasa inversa análogos de nucleósido
ITINN: inhibidores de la transcriptasa inversa no análogos de nucleósido
QD: una vez al día
BID: dos veces al día

10. Bibliografía

[1] Clotet, B y Fuster, D. 2005. Protocolo ensayo clínico multicéntrico, abierto, prospectivo, aleatorizado para evaluar la efectividad de abacavir 600 mg+ lamivudina 300 mg en pauta qd+ efavirenz 600 mg qd versus kaletra 400/100 mg bid como tratamiento antirretroviral de Inicio, p. 13-15.

[2] Marinelli, M. 2011. Missing Data in Clinical trials. [Master´s degree thesis, Universitat Politècnica de Catalunya and Universitat de Barcelona]

[3] Delgado, R. Características virológicas del VIH, Enferm Infecc Microbiol Clin. 2011;29(1):58–65

[4] James D. Dziuraa, Lori A. Strategies for dealing with Missing data in clinical trials: From design to Analysis, Yale journal of biology and medicine 86 (2013), pp.343-358.

[5] Wang P, Li Y, Reddy C. Machine learning for survival analysis: A survey. 2017. Article 1.

[6] https://www.emyaccion.com/emyaccion-articles/importancia-de-los-ensayos-clinicos/

[7] https://www.bitrix24.es/about/blogs/desarrollo-de-negocios/las-mejores-prcticas-del-diagrama-de-gantt-en-la-gestin-de-proyectos.php

[8] https://campus.uoc.edu/tren/trenacc/web/GAT_EXP.PLANDOCENTE?any_academico=20211&cod_asignatura=M0.167&idioma=CAS&pagina=PD_PREV_SECRE&cache=S

[9] https://www.rdocumentation.org/packages/Hmisc/versions/4.5-0

[10] Tang F, Ishwaran H. Random Forest Missing Data Algorithms. Stat Anal Data Min. 2017 Dec;10(6):363-377. doi: 10.1002/sam.11348. Epub 2017 Jun 13. PMID: 29403567; PMCID: PMC5796790.

[11] Austin PC, White IR, Lee DS, van Buuren S. Missing Data in Clinical Research: A Tutorial on Multiple Imputation. Can J Cardiol. 2021 Sep;37(9):1322-1331. doi: 10.1016/j.cjca.2020.11.010. Epub 2020 Dec 1. PMID: 33276049.

[12] https://www.isciii.es/InformacionCiudadanos/DivulgacionCulturaCientifica/DivulgacionISCIII/Paginas/Divulgacion/TiposdeEstudiosClinicos.aspx (Fecha acceso: 28/10/2021)

[13] https://www.conprueba.es/sites/default/files/informes/2020-06/TIPOLOG%C3%8DA%20DE%20ESTUDIOS%20CL%C3%8DNICOS_1.pdf (Fecha acceso: 28/10/2021)

[14] https://www.cancer.gov/espanol/publicaciones/diccionarios/diccionario-cancer (Fecha acceso: 28/10/2021)

[15] Métodos de investigación clínica y epidemiología, J.M. Argimon, J. Jiménez Villa., 4a ed., Elsevier, Barcelona, Espana˜ (2013). 402 p. ISBN: 978-84-8086-941-6

[16] Delgado M, Llorca J. Estudios longitudinales: concepto y particularidades. Rev Esp Salud Pública 2004; 78: 141-148. N.º 2 - Marzo-Abril 2004

[17] https://www.flsida.org/es/vih-sida (Fecha acceso: 30/10/2021)

[18] https://www.unaids.org/es/resources/fact-sheet. (Fecha acceso: 30/10/2021)

[19] https://www.fisterra.com/mbe/investiga/supervivencia/analisis_supervivencia2.pdf (Fecha acceso: 13/11/2021)

[20] Rebasa, P. Conceptos básicos del análisis de supervivencia. Cir Esp. 2005;78(4):222-30

[21] https://rpubs.com/ydmarinb/429757 (Fecha acceso: 20/11/2021)

[22] https://rstudio-pubs-static.s3.amazonaws.com/375297_34390ade0ddb4dd2bbe3bf1abf884dfe.html (Fecha acceso: 21/11/2021)

[23] https://rpubs.com/ItzelEscutia/524803 (Fecha acceso: 26/11/2021)

[24] J Martin Bland, Douglas G Altman. The logrank test. BMJ VOLUME 328 1 MAY 2004.

[25] https://cran.r-project.org/web/packages/survival/survival.pdf (Fecha acceso: 03/12/2021)

[26] https://cran.r-project.org/web/packages/randomForestSRC/randomForestSRC.pdf (Fecha acceso: 06/12/2021)

[27] https://www.researchgate.net/figure/Graphical-representations-of-a-missing-completely-at-random-MCAR-b-missing-at_fig3_312155483 (Fecha acceso: 12/12/2021)

[28] https://eprints.ucm.es/id/eprint/43961/1/TFM_PlanchueloGomez.pdf (Fecha acceso: 12/10/2021)

[29] http://openaccess.uoc.edu/webapps/o2/handle/10609/127626 (Fecha acceso: 20/10/2021)

[30] https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5796790/ (Fecha acceso: 09/11/2021)

[31] https://iagolast.github.io/blog/2019/01/13/kaplan-meier.html (Fecha acceso: 15/11/2021)

[32] https://stefvanbuuren.name/fimd/ (Fecha acceso: 10/10/2021)

11. Anexos

Anexo 1: pardo_beatriz_TMF_codigo_final_anexo1 (en pdf y rmd)