Missing data analysis in longitudinal data. How to analyze it?

Jorge J. Curto García
M0.185 - TFM-Estadística y Bioinformática
Máster Universitario en Bioinformática y Bioestadística UOC-UB

Nombre Profesor/a: **Núria Pérez**

Barcelona, 10 de Enero 2018

AGRADECIMEINTOS

A Nuria Pérez por su ayuda y soporte durante los últimos meses.
A mis padres y mi hermana porque siempre están.
A Gemma por todo. Sin tu apoyo no hubiera sido posible. Gracias.

FICHA DEL TRABAJO FINAL

Título del trabajo:	*Missing data analysis in longitudinal data. How to analyze it?*
Nombre del autor:	*Jorge Juan Curto García*
Nombre del consultor/a:	*Nuria Pérez Álvarez*
Nombre del PRA:	*Alexandre Sánchez Pla*
Fecha de entrega (mm/aaaa):	*01/2018*
Titulación:	*Máster Universitario en Bioinformática y Bioestadística UOC-UB*
Área del Trabajo Final:	*Bioestadística*
Idioma del trabajo:	*Castellano*
Palabras clave	*Missing data, longitudinal data, R, coarsened data[a], methodology.*

Resumen del Trabajo (máximo 250 palabras):

Mediante este trabajo se pretende caracterizar los estudios con datos longitudinales y los problemas derivados de los análisis en los que se presentan datos faltantes. Apoyándose en los grandes avances en la capacidad computacional que permiten la aplicación de algoritmos más complejos, en los últimos años se han desarrollado nuevos métodos de tratamiento de datos faltantes en el contexto del análisis de datos longitudinales. Se pretende indagar en los distintos tipos de datos faltantes y en la metodología disponible para abordar su análisis en el ámbito de datos longitudinales, para identificar bondades y limitaciones de dichos métodos. En la fase final del trabajo se presentará una ejemplificación de la aplicación de los métodos estudiados mediante el análisis de una base de datos longitudinales en el ámbito de la biomedicina, generando un informe estadístico dinámico (utilizando software de licencia libre: R y Markdown).

Abstract (in English, 250 words or less):

In this work, we intend to characterize the studies with longitudinal data and the problems derived from the analyzes in which missing data are presented. In recent years, based on the great advances in computational capacity that allow the application of more complex algorithms, there have been developed new methods of processing missing data in the context of longitudinal data analysis. The aim of this work is to investigate the different types of missing data and the available methodology to address their analysis in the longitudinal data field, in order to identify benefits and limitations of these methods. In the final phase of

[a] *Datos gruesos*

the work, an exemplification of the application of the methods studied will be presented through the analysis of a longitudinal database in the field of biomedicine, generating a dynamic statistical report (using free license software: R and Markdown).

Índice de contenidos

Lista de figuras:

Lista de tablas:

1.Introducción

1.1.Contextualización y justificación del trabajo

1.1.1.Descripción general

En este trabajo se pretende caracterizar los estudios con datos longitudinales y los problemas derivados de los análisis en los que se presentan datos faltantes. Se indagará en los distintos tipos de datos faltantes y en la metodología disponible[(2)] para abordar su análisis en el ámbito de datos longitudinales, para tratar de identificar bondades y limitaciones de dichos métodos. Se realizará una búsqueda de una base de datos reales, que permita ejemplificar la aplicabilidad de las técnicas de tratamiento de los datos faltantes.

1.1.2.Justificación del TFM

Uno de los principales focos de dificultades en el análisis de datos longitudinales [(3)] es la presencia de valores perdidos o datos faltantes (missing data) [(4)]. Apoyándose en los grandes avances en la capacidad computacional que permiten la aplicación de algoritmos más complejos, en los últimos años se han desarrollado nuevos métodos de tratamiento de datos faltantes en el contexto del análisis de datos longitudinales[(1)]. Mediante este trabajo se pretende hacer un balance de los métodos de reciente desarrollo y abordar un análisis de datos reales para ejemplificar la aplicación de dichos métodos en comparación con la sustitución o imputación de datos faltantes por valores plausibles como la media de la variable para todas las observaciones disponibles o la predicción obtenida a partir de la regresión del resto de variables para un individuo determinado [(2)].

Siempre que sea posible todos los análisis estadísticos se llevarán a cabo utilizando el lenguaje de programación R. La elección de R se justifica por el continuo crecimiento en la aplicabilidad y disponibilidad de las distintas técnicas estadísticas a través de las librerías accesibles de manera gratuita en la red, y que son constantemente revisadas y actualizadas por investigadores especializados.

La temática del presente TFM aplicada a datos reales se considera de suma importancia para la capacitación de los investigadores que se dedican al análisis de datos longitudinales y permitirá a su vez la consolidación de los conocimientos en bioestadística y bioinformática adquiridos durante el máster.

1.2.Objetivos

1.2.1. Objetivo general

Contextualizar el problema que generan los datos perdidos en el análisis de datos longitudinales y describir los métodos actuales disponibles para abordar dicho problema.

1.2.2.Objetivos específicos

1. Caracterización de los estudios con datos longitudinales.
2. Definición de datos faltantes
3. Contextualización del problema de los datos perdidos en el análisis de los datos longitudinales.
4. Identificación los métodos actuales de tratamiento de datos perdidos y determinar las bondades y limitaciones de estos métodos.
5. Ejemplificación de la aplicación de los métodos estudiados mediante el análisis de una base de datos longitudinales en el ámbito de la biomedicina.

1.3.Enfoque

Este TFM contiene dos partes diferenciadas: una primera ***sección teórica*** que engloba los 3 primeros objetivos y una segunda ***sección práctica*** (objetivo 4) donde se pretende aplicar los conocimientos adquiridos en la primera parte del TFM.

Se iniciará el trabajo llevando a cabo una recopilación de bibliografía acerca del análisis de datos longitudinales y la problemática que se genera con la presencia de datos faltantes. Mediante una revisión exhaustiva de la bibliografía disponible se pretende identificar los distintos métodos actuales de tratamiento de los datos faltantes, así como las distintas librerías de R que permitan el tratamiento de los mismos en el análisis de datos longitudinales.

Con el fin de llevar a cabo la sección práctica del trabajo, se contactará con expertos de diversos campos de investigación biomédica para tratar de obtener acceso a una base de datos longitudinales que se pueda utilizar para la ejemplificación de la aplicación de los métodos estudiados. En caso de no disponer de una base de datos a través de dichos expertos, se realizará una búsqueda online de bases de datos de acceso libre.

Se implementará un análisis descriptivo de los datos y un análisis de los tipos de datos faltantes presentes en la base de datos. Se generará un informe estadístico dinámico (utilizando R y Markdown) que contendrá aquellos métodos de tratamiento de datos faltantes identificados previamente y que sean aplicables a los datos contenidos en la base de datos disponible. El informe estadístico se incluirá como anexo de la memoria final.

1.4.Planificación temporal con hitos y temporalización

1.4.1.Tareas

Objetivo 1: Caracterización de los estudios con datos longitudinales.

- Definición de datos longitudinales.
- Métodos de análisis de datos longitudinales.

Objetivo 2: Contextualización del problema de los datos perdidos en el análisis de los datos longitudinales mediante una búsqueda bibliográfica exhaustiva.

- Definición de datos perdidos.
- Tipos de datos perdidos.
- Presencia de datos perdidos en el análisis de datos longitudinales.

Objetivo 3: Identificar los métodos actuales de tratamiento de datos perdidos, determinando las bondades y limitaciones de estos métodos (también mediante búsqueda bibliográfica exhaustiva).

- Métodos de tratamiento de datos perdidos en el análisis de datos longitudinales.
- Bondades y limitaciones de los métodos actuales.
- Identificación de las librerías de R disponibles para el tratamiento de datos perdidos.

Objetivo 4: Ejemplificación de la aplicación de los métodos estudiados mediante el análisis de una base de datos biomédicos.

- Búsqueda de una base de datos biomédicos.
- Análisis descriptivo de los datos mediante R.
- Identificación del tipo de datos faltantes.
- Aplicación de los métodos de tratamiento de datos perdidos disponibles mediante la utilización de librerías de R.
- Generación de un informe de resultados estadísticos con R y Markdown.

1.4.2.Calendario

	PEC 1 (2 semanas)		PEC 2 Desarrollo del trabajo Fase I (5 semanas)					PEC 3 Desarrollo del trabajo Fase II (4 semanas)			Memoria y presentación del trabajo final (2 semanas)		Elaboración de la presentación (1 semana)	Defensa pública (1 semana y 1/2)	
	3-16/10/2017		17-23/10	24-30/10	31-6/11	7-13/11	14-20/11	21-27/10	28/11-4/12	5-18/12	19/12-02/01		03/01-10/01	11/01-22/01	
Plan de trabajo															
Objetivo 1: Definición de datos longitudinales															
Objetivo1: Métodos de análisis de datos longitudinales															
Objetivo 2: Definición de datos perdidos															
Objetivo 2: Tipos de datos perdidos															
Objetivo 2: Presencia de datos perdidos en el análisis de datos longitudinales															
Objetivo 3: Métodos de tratamiento de datos perdidos en el análisis de datos long.															
Objetivo 3: Bondades y limitaciones de los métodos actuales															
Objetivo 3: Librerías de R disponibles para el tratamiento de datos perdidos															
Objetivo 4: Búsqueda de una base de datos															
Objetivo 4: Análisis descriptivo de los datos mediante R															
Objetivo 4: Identificación del tipo de datos faltantes															
Objetivo 4: Apl. de los métodos de tratamiento de datos perdidos disponibles															
Redacción de la memoria y presentación															
Elaboración de la presentación															
Defensa pública- Tribunal TFM															

1.4.3.Hitos

Hito 1: Entrega del plan de trabajo **(16/10/2017)**

Hito 2: Datos longitudinales: Definición y métodos de análisis

Hito 3: Datos perdidos: Definición, tipos y presencia en datos longitudinales

Hito 4: Selección y preparación de la base de datos para análisis

Hito 5: Entrega PEC2 **(20/10/2017)**

Hito 6: Datos perdidos en datos longitudinales: Métodos, bondades y limitaciones

Hito 7: Análisis de datos y redacción de resultados

Hito 8: Entrega PEC3 **(18/12/2017)**

Hito 9: Entrega de la memoria y presentación **(02/01/2018)**

Hito 10: Elaboración de la presentación **(del 03/01/2018 al 10/01/2018)**

Hito 11: Defensa pública **(del 11/01/2018 al 22/01/2018)**

1.4.4. Análisis de riesgos

Son varios los factores que han supuesto un riesgo para no cumplir con la temporalización planteada:

- Software para el tratamiento de datos faltantes: Problemas derivados en el acceso a librerías de R, problemas con actualización de últimas versiones del software disponible y dificultades técnicas en su aplicabilidad a los datos analizados.

- Selección de la base de datos: Problemas derivados en el acceso a la base de datos seleccionada y complicaciones derivadas de la transformación de la base de datos a formato legible por R.

- Tiempo: Problemas de compatibilidad con el horario, carga laboral, viaje imprevisto o enfermedad personal o de familiar cercano. Retraso en el análisis por la complejidad de los métodos empleados y/o la falta de accesibilidad a los métodos para el análisis estadístico de los datos.

1.5.Resultados esperados

1.5.1.Plan de trabajo

Descripción detallada y temporalización de los objetivos, tareas que se llevarán a cabo durante el TFM.

1.5.2.Memoria

Presentación detallada por escrito de los resultados y conclusiones obtenidas durante el TFM. El documento final contendrá los siguientes apartados: Introducción, métodos, resultados, discusión, conclusiones y anexos.

1.5.3.Producto

Informe de resultados estadísticos generado con R y Markdown (se incluirá en la memoria como un anexo), basado sobre una base de datos reales.

1.5.4.Presentación en diapositivas y virtual

Diapositivas y vídeo de 20 minutos que contendrá una presentación oral resumiendo los resultados más interesantes y las conclusiones.

.

1.5.5.Estructuración del proyecto

1.5.5.1. PEC 1

Plan de trabajo

1.5.5.2. PEC 2

Se pretenden llevar a cabo las tareas correspondientes a los objetivos 1, 2 y 3 (secciones teóricas) y parte del objetivo 4 (sección práctica).

Objetivo 1: Caracterización de los estudios con datos longitudinales.

Objetivo 2: Contextualización del problema de los datos perdidos en el análisis de los datos longitudinales.

Objetivo 3: Identificación de los métodos actuales de tratamiento de datos perdidos, determinando las bondades y limitaciones de estos métodos.

Objetivo 4: Búsqueda y preparación de la base de datos.

1.5.5.3. PEC 3

Objetivo 4 (sección práctica): Ejemplificación de la aplicación de los métodos estudiados mediante el análisis de una base de datos reales con el software estadístico R.

1.5.5.4. PEC 4: Memoria

Presentación detallada por escrito de los resultados y conclusiones obtenidas durante el TFM.

1.5.5.5. PEC 5: Presentación en diapositivas y virtual

Diapositivas y vídeo de 20 minutos que contendrá una presentación oral resumiendo los resultados más interesantes y las conclusiones.

2.Estudios con datos longitudinales

2.1.Definición de datos longitudinales.

En muchos estudios médicos el interés de los investigadores se basa en la evolución temporal de una determinada variable, más allá de la observación de la variable en un instante dado[3]. El interés de los investigadores puede estar en el cambio que se produce en una variable determinada como el peso en neonatos durante las primeras semanas de vida o la evolución de un marcador como los CD4, en pacientes VIH+, durante los meses siguientes al inicio de un tratamiento con medicamentos antiretrovirales. Cuando este es el caso y la variable, o variables de interés, se recogen para cada individuo en más de dos mediciones repetidas a lo largo del tiempo, obtendremos un conjunto de datos longitudinales[5,6]. Precisamente es el "tiempo" el principal recurso utilizado en los estudios de datos longitudinales, midiendo los potenciales cambios de las variables en función de los posibles factores que hayan estado actuando durante dicho tiempo y con la condición necesaria de que los individuos observados sean los mismos durante todo el seguimiento[7]. Cabe destacar que hay datos longitudinales que no se corresponden a una escala temporal, por ejemplo si hay un eje que ordene los datos mediante una distancia.

Los datos longitudinales pueden presentarse tanto en estudios observacionales (neonatos y su peso) como en estudios con intervención (pacientes VIH+ y tratamiento antirretroviral) [3, 4,10]. Los datos se pueden recoger tanto de manera prospectiva como retrospectiva y hay tres elementos claves que caracterizan un estudio de datos longitudinales[4]: el seguimiento, más de dos medidas y un análisis que las tenga en cuenta.

Existe una amplia variabilidad entre los estudios en la temporalización y el número de observaciones para cada individuo[3]. Los estudios que recogen este tipo de datos, generalmente, se diseñan de manera que las mediciones de las variables para cada individuo se lleven a cabo en los mismos intervalos temporales. Aunque no siempre es el caso, para ciertas patologías, los cambios en el marcador que mide la respuesta pueden ser más sensibles al inicio de un tratamiento y es conveniente que el tiempo entre las mediciones iniciales sea menor que en las mediciones posteriores.

2.2.Objetivos de los estudios de datos longitudinales.

El diseño longitudinal se define como un procedimiento cuyo objetivo es analizar los "patrones interindividuales de cambio intraindividual"[15]. Se entiende por variabilidad interindividual (o diferencias entre individuos) la diversidad que presentan los individuos de una misma población en un momento dado (corte transversal) y estas diferencias deben ser consistentes a través de las situaciones y temporalmente estables. En cuanto a los cambios intraindividuales engloban las diferencias que puedan darse en un sujeto bajo las mismas circunstancias en periodos de tiempo. En el análisis de estudios longitudinales los objetivos deberán incluir[11]:

a) estudio directo de cambio intraindividual

b) la identificación directa de las diferencias interindividuales en el cambio intraindividual,

c) el análisis de la relación entre los cambios intra e interindividuales y

d) el estudio de las variables que influyen en el cambio intra e intraindividual.

Por lo tanto se analiza el cambio en función del tiempo y para ello se recogen datos longitudinales de una muestra dada de sujetos que es medida en sucesivas ocasiones en la misma variable respuesta[16].

2.3.Métodos de análisis de datos longitudinales.

Si la metodología de análisis clásico fuese aplicada al análisis de datos longitudinales presentaría limitaciones. Las distintas mediciones de un mismo individuo recogidas de manera correlativa suelen presentar correlación y por lo tanto requieren de un tratamiento específico que permita realizar inferencias válidas.

Los métodos de análisis de datos longitudinales se engloban dentro del contexto de los modelos lineales generalizados. En ellos se utilizan las herramientas convencionales de regresión para relacionar el efecto con las diferentes exposiciones y se tiene en cuenta la correlación de las medidas entre sujetos[4]. Mediante estos modelos es posible tratar covariables dependientes del tiempo que pueden a la vez influir sobre la exposición en estudio y ser influidas por ella (variables que se comportan simultáneamente como confundidoras e intermedias entre exposición y efecto).

La metodología para el análisis de datos longitudinales difiere en función de la naturaleza de la variable respuesta. En el caso en que la variable de interés fuese cuantitativa siguiendo una distribución normal, se podrán utilizar técnicas de análisis multivariante, análisis de la variancia de medidas repetidas, análisis de curvas de crecimiento, modelos de efectos mixtos o modelos de ecuaciones de estimación generalizada (GEE). Por otro lado, en los casos en que la variable respuesta es cualitativa se pueden utilizar modelos mixtos generalizados, modelos log-lineales y los basados en GEE[11]. Conviene resaltar que en la mayoría de los métodos mencionados las estimaciones de los coeficientes se basan en el método de máxima verosimilitud, a excepción de los métodos basados en GEE que basan en estimaciones por pseudo verosimilitud.

3.Datos perdidos.

3.1.Definición de datos perdidos.

La calidad de los datos recogidos es esencial a la hora de poder llevar a cabo el análisis estadístico de los mismos y a menudo los investigadores se enfrentan con diversos problemas en el seguimiento de los pacientes.

Uno de los principales focos de dificultades en el análisis de datos longitudinales es la presencia de valores perdidos o datos faltantes (missing data)[4]. Los valores perdidos pueden afectar desde algunas de las variables del estudio y algunos de los individuos de la muestra hasta la totalidad de los datos de algunos de los individuos en la muestra[8] y los motivos por los que se producen son muy diversos, desde la falta de consentimiento por parte del paciente que permita al investigador acceder a un determinado resultado analítico, pasando por un problema técnico a la hora de analizar una determinada muestra o la no asistencia del paciente a una visita programada.

3.2.Tipos de datos perdidos.

Cuando se lleva a cabo una investigación, la situación ideal es obtener una base de datos completa. Pero esto no es lo habitual y la no respuesta (datos perdidos) pueden presentarse en dos formas[17]:

La no respuesta total: datos perdidos en todas las variables de un registro (individuo o sujeto de análisis).

La no respuesta parcial: datos perdidos en una o más variables sin llegar a la ausencia completa.

Entendiendo que los datos perdidos aparecen en las muestras por razones fuera del control del investigador, antes de tomar una decisión acerca de su presencia e influencia en los análisis posteriores, es imprescindible establecer los supuestos sobre los procesos que hayan generado dichos datos perdidos[2,12]. Rubin en su trabajo de 1976[12], establece tres mecanismos de pérdidas de datos en función de la relación entre la probabilidad de datos faltantes y las variables recogidas:

- Datos perdidos completamente al azar (***MCAR*** = missing completely at random): la probabilidad de que un individuo presente un valor perdido en una variable no depende ni de otras variables de la muestra ni de los valores de la propia variable con valores perdidos, es decir que las características de los individuos con información son similares a las de los individuos sin información. Un ejemplo de MCAR sería: *Se asume que el colesterol medio de los individuos para los que no se dispone de la variable colesterol, sería similar al de los individuos para los que si se dispone del dato. No habría diferencias en el resto de las variables entre los individuos para los que se dispone y no del colesterol.*

- Datos perdidos al azar (***MAR***= missing at random): la probabilidad de que un valor no se observe depende de los valores de los datos observados, pero no de los faltantes. Un ejemplo de MAR sería: *La pérdida de valores en la variable colesterol depende de la dieta de los pacientes, pero dentro de cada tipo de dieta, la probabilidad de ser un dato perdido no se relaciona con los valores del colesterol.* No es posible probar si la condición MAR se satisface ya que al desconocer la información faltante en la variable de interés, no se pueden hacer comparaciones entre los sujetos con y sin valores faltantes en dicha variable.

- Datos perdidos no ignorables o no debidos al azar (***MNI***=missing non-ignorable, o ***MNAR***=missing not at random): la probabilidad de que un valor no se observe en una variable depende del verdadero valor del dato perdido o de variables no observables. Este caso es el más habitual y a su vez el más difícil de modelizar[2]. Un ejemplo de MNAR sería: *Ajustando por el resto de variables disponibles, el colesterol medio de los individuos con datos faltantes es más alto que el del resto de los individuos.*

Para estudios en los que se clasifiquen los datos perdidos como MCAR o MAR, los mecanismos de las pérdidas se pueden considerar como pérdidas 'ignorables', y los métodos de verosimilitud e imputación múltiple inferirán correctamente los resultados sin necesidad de modelar dichas pérdidas. Aunque en la práctica, es difícil justificar la asunción de independencia de MCAR o MAR y en muchas ocasiones MNAR deberá ser considerado, teniendo en cuenta que no ocurren de manera aleatoria, sino que siguen un patrón, que será necesario analizar antes de tomar una decisión acerca de una posible imputación de valores [17,18].

Datos gruesos (coarsening): En ocasiones, aunque como ya se ha comentado la situación ideal es obtener una base de datos completa, debido a la idiosincrasia tanto de los individuos que participan en el estudio como de la información a recoger, no es posible obtener los datos exactos (en el sentido de precisos) de una variable. Este podría ser el caso en que las variables se recojan categorizadas en intervalos, por ejemplo, el salario que cobran una serie de individuos: >50000€, 50000-40001€, 40000-30001€, 30000-20001€, ≤20000€. Otro caso de pérdida de información por *datos gruesos* se produce cuando las variables se redondean, por ejemplo en el caso en que se recogen las edades de niños, que puede redondearse a años si tienen más de tres años de edad, pero para los niños más pequeños, se podría obtener en fracciones de un año (mitades o cuartos), y para niños menores de un año, se podría especificar en meses o incluso semanas[19].

3.3.Presencia de datos perdidos en datos longitudinales

La presencia de valores perdidos pone al estadístico frente la disyuntiva de si los datos disponibles para un individuo se pueden utilizar o no en un análisis

determinado. Algunos métodos estadísticos aplicables a datos longitudinales como el análisis de varianza con medidas repetidas[9, 10,11], presentan la limitación de tener en cuenta solo los individuos con las mismas mediciones a lo largo del tiempo, por lo que aquellos individuos con algún dato faltante serían excluidos del análisis. La reducción del número total de casos o de una o más variables afectaría a la precisión de los resultados y los posibles sesgos debidos a la posibilidad de pérdida de valores en sujetos con características comunes influenciarían en la validez del estudio[8]. Aunque ya hay autores importantes que desde hace varias décadas comenzaron a investigar métodos más avanzados[4,18] para hacer frente a dichos problemas, hasta hace relativamente poco tiempo, la sustitución o imputación de datos faltantes por valores plausibles como la media de la variable para todas las observaciones disponibles o la predicción obtenida a partir de la regresión del resto de variables para un individuo determinado[1], eran los métodos más utilizados.

4.Tratamiento de datos perdidos en el análisis de datos longitudinales.

4.1.Métodos de tratamiento de datos perdidos en el análisis de datos longitudinales.

El tratamiento de los datos perdidos tiene como fin el reducir su posible influencia en las variables exploradas dentro de cualquier tipo de análisis. En el desarrollo de un estudio de investigación (con datos longitudinales o no) hay tres momentos fundamentales en los que deben ser tenidos en cuenta[8]:

1)Diseño del estudio: previendo que el tamaño muestral no se alcance durante el reclutamiento o que haya pérdidas de datos o sujetos durante el seguimiento, se debe añadir una proporción variable (usualmente 10-15%) al tamaño muestral requerido.

2)Recogida de datos: una adecuada monitorización de los datos debe incrementar la calidad de los mismos permitiendo la recuperación de posibles datos perdidos.

3)Análisis de los datos: Se deberá caracterizar el número de casos afectados en las variables de interés y el tipo de valores faltantes, para analizar el posible efecto que estos tengan sobre la precisión y validez del análisis final. Sería interesante poder disponer del motivo de la pérdida de datos. En este punto se decide el tipo de tratamiento de los datos perdidos previa a la estimación de los estadísticos que den respuesta a los objetivos del estudio.

Los métodos de tratamiento de datos perdidos han ido evolucionando desde la primera publicación de Wilks[20] en 1932, en la que proponía la ***sustitución de los datos faltantes en una variable por la media*** del resto de datos disponibles en dicha variable. Utilizando la clasificación de Graham[21], este

método se engloba, en los ***métodos antiguos.*** A continuación presentamos una clasificación que se subdivide a su vez en dos categorías: **Eliminación e imputación simple**[22].

Métodos de eliminación:

- ***Eliminación de los casos (listwise)***: eliminando del análisis todos los casos con al menos un dato faltante, también se denomina ***análisis de casos completos.***
- ***Eliminación por pares (pairwise)***: eliminando el sujeto solo en los análisis para los que se utilice la variable con datos faltantes pero manteniéndolo en el resto de análisis. Este tipo de análisis también se denomina de ***Casos Disponibles***.

Métodos de imputación simple:

- ***Imputación de la media (o imputación de la media incondicional):*** se reemplazan los valores perdidos en una variable por el valor de la media del resto de valores disponibles en dicha variable.
- ***Imputación por regresión simple (media condicional)***: se reemplazan los valores perdidos por el valor predicho a partir de la aplicación de la regresión simple sobre el resto de datos conocidos en el sujeto.
- ***Imputación por el vecino más cercano***: se identifica la distancia entre la variable a imputar y cada una de las unidades restantes (*x* o variables auxiliares) mediante alguna medida de distancia, entonces se determina la unidad más cercana a *y*, usando el valor de esta unidad cercana para imputar el faltante.
- ***Imputación por regresión estocástica***: también se predicen los valores de las variables incompletas a partir de las variables completas pero añadiendo un paso extra a la predicción de los valores por medio de un término residual (error) que sigue una distribución normal. Al valor imputado se le añade un valor aleatorio, que se genera a partir de una distribución condicional con estimadores máximo verosímiles para la distribución desconocida de los parámetros basado en los datos perdidos[21].
- ***Imputación por la observación previa (LOCF Last Observation Carried Forward)***: aplicada en datos longitudinales, se imputa la observación faltante en un individuo y con el valor que le precede del mismo individuo.
- ***Imputación Hot-Deck***: Utilizado habitualmente en el ámbito de las encuestas poblacionales, los individuos son clasificados en factores a partir de sus características sociodemográficas y se imputa la observación faltante en un individuo a partir de los valores de individuos "similares"[22,23].

-***Imputación por el vecino más cercano***: se puede entender como una variante del método Hot-Deck. A partir de las características sociodemográficas y/u otros factores[22,24], se selecciona el individuo "más cercano" al individuo con un dato faltante y el dato correspondiente a dicho individuo más cercano es el imputado en lugar del dato faltante.

Los siguientes métodos se clasifican como ***métodos modernos***[21]:

-***Algoritmo de expectación-maximización (EM)***[22,25]: Este método está basado en la función de máxima verosimilitud y permite obtener estimaciones máximo verosímiles de los parámetros en el caso de datos incompletos con unas estructuras determinadas. Es un proceso iterativo, en el cual se repiten dos pasos en cada iteración: en el primer paso (*Expectación)* se imputan valores (mediante el uso de ecuaciones de regresión) y en el segundo paso (*Maximización)* los valores de media y la matriz de covarianza son calculados nuevamente pero utilizando los valores imputados y no perdidos. El proceso se repite hasta que las estimaciones convergen.

-***Algoritmo de imputación múltiple (MI)***[26-36]: Se divide en tres fases: imputación, análisis y combinación de los resultados. En la fase de reemplazamiento de los datos perdidos (o imputación), se generan *m* copias del conjunto de datos y cada una de ellas se imputan diferentes estimaciones de los valores perdidos. Existen varias formas para realizar la imputación siendo las estrategias más populares la sustitución de valores perdidos utilizando regresión de imputación (patrón de pérdidas univariado), el método de Markov Chain Monte Carlo (MCMC) cuando los datos perdidos son arbitrarios o el algoritmo Predictive Mean Matching[28,29] (Equiparación de media predictiva, PMM) especialmente efectivo para imputar variables cuantitativas cuando no siguen la distribución normal. Además a diferencia de otros algoritmos ajusta los valores imputados al rango (máximo y mínimo) de los valores de la variable e imputa correctamente los casos de variables discretas, siendo muy eficaz a la hora de mantener la variabilidad original de los datos. Esto se consigue porque el algoritmo "toma prestados" datos reales de casos para los que si se dispone del dato[28]. Mediante el método MI, cada una de las imputaciones genera un conjunto de datos diferentes los cuales se analizan por separado mediante técnicas estadísticas tradicionales, obteniéndose *m* estimaciones y sus errores estándar. Finalmente se combinan los *m* resultados obtenidos, y la estimación global es el promedio de las *m* estimaciones.

-***Fully Conditional Specification (FCS)***[30,31]: Mediante este método se genera un conjunto de datos completo para cada variable que presenta datos faltantes tras una fase de "aprendizaje" (*learning phase*), permitiendo la modelación de las interacciones entre variables. El

proceso se repite hasta obtener un conjunto de datos completos que son combinados y analizados como datos múltiples imputados.

-***Full Information Maximum Likelihood (FIML)***[31,34]: estima el logaritmo de verosimilitud para cada individuo basado en las variables presentes en el modelo, estimando como debería ser el modelo hipotético (parámetros y los errores estándar), en un solo paso, sin necesidad de conocer cómo serían realmente las respuestas perdidas. Este objetivo se logra mediante el uso de las respuestas observadas para complementar la pérdida de información debido a las respuestas faltantes.

En un artículo reciente, Roberts et al.[30] plantea la clasificación de los métodos para el tratamiento análisis de datos faltantes en métodos a nivel variable (eliminación de casos y por pares) y métodos analíticos (el resto).

Y a su vez, dentro de los métodos analíticos incluye la subdivisión entre métodos no-estocásticos o de imputación no aleatoria (todos los métodos de imputación simple excepto la regresión estocástica) y los métodos estocásticos (regresión estocástica, EM, MI, FCS FIML).

Finalmente existen otros métodos que no implican ni eliminación ni imputación, son aquellos métodos que usan modelos en los que se permite la presencia de datos perdidos como “weighed” GEE (válido para datos MCAR o MAR) o que modelan por un lado los datos faltantes y por otro los datos observados, como los modelos Pattern-Mixture.

4.2.Bondades y limitaciones de los métodos actuales.

Algunos métodos como el de ***análisis de casos completos*** son estrategias de simple aplicación a cualquier análisis estadístico con datos faltantes, pero que pueden generar dudas en cuanto a si la muestra a analizar, una vez eliminados los casos con observaciones faltantes, es una muestra aleatoria de la muestra original o está sesgada por la naturaleza de los datos perdidos. En la aplicación de los métodos de ***eliminación*** (casos completos y casos disponibles) se requiere asumir que el mecanismo de los datos perdidos es MCAR[22] y pese a la pérdida de eficacia (menor tamaño muestral y menor potencia) la muestra sería insesgada. Pero en el análisis de datos reales este es un caso poco habitual y generalmente el investigador se enfrenta a pérdidas MAR (o incluso MNAR) y en esos casos el análisis de las observaciones completas puede producir estimaciones sesgadas. En el caso específico de la eliminación por pares (casos disponibles), al involucrar dentro de un mismo estudio el análisis de submuestras con diferentes tamaños muestrales puede producir problemas con las medidas de asociación entre las variables[22].

Cuando se utiliza la ***sustitución de los datos faltantes en una variable por la media (incondicional),*** se produce una reducción de la varianza y además se distorsionan los valores de covarianzas y correlaciones, llegando a ser el método “*a evitar*” para algunos autores debido al sesgo que produce en los resultados[21,22].

En la aplicación de los métodos de ***imputación por regresión*** es posible la utilización de covariables para la estimación de los valores perdidos que aunque inicialmente disminuiría el sesgo también reduciría la variabilidad estadística[33].

En el caso de la imputación por LOCF se llevan a cabo asunciones no realistas acerca de las trayectorias temporales de los individuos[23], ya que se asume que los valores no cambian significativamente después de la última medida observada o intermitentemente en un periodo en que se hallan detectado valores faltantes. De modo que puede llevar a conclusiones erróneas acerca del comportamiento a lo largo del tiempo distorsionando diferencias entre grupos así como las estimaciones de los parámetros incluso en el caso de MCAR[22].

En todos los métodos no estocásticos (de imputación no aleatoria) se requiere asumir que el mecanismo de los datos perdidos es MCAR o MAR y en general introducen sesgo estadístico reduciendo la variabilidad; además se deben tener en cuenta problemas potenciales con los errores tipo I y tipo II si se decide utilizar este tipo de métodos[30].

En general en todos los métodos antiguos o convencionales [21,23,33,35], incluso en los casos en que se consigue evitar el sesgo en la estimación de los parámetros, se infraestiman los errores estándar[23]. Esto se debe a que los métodos de análisis asumen que los datos analizados son reales, pero el proceso de imputación introduce una variabilidad adicional en la muestra a través de los valores imputados, que no se tiene en cuenta de manera adecuada en dichos análisis[35].

Frente a las limitaciones en los métodos presentadas anteriormente y para aprovechar las bondades de los datos perdidos al azar, los métodos estocásticos generan aleatoriamente múltiples conjuntos de datos basados en los valores observados[28,34,36] y reducen potencialmente la posibilidad de que se produzca sesgo estadístico mientras que a su vez maximizan la variabilidad[21,30].

4.3. Identificación de las librerías de R disponibles para el tratamiento de datos perdidos.

Existen multitud de paquetes diferentes en R que implementan técnicas para el análisis de los datos faltantes, una recopilación bastante extensa se puede encontrar en: http://www.stefvanbuuren.nl/mi/Software.html (último acceso 28-12-2017). A continuación se detallan varios de ellos:

- ***MissingDataGUI***[37]***:*** un paquete que permite obtener resúmenes numéricos y gráficos para los datos perdidos de variables cuantitativas y categóricas. Contiene una variedad de métodos de imputación, incluyendo imputaciones univariadas (valores fijos o aleatorios), imputaciones multivariadas (por ejemplo: *"por el vecino más cercano")* e imputaciones múltiples. También incluye métodos que permiten hacer imputaciones condicionadas a una variable categórica.

- ***VIM***[38,39]***:*** Un paquete que también contiene herramientas para la visualización de los datos faltantes y/o imputados para ser utilizados en la exploración de los datos y la estructura de los datos faltantes y/o imputados. Dependiendo de la estructura de los valores faltantes, los métodos correspondientes pueden ayudar a identificar el mecanismo que genera los valores perdidos y permite explorar los datos incluyendo los valores faltantes. Además, la calidad de imputación puede ser explorada visualmente utilizando varios métodos gráficos univariantes, bivariantes y multivariantes. Un interfaz de usuario gráfico disponible en el paquete ***VIMGUI***[38] permite un fácil manejo de los métodos gráficos implementados.

- ***MICE***[40] *(Multivariate Imputation via Chained Equations*): Permite realizar imputaciones múltiples utilizando *Fully Conditional Specification* (FCS), implementado por el algoritmo MICE, por lo tanto cada variable tiene su propio modelo de imputación. Se incluyen modelos de imputación para datos continuos (predictive mean matching, normal), datos binarios (regresión logística), datos categóricos no ordenados (regresión logística politómica) y datos categóricos ordenados (odds proporcionales). MICE también puede imputar datos continuos de dos niveles (modelo normal, pan, variables de segundo nivel). La imputación pasiva puede utilizarse para mantener la coherencia entre las variables. Existen varios gráficos de diagnóstico para inspeccionar la calidad de las imputaciones.

- ***Amelia (Amelia II)***[41]: Este paquete imputa de forma múltiple datos faltantes en una sección cruzada única (como una encuesta), de una serie temporal (como variables recogidas anualmente en un país), o de serie temporal-cross-sectional (tales como variables recogidas a lo largo de varios años en varios países). Amelia II implementa un algoritmo basado en bootstrap y generalmente es considerablemente más rápido que otros métodos y puede manejar más variables. Incluye también herramientas de diagnóstico útiles para valorar el ajuste de los modelos de imputación múltiple.

- ***Hmisc***: Contiene muchas funciones útiles para el análisis de datos, gráficos de alto nivel, operaciones de utilidad, funciones para calcular el tamaño y la potencia de la muestra, importación y anotación de conjuntos de datos, imputación de valores perdidos, elaboración avanzada de tablas, agrupación de variables, manipulación de cadenas de caracteres, conversión de objetos R en LaTeX y código html y variables de recodificación.

  http://biostat.mc.vanderbilt.edu/wiki/Main/Hmisc (último acceso 28-12-2017)

- ***mi***[42]: Este paquete ofrece funciones para la manipulación de datos, imputación de valores faltantes en un marco aproximado bayesiano, diagnósticos de los modelos utilizados para generar las imputaciones, mecanismos de confianza para validar algunos de los supuestos del algoritmo de imputación y funciones para analizar conjuntos de datos

multiplicados imputados con el grado apropiado de incertidumbre de muestreo.

5.Ejemplificación de la aplicación de los métodos estudiados mediante el análisis de una base de datos biomédicos longitudinales.

El objetivo del presente TFM, es la aplicación de diferentes técnicas de imputación de valores perdidos en un conjunto de datos longitudinales reales utilizando el programa estadístico R[43].

Una limitación que surgió a la hora de decidir qué base de datos escoger para la realización del presente TFM es que en la mayoría de los casos los investigadores publican bases de datos completas. Finalmente se ha optado por seleccionar una base de datos con un número significativo de registros y se generaron los datos perdidos de manera aleatoria. De este modo se reformula el objetivo de esta sección y a modo de análisis de sensibilidad se plantea el tratar de comparar los resultados originales publicados (base de datos completa) con los resultados obtenidos tras haber generado distintos escenarios de datos faltantes y haber tratado dichos datos con varios de los métodos presentados anteriormente (ver sección 4). Los distintos escenarios de datos faltantes, se plantearon generando de manera aleatoria un 10%, 20% y 30% de valores perdidos en las variables principales de la base de datos original.

El resumen del plan de análisis de los resultados que se presentan a continuación es el siguiente:

- Selección, presentación y análisis descriptivo de la base de datos seleccionada.
- Generación de nuevas bases de datos (*dataframes* en R) con datos faltantes generados de manera aleatoria, 3 escenarios: 10%, 20% y 30%.
- Tratamiento de los datos faltantes para cada una de las bases de datos generadas anteriormente.
- Reproducción de los análisis originales publicados por los autores (base de datos completa).
- Reproducción de los análisis originales publicados por los autores con las bases de datos generadas tras el tratamiento de los datos faltantes.
- Comparación de los resultados de los dos pasos anteriores.

Todos los resultados que se presentan a continuación se pueden hallar detallados en el informe adjunto y que ha sido generado mediante RStudio (Versión 1.1.383 – © 2009-2017 RStudio, Inc.)[44], el paquete Knitr [45], trabajando bajo la versión 3.4.3 de R (RGui 64-bit). Además del informe final, se adjunta la sintaxis con formato Markdown[46] que permitiría a cualquier lector que esté interesado, generar dicho informe. Para los análisis estadísticos se fijó el nivel de significación en α=0.05.

5.1. Búsqueda de una base de datos biomédicos longitudinales.

Existen diversos repositorios online que recopilan bases de datos, uno de ellos es "**The Dryad Digital Repository" (**http://datadryad.org/ último acceso 28-12-2017), que es un recurso comisariado que permite acceder libremente a los datos utilizados en diversas publicaciones científicas. A través de **Dryad**, se obtuvo acceso a una base de datos correspondiente a un estudio longitudinal que incluye resultados del test de *Ruff Figural Fluency/ Test de Fluidez de diseños de Ruff (RFFT)* en 2515 participantes del estudio *Prevention of Renal and Vascular ENd-stage Disease (PREVEND)*, de la ciudad de Groningen, en los Países Bajos[47]. La cumplimentación del RFFT se recogió en tres ocasiones durante un periodo de seguimiento de 6 años. El RFFT es una prueba cognitiva que evalúa la función ejecutiva (FE) midiendo la programación visomotora y asociándose con la actividad frontal derecha[47]. La FE engloba habilidades cognitivas propias de la corteza prefrontal, entre los cuales se incluyen la anticipación, la elección de objetivos, la planeación, la selección de la conducta, la autorregulación, el autocontrol y el uso de la retroalimentación[48,49]. Además de los resultados del RFTT, la base de datos contiene la edad de los participantes en la primera cumplimentación del test, el género, el nivel de educación y el tiempo transcurrido entre cada cumplimentación del test. Estudios longitudinales han demostrado que las pruebas repetidas mejoran el rendimiento en RFFT dificultando la interpretación de los resultados de la prueba en el entorno clínico. Por este motivo los autores investigaron el efecto de las mediciones consecutivas sobre el rendimiento en el RFFT mediante la utilización de un modelo de regresión lineal multivariado en el que incluyeron la edad, género, nivel de educación y el término de interacción entre el número de medición consecutivo y la edad como variables independientes. El test RFFT consiste de cinco partes que a su vez contienen 35 patrones de cinco puntos y, de forma resumida, la tarea de los participantes es dibujar tantos diseños únicos en cada parte como les sea posible durante un minuto, conectando los puntos evitando repetir diseños. El rendimiento en el RFFT se expresa como ***el número total de diseños únicos*** en las cinco partes y el número total de repeticiones de los diseños o ***errores perseverantes.*** En el estudio PREVEND los resultados (número de diseños únicos y errores perseverantes) fueron analizados por dos examinadores independientes y en caso de no coincidir, se introdujo a un tercer examinador y se promediaron los resultados de los dos examinadores con resultados más concordantes[47].

Los autores[47] reportaron un aumento significativo ($p_{tendencia}<0.001$) en el número medio (DE) de diseños únicos en el RFFT incrementándose de 73 (26) en la primera medición, a 79 (27) en la segunda medición y a 83 (26) en la tercera. Además dicho aumento se asoció negativamente con la edad

(disminuyó con 0,23 por incremento de un año) y no se halló relación con el nivel educativo. Resultados similares se obtuvieron para los errores perseverantes: la mediana (IQR) en la primera medición fue 7 (3-13), en la segunda medición 7 (4-14) y aumentando en la tercera medición 8 (4-15) ($p_{tendencia}$=0.002).

5.2.Análisis descriptivo de los datos mediante R.

Inicialmente se habían reclutado a 4158 participantes, pero solo un total de 2515 (61%) completaron las tres medidas del test RFFT y son los casos incluidos en la base de datos disponible a través de ***Dryad*** (http://datadryad.org/bitstream/handle/10255/dryad.77302/RFFT%20longitudinal%20data%20Groningen%20the%20Netherlands.sav?sequence=1). La base de datos está compuesta por las siguientes variables:

Tabla 1. Variables

VARIABLE	Descripción	Tipo
Casenr	Nº de individuo	entero
Age	Edad en la 1ª medida (años)	continua
Gender	Género	categórica
Education	Nivel de educación	categórica
Measurement	Nº de medida consecutiva	ordinal
Unique	Nº de diseños únicos	entero
Perseverative	Nº de errores perseverantes	entero
Interval	Tiempo desde la medida previa (años)	continua

El 53.04% de los individuos eran hombres con una edad media de 52.6 años y el 38.33% habían completado estudios universitarios. Basalmente (en la 1ª medición) os resultados del test RFFT fueron los siguientes: la media del número de diseños únicos fue de 72.9 (con una desviación estándar [DE] de 25.6) y la media del número de errores perseverantes fue de 11.9 (con DE de 15.8).

Figura 1. Frecuencias variables basales cualitativas (I)

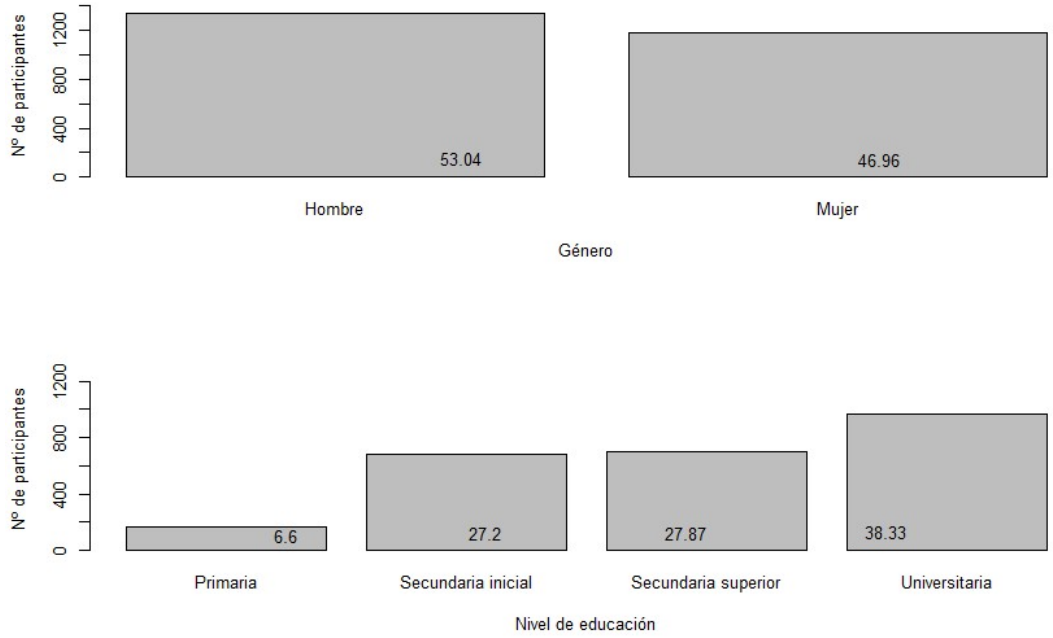
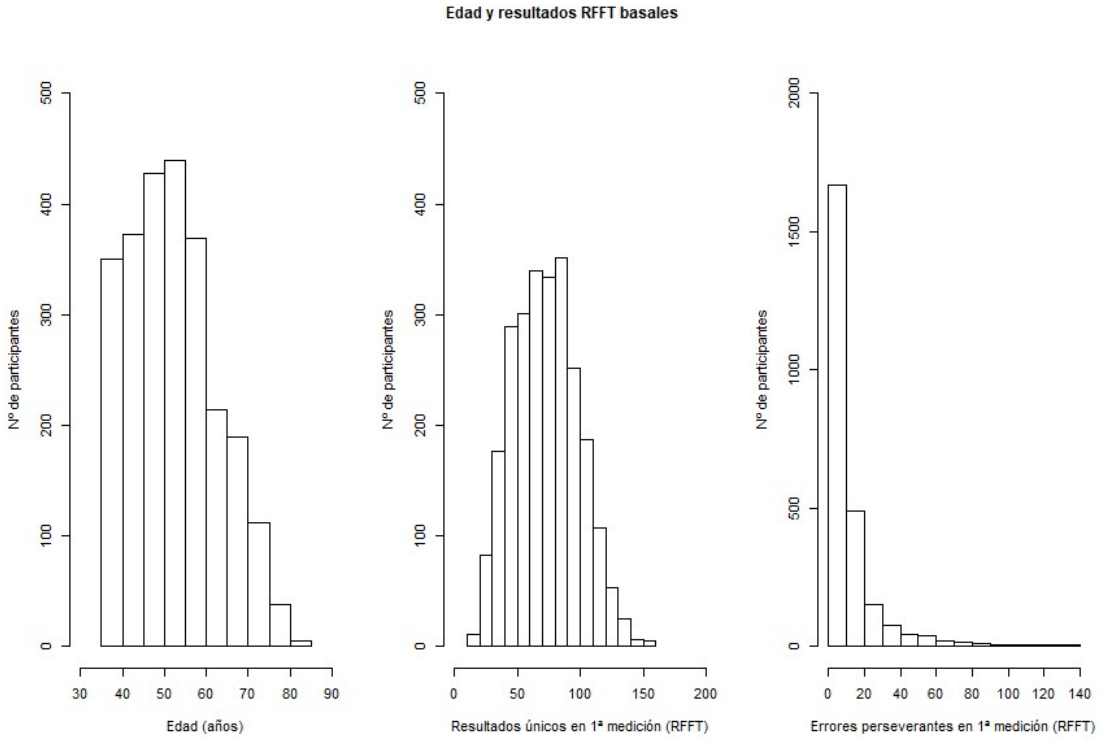


Figura 2. Histograma variables cuantitativas



5.3. Generación de datos faltantes.

Los autores analizan por separado tanto el número de diseños único como el número de errores perseverantes y sus resultados se presentan de manera independiente. De modo que a la hora de generar las bases de datos (dataframes) con datos faltantes se han generado, y posteriormente analizado, por separado ambas variables. Para plantear diversos escenarios en cuanto a la presencia de datos faltantes, se han generado de manera aleatoria un 10%, 20% y 30% de valores perdidos respectivamente en cada una de las variables que determinan las mediciones de las variables de interés. De modo que finalmente se han generado 6 bases de datos (dataframes):

- 10% - *Diseños únicos*: se han generado un total de 10.14%, 9.9% y 10.93% de datos faltantes, respectivamente en cada una de las 3 mediciones.
- 10% - *Errores perseverantes*: se han generado un total de 10.26%, 10.02% y 8.83% de datos faltantes, respectivamente en cada una de las 3 mediciones.
- 20% - *Diseños únicos*: se han generado un total de 20.12%, 20.4% y 21.07% de datos faltantes, respectivamente en cada una de las 3 mediciones.
- 20% - *Errores perseverantes*: se han generado un total de 19.68%, 21.19% y 18.61% de datos faltantes, respectivamente en cada una de las 3 mediciones.
- 30% - *Diseños únicos*: se han generado un total de 29.86%, 30.02% y 31.29% de datos faltantes, respectivamente en cada una de las 3 mediciones.
- 30% - *Errores perseverantes*: se han generado un total de 29.9%, 31.77% y 28.43% de datos faltantes, respectivamente en cada una de las 3 mediciones.

5.4.Identificación del tipo de datos faltantes.

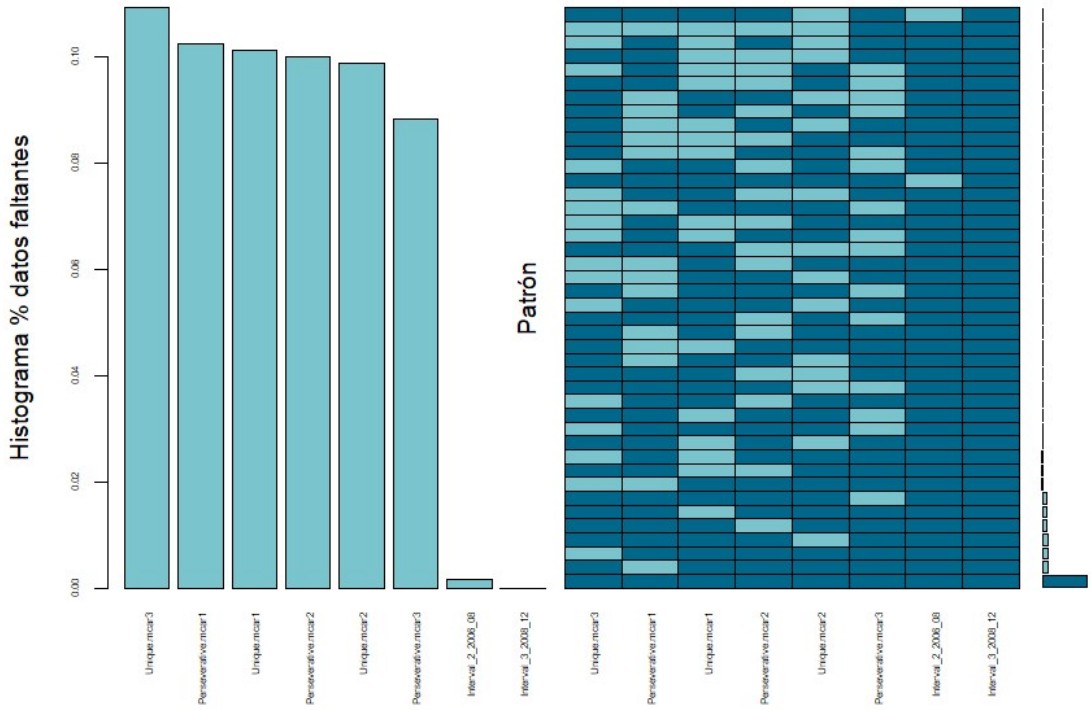
Para cada uno de los 3 escenarios de generación de datos faltantes (10%, 20% y 30%) se llevó a cabo un análisis de los patrones de los datos faltantes y de su posible relación con las características basales.

5.4.1.Análisis de los patrones de los datos faltantes (10%).

A partir de la base de datos generada con 10% de datos faltantes en cada una de las 3 mediciones del número de diseños únicos y del número de errores perseverantes, el total de pacientes que presenta al menos un valor perdido en alguna de los valores fue el 46.64%, mientras que en las 3 mediciones de diseños únicos fue el 27.83% y en las 3 mediciones de errores perseverantes fue el 26.6%.

El paquete **VIM**[39] permite la visualización de los datos faltantes para el análisis de los posibles patrones. Aunque ya se sabía a priori que en la base de datos original no había datos faltantes, se analizaron los posibles patrones sin encontrar ningún resultado destacable (resultados no mostrados). Para el escenario del 10% de datos faltantes, a partir del gráfico que proporciona la función *"**aggr**" (paquete VIM)*, se puede observar que ninguno de los patrones que involucran más de una variable presenta una frecuencia superior al resto (panel derecho de la gráfica). En el histograma (panel izquierdo), se representan las frecuencias relativas (%) de datos faltantes en las variables de interés. Dichas frecuencias ya se habían reportado en la sección anterior.

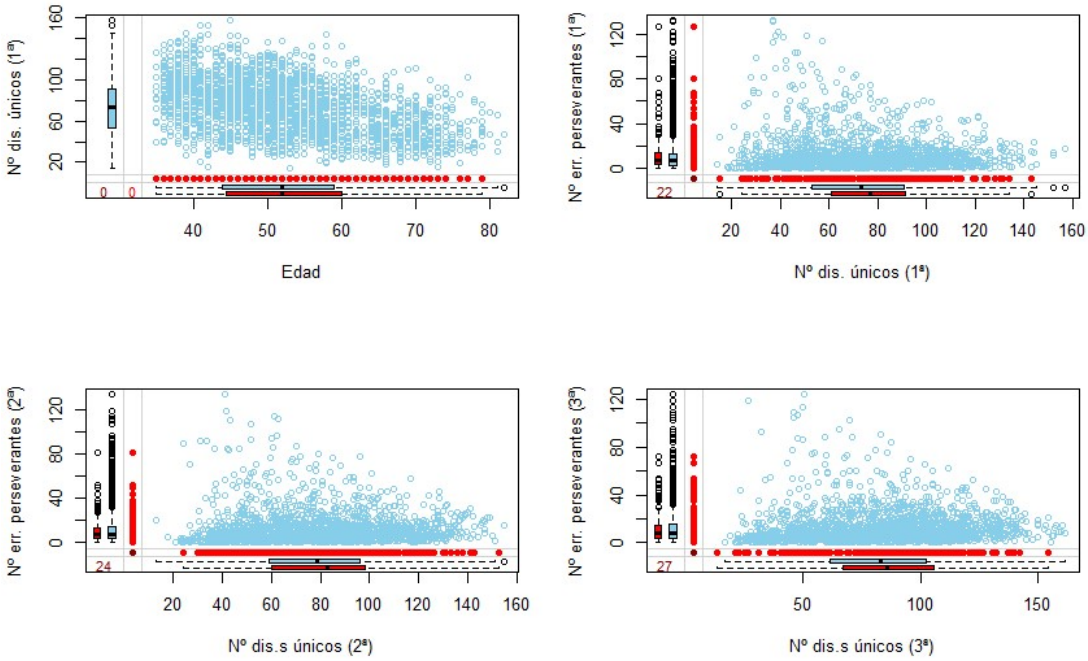
Figura 3. Patrón datos faltantes tras generar 10% de datos faltantes



El paquete ***BaylorEdPsych***[50,51], contiene el una función (***LittleMCAR***) que permite aplicar el test *Little´s MCAR-test*, para analizar la hipótesis nula de que *los datos faltantes son MCAR*, donde un resultado significativo indicaría que se debe rechazar dicha hipótesis. Se ha aplicado el test sobre las variables de interés para el estudio: edad, género, nivel de educación, 1ª medición diseños únicos, 2ª medición diseños únicos, 3ª medición diseños únicos, 1ª medición errores perseverantes, 2ª medición errores perseverantes y 3ª medición errores perseverantes A partir del resultado de test de Little's, no se puede rechazar la hipótesis nula (*p-valor*=0.94421>α=0.05), y por lo tanto, como era de esperar al haber sido generados de manera aleatoria, los datos faltantes cumplen las características de MCAR.

El paquete **VIM**[39] dispone de gráficos alternativos (por ejemplo: *marginplot*) que permiten analizar en profundidad la posible relación de la presencia de datos faltantes por pares de variables, para intentar detectar patrones ocultos.

Figura 4. Marginplot: relación entre pares de variables tras generar 10% de datos faltantes



Los gráficos anteriores permiten comparar la distribución de los datos faltantes por pares de variables. De modo que en el primer gráfico, la fila de puntos rojos indicaría como se distribuirían los datos faltantes de la variable *nº de diseños únicos 1ª* en función de los valores de la variable *edad*, y los puntos azules serían el resto de valores. El hecho de que los boxplots representados en el eje x (*edad*) sean similares indica que los valores de edad con y sin valores faltantes en el nº de diseños únicos, no presentan diferencias y es un indicativo de que los datos cumplen las premisas de **MCAR**. Igualmente entre el resto de pares de variables representadas, visualmente tampoco se han hallado diferencias entre la presencia o ausencia de datos faltantes.

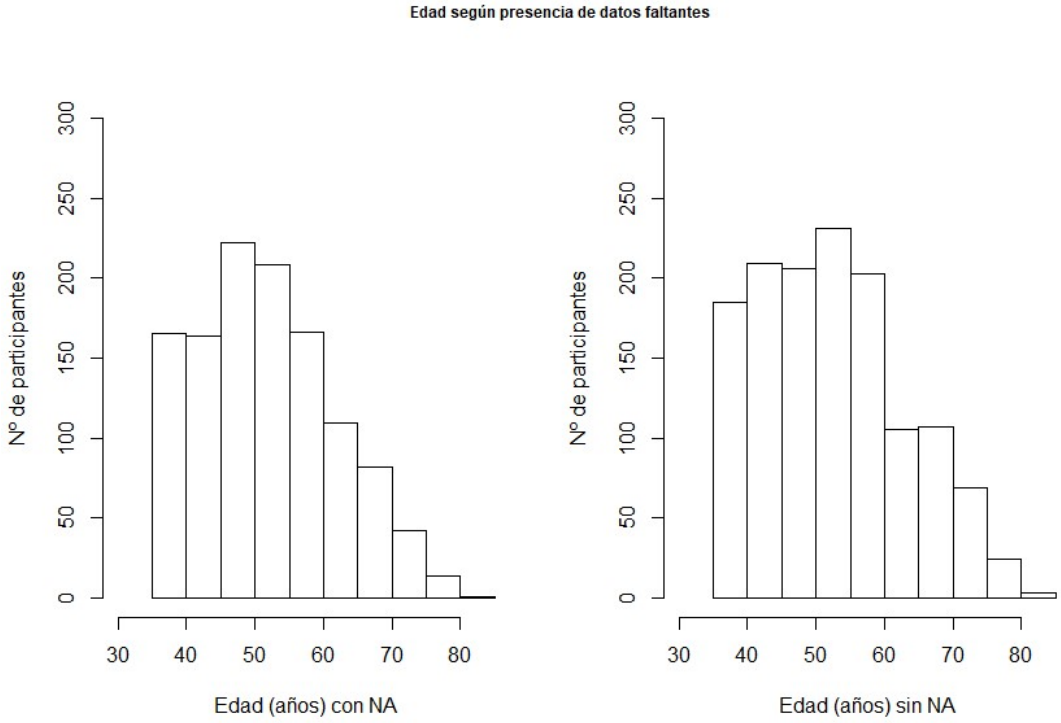
5.4.2.Características basales Vs datos faltantes todas las variables (10%)

A continuación se presentan los resultados obtenidos tras analizar las posibles diferencias en las características basales en función de la ausencia o presencia de al menos un dato faltante. Como ya se ha indicado previamente el 46.64% de los participantes presentaban al menos un dato faltante.

Edad vs al menos un dato faltante en alguna de las variables (10%)

No se hallaron diferencias estadísticamente significativas en la edad entre los casos con y sin algún dato faltante. No se pudo asumir la normalidad de los datos (ambos p-valores en el test Saphiro-Wilks son significativos, *p-valor*<0.001<α=0.05. Por lo tanto para analizar si hay diferencias en las edades entre ambos grupos de pacientes, se debe que utilizar el resultado obtenido en el test de Wilcoxon y se puede afirmar que no se han hallado diferencias estadísticamente significativas en las edades entre los grupos definidos en función de la presencia/ausencia de datos faltantes (*p-valor*=0.192> α=0.05). Los histogramas que se presentan a continuación confirman que ambas distribuciones de la edad en función de la presencia y ausencia de datos perdidos son muy similares.

Figura 5. Edad en función de la presencia al menos un dato faltante (10%)

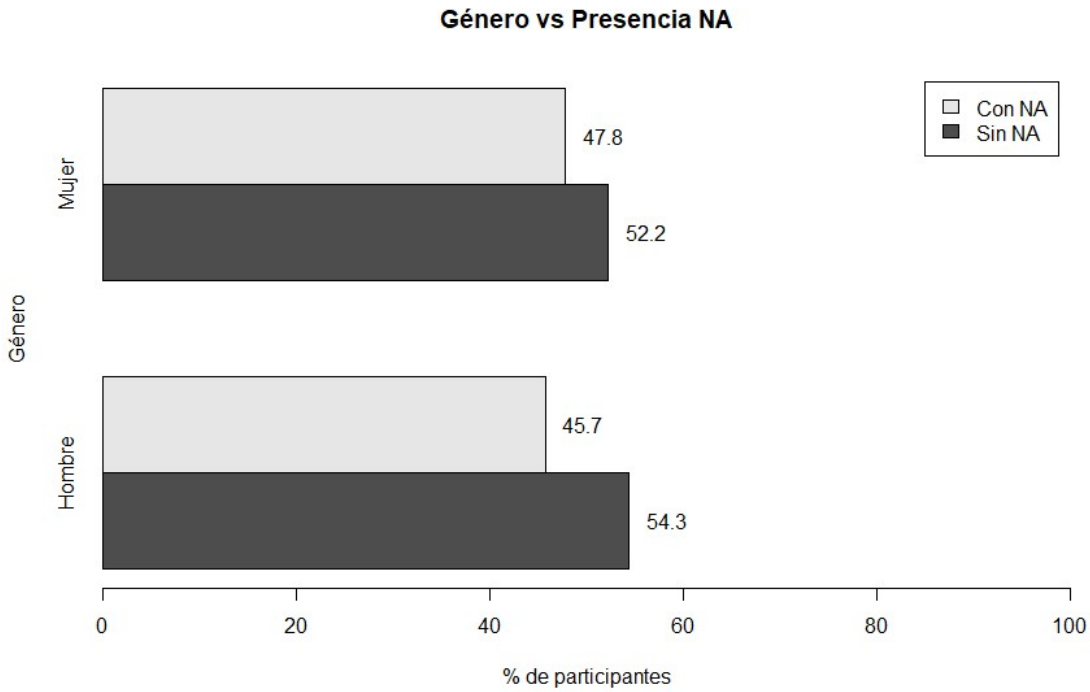


Género y nivel de educación vs al menos un dato faltante (10%)

Tanto para analizar si hay diferencias en la proporción de datos faltantes entre géneros y entre los grupos del nivel de educación, se utiliza el test de Chi-cuadrado o en su defecto el test de Fisher.

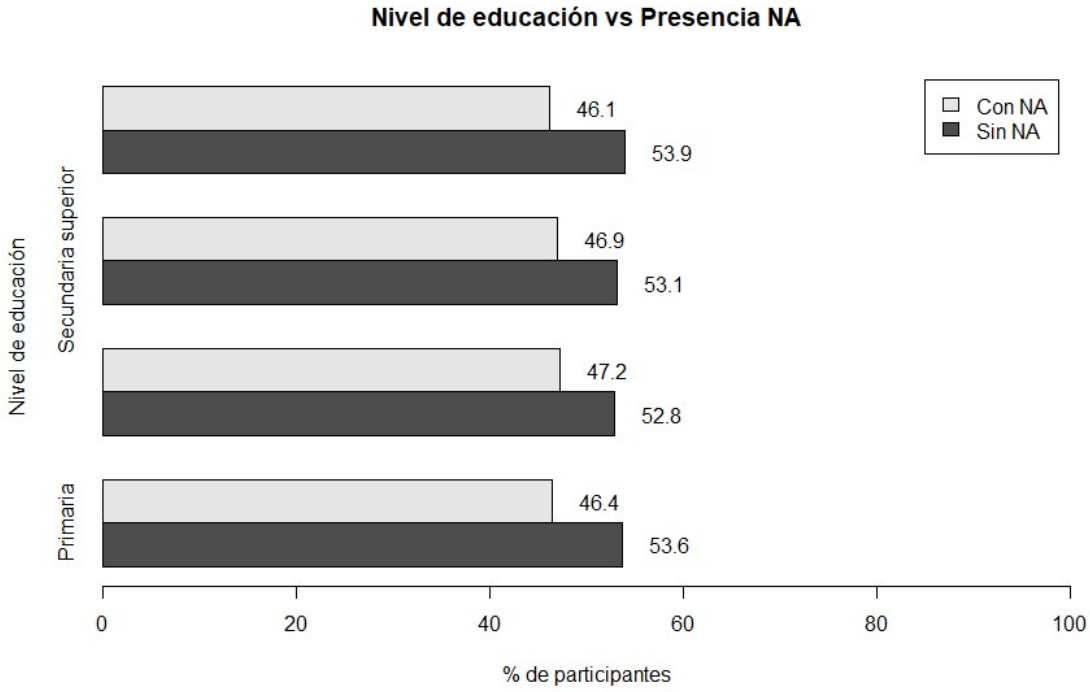
El % de casos con al menos un dato perdido en los hombres fue del 45.7% y en las mujeres del 47.8%, tras aplicar el test de Chi-cuadrado, no se hallaron diferencias estadísticamente significativas (p=0.291).

Figura 6. Género en función de la presencia de datos faltantes (10%)



De modo similar, también se utilizó el Test de Chi-cuadrado para comparar las proporciones de al menos un dato faltante entre las categorías del nivel de educación. Dichas proporciones eran muy similares (ver gráfico siguiente) y como el p-valor=0.969>α=0.05, no se pudo rechazar la hipótesis nula de ausencia de diferencias, por lo tanto el % de datos faltantes no varió entre los niveles de educación.

Figura 7. Nivel de educación en función de la presencia de datos faltantes (10%)



Todos los resultados anteriores corroboraron la ausencia de patrones de los datos faltantes y de relación entre la presencia y ausencia de datos faltantes con las características basales. Además el resultado obtenido mediante el test *Little´s MCAR-test*, permitió asumir que los datos faltantes generados aleatoriamente en el escenario del 10% son *MCAR.*

5.4.3.Características basales Vs datos faltantes en diseños únicos (10%)

El 27.83% de los participantes presentaron al menos un dato faltante en alguna de las 3 mediciones de diseños únicos. De modo similar a la sección anterior, en ninguno de los caso se hallaron diferencias estadísticamente significativas en las características basales entre los participantes con y sin datos faltantes en alguna de las 3 mediciones del número de diseños únicos (ver informe para más detalles de los test estadísticos).

Figura 8. Edad en función de la presencia de datos faltantes en diseños únicos (10%)

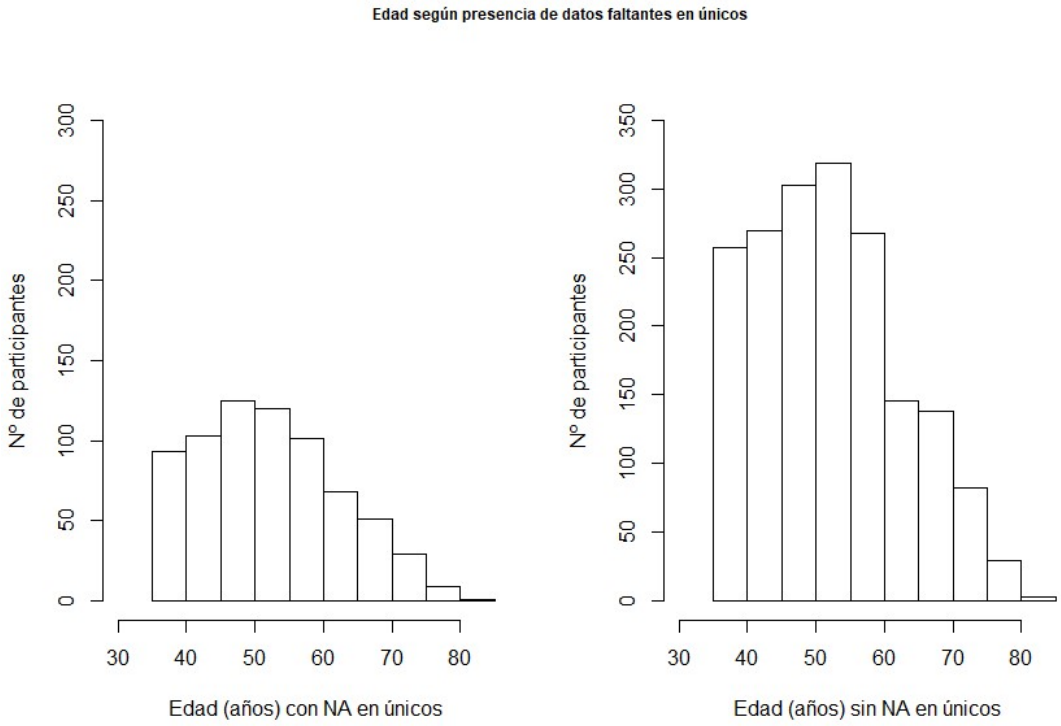


Figura 9. Género en función de la presencia de datos faltantes en diseños únicos (10%)

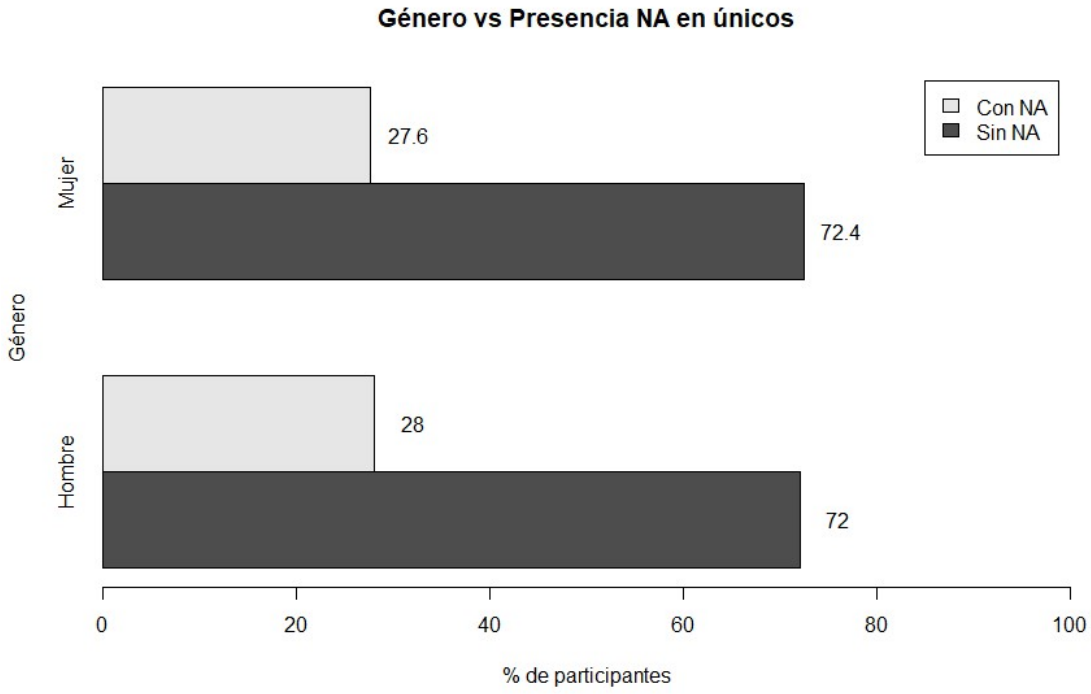
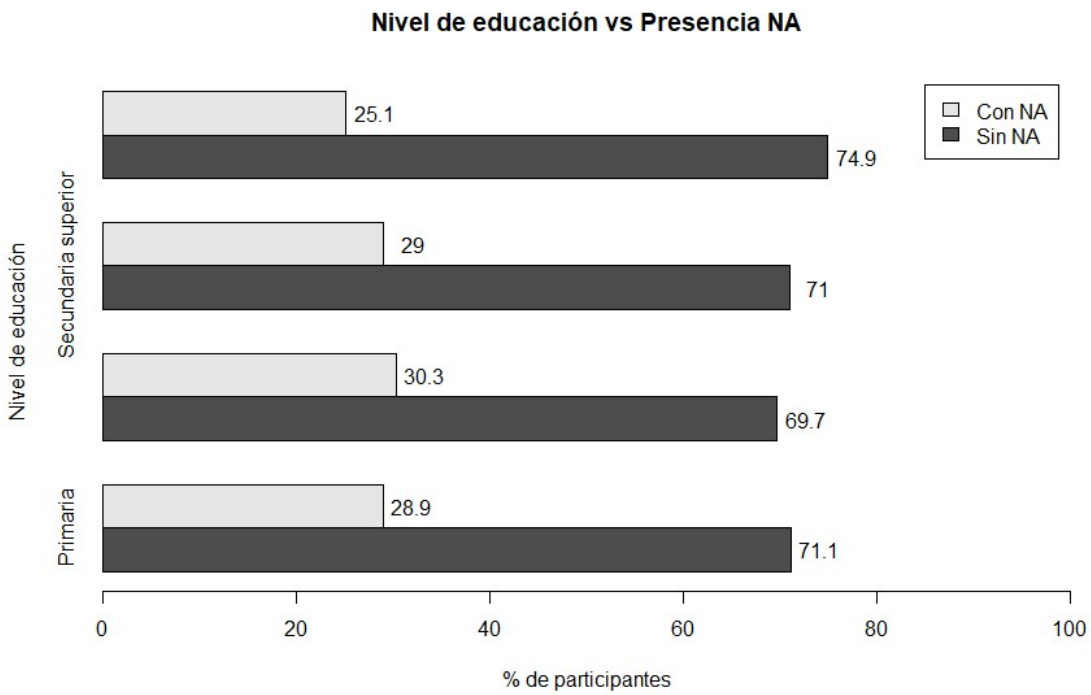


Figura 10. Nivel de educación en función de la presencia de datos faltantes en diseños únicos (10%)



5.4.4.Características basales Vs datos faltantes en errores perseverantes (10%)

El 26.6% de los participantes presentaron al menos un dato faltante en alguna de las 3 mediciones de errores perseverantes. De modo similar a la sección anterior, en ninguno de los caso se hallaron diferencias estadísticamente significativas en las características basales entre los participantes con y sin datos faltantes en alguna de las 3 mediciones del número de errores perseverantes (ver informe para más detalles de los test estadísticos).

Figura 11. Edad en función de la presencia de datos faltantes en errores perseverantes (10%)

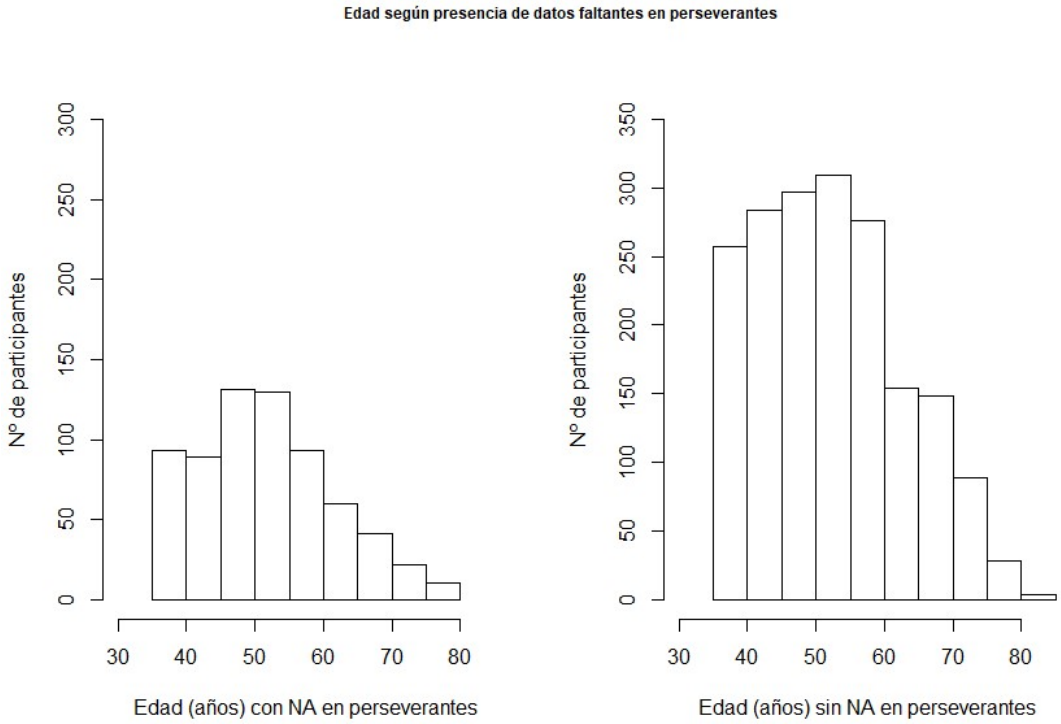


Figura 12. Género en función de la presencia de datos faltantes en errores perseverantes (10%)

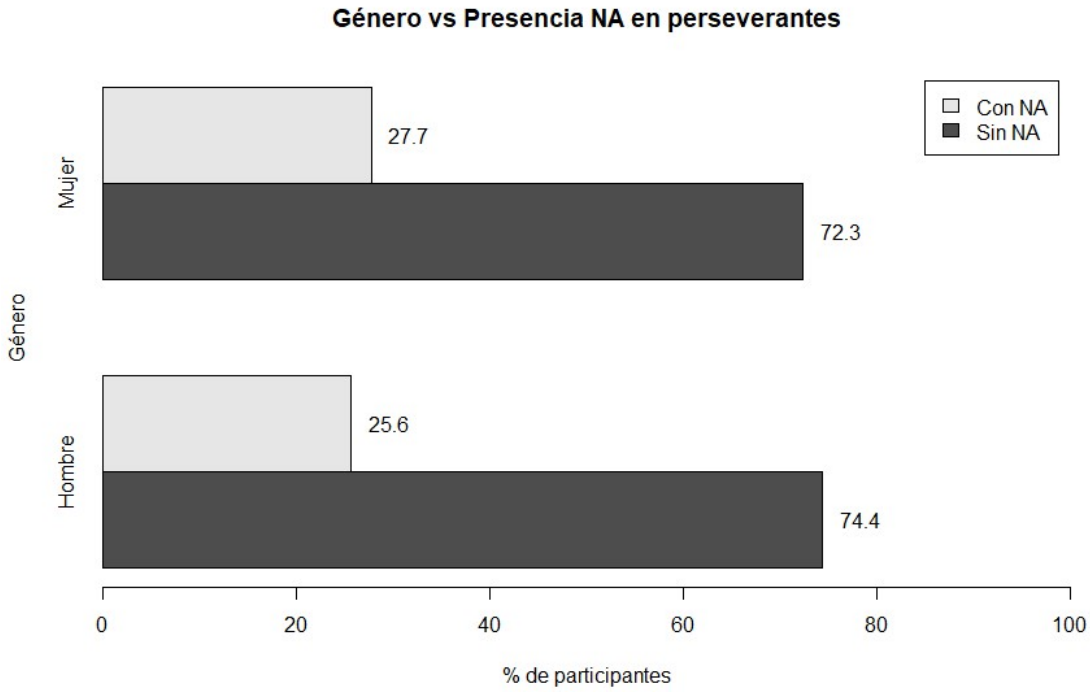
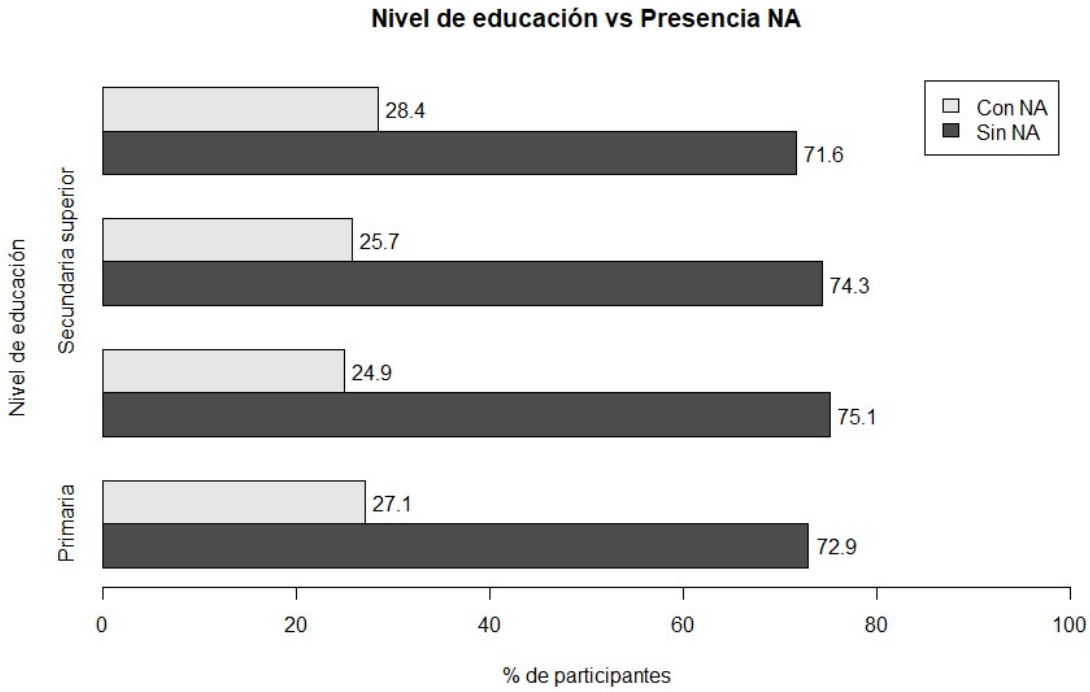


Figura 13. Nivel de educación en función de la presencia de datos faltantes en errores perseverantes (10%)

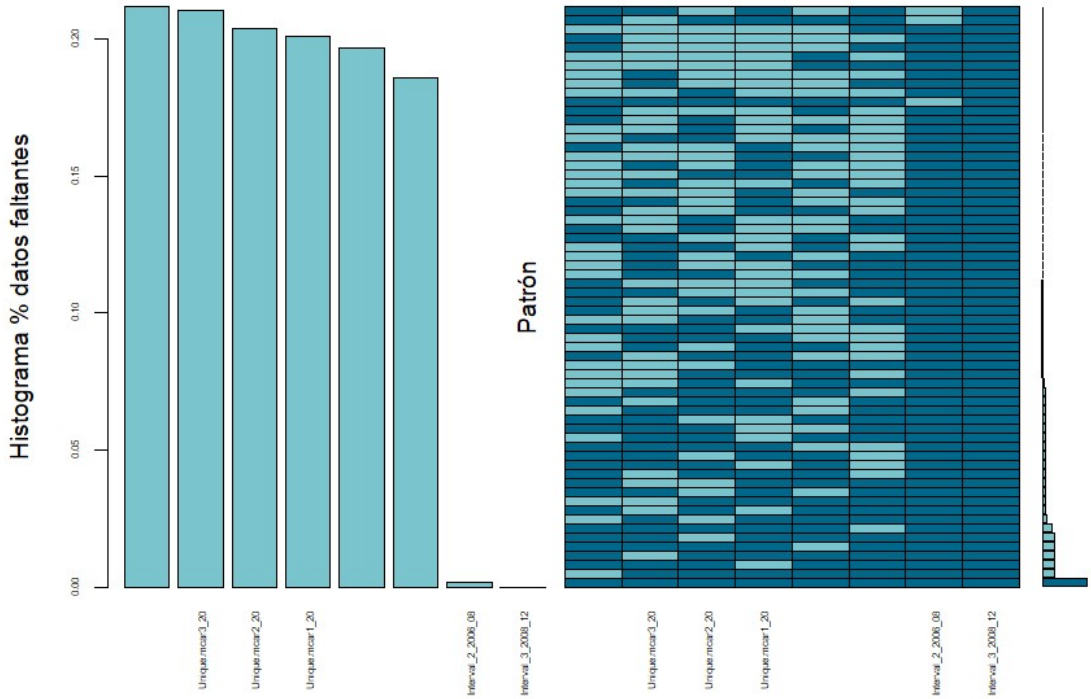


5.4.5.Análisis de los patrones de los datos faltantes (20%).

A partir de la base de datos generada con 20% de datos faltantes en cada una de las 3 mediciones del número de diseños únicos y del número de errores perseverantes, el total de pacientes que presenta al menos un valor perdido en alguna de los valores fue el 74.71%, mientras que en las 3 mediciones de diseños únicos fue el 50.5% y en las 3 mediciones de errores perseverantes fue el 48.43%.

Para el escenario del 20% de datos faltantes, a partir del gráfico siguiente, se puede observar que ninguno de los patrones que involucran más de una variable presenta una frecuencia superior al resto (panel derecho de la gráfica). En el histograma (panel izquierdo), se representan las frecuencias relativas (%) de datos faltantes en las variables de interés. Dichas frecuencias ya se habían reportado en una sección previa.

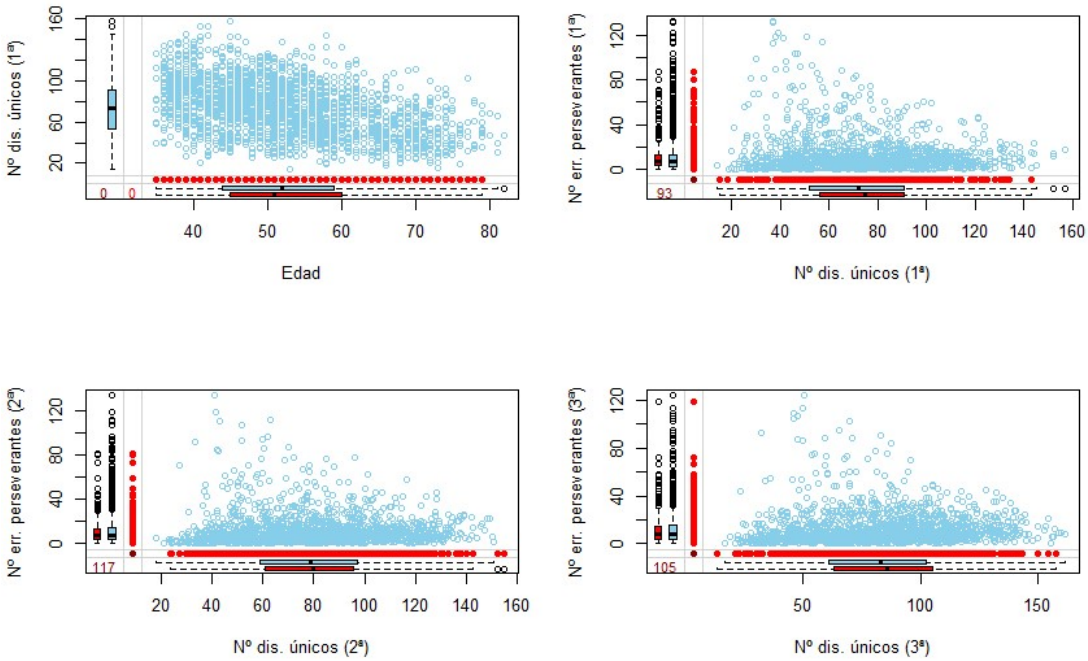
Figura 14. Patrón datos faltantes tras generar 20% de datos faltantes



Se ha aplicado el test sobre las variables de interés para el estudio: edad, género, nivel de educación, 1ª medición diseños únicos, 2ª medición diseños únicos, 3ª medición diseños únicos, 1ª medición errores perseverantes, 2ª medición errores perseverantes y 3ª medición errores perseverantes A partir del resultado de test de Little's, no se puede rechazar la hipótesis nula (*p-valor*=0.99304>α=0.05), y por lo tanto, como era de esperar al haber sido generados de manera aleatoria, los datos faltantes cumplen las características de MCAR.

De nuevo utilizamos el *marginplot* del paquete **VIM**[39] para analizar en profundidad la posible relación de la presencia de datos faltantes por pares de variables, para intentar detectar patrones ocultos.

Figura 15. Marginplot: relación entre pares de variables tras generar 20% de datos faltantes



Los gráficos anteriores permiten comparar la distribución de los datos faltantes por pares de variables. De modo que en el primer gráfico, la fila de puntos rojos indicaría como se distribuirían los datos faltantes de la variable *nº de diseños únicos 1ª* en función de los valores de la variable *edad*, y los puntos azules serían el resto de valores. El hecho de que los boxplots representados en el eje x (*edad*) sean similares indica que los valores de edad con y sin valores faltantes en el nº de diseños únicos, no presentan diferencias y es un indicativo de que los datos cumplen las premisas de **MCAR**. Igualmente entre el resto de pares de variables representadas, visualmente tampoco se han hallado diferencias entre la presencia o ausencia de datos faltantes, corroborando la ausencia de patrones en los datos faltantes.

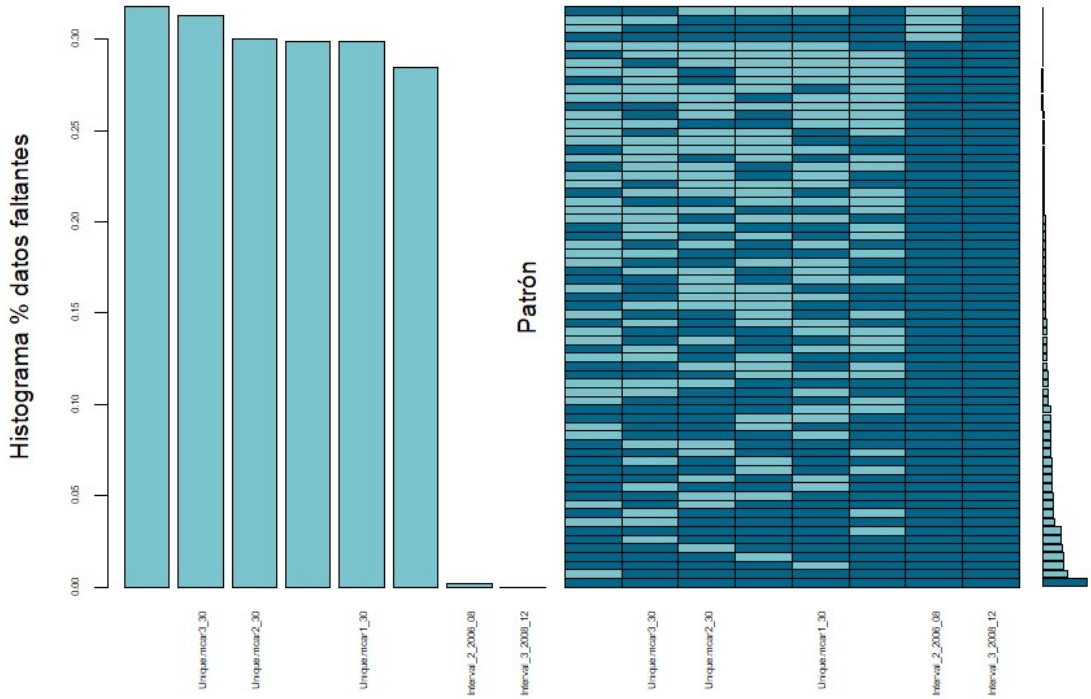
De modo similar la escenario generado con el 10% de datos faltantes, en el escenario generado con 20% de datos faltantes, también se analizaron las posibles relaciones entre la presencia y ausencia de datos faltantes y las características basales (conjuntamente y por separado para diseños únicos y errores perseverantes) y en ninguno de los caso se hallaron diferencias estadísticamente significativas (ver informe para más detalles de los test estadísticos).

5.4.6. Análisis de los patrones de los datos faltantes (30%).

A partir de la base de datos generada con 30% de datos faltantes en cada una de las 3 mediciones del número de diseños únicos y del número de errores perseverantes, el total de pacientes que presenta al menos un valor perdido en alguna de los valores fue el 89.07%, mientras que en las 3 mediciones de diseños únicos fue el 66.92% y en las 3 mediciones de errores perseverantes fue el 66.88%.

Para el escenario del 30% de datos faltantes, a partir del gráfico siguiente, se puede observar que ninguno de los patrones que involucran más de una variable presenta una frecuencia superior al resto (panel derecho de la gráfica). En el histograma (panel izquierdo), se representan las frecuencias relativas (%) de datos faltantes en las variables de interés. Dichas frecuencias ya se habían reportado en una sección previa.

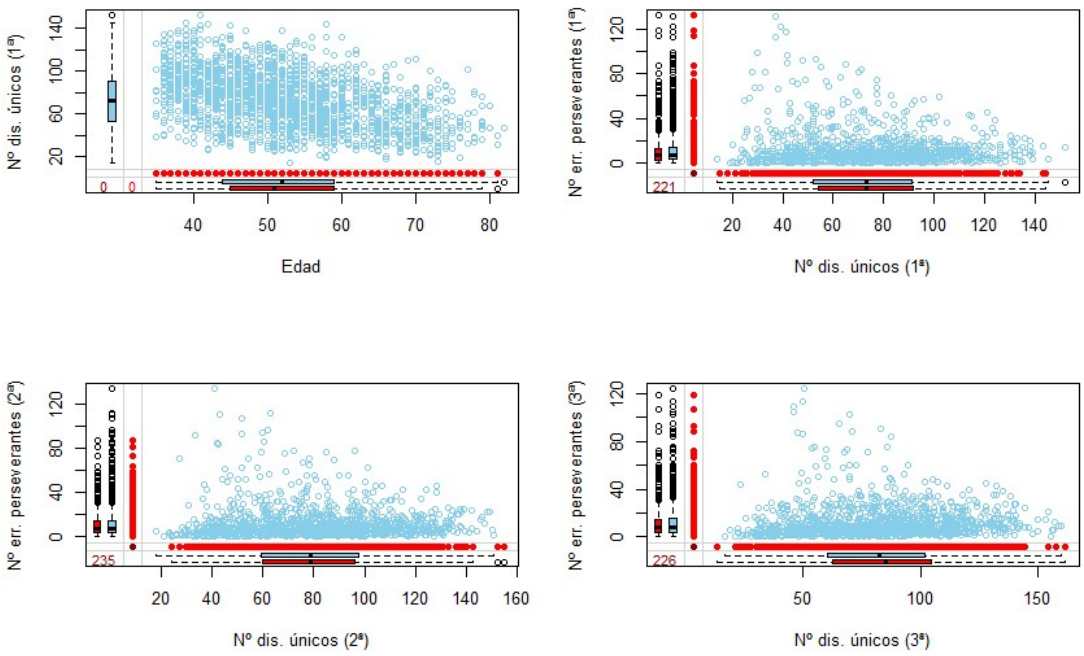
Figura 16. Patrón datos faltantes tras generar 30% de datos faltantes



Se ha aplicado el test sobre las variables de interés para el estudio: edad, género, nivel de educación, 1ª medición diseños únicos, 2ª medición diseños únicos, 3ª medición diseños únicos, 1ª medición errores perseverantes, 2ª medición errores perseverantes y 3ª medición errores perseverantes A partir del resultado de test de Little's, no se puede rechazar la hipótesis nula (*p-valor*=0.51085>α=0.05), y por lo tanto, como era de esperar al haber sido generados de manera aleatoria, los datos faltantes cumplen las características de MCAR.

De nuevo utilizamos el *marginplot* del paquete **VIM**[39] para analizar en profundidad la posible relación de la presencia de datos faltantes por pares de variables, para intentar detectar patrones ocultos.

Figura 17. Marginplot: relación entre pares de variables tras generar 30% de datos faltantes



Los gráficos anteriores permiten comparar la distribución de los datos faltantes por pares de variables. De modo que en el primer gráfico, la fila de puntos rojos indicaría como se distribuirían los datos faltantes de la variable *nº de diseños únicos 1ª* en función de los valores de la variable *edad*, y los puntos azules serían el resto de valores. El hecho de que los boxplots representados en el eje x (*edad*) sean similares indica que los valores de edad con y sin valores faltantes en el nº de diseños únicos, no presentan diferencias y es un indicativo de que los datos cumplen las premisas de **MCAR**. Igualmente entre el resto de pares de variables representadas, visualmente tampoco se han hallado diferencias entre la presencia o ausencia de datos faltantes, corroborando la ausencia de patrones en los datos faltantes.

De modo similar la escenario generado con el 10% de datos faltantes, en el escenario generado con 30% de datos faltantes, también se analizaron las posibles relaciones entre la presencia y ausencia de datos faltantes y las características basales (conjuntamente y por separado para diseños únicos y errores perseverantes) y en ninguno de los caso se hallaron diferencias estadísticamente significativas (ver informe para más detalles de los test estadísticos).

5.5. Tratamiento de datos faltantes.

Como ya se ha detallado previamente en la sección 4 de la presente memoria, existen diversas técnicas para el tratamiento de datos faltantes. El objetivo de este TFM, además de haber detallado las diversas técnicas existentes, es la aplicación de los métodos estudiados mediante el análisis de una base de datos biomédicos. A modo de ejemplo se han seleccionado algunos de los métodos más habitualmente utilizados por la comunidad científica:

- En primer lugar se propone uno de los métodos de eliminación: Eliminación de los casos (listwise).

- En segundo lugar se propone utilizar dos métodos de imputación simples: Reemplazamiento por la media e Imputación por regresión simple (media condicional).

- Finalmente se propone uno de los métodos modernos sobre los que más investigación se está llevando a cabo en los últimos años como es el Algoritmo de imputación múltiple (MI)[26-36] (método Predictive Mean Matching o Equiparación de media predictiva, PMM[28,29]).

Los 4 métodos se han aplicado a cada uno de los 3 escenarios: 10%, 20% y 30% de datos faltantes, para analizar las posibles diferencias entre ellos en función de la frecuencia relativa de datos faltantes.

5.5.1. Eliminación de los casos (listwise).

Para cada uno de los 3 escenarios (10%, 20% y 30%) se generaron dos bases de datos (dataframes) por separado para los casos de Nº de diseños únicos y Nª de errores perseverantes, en los que se procedió a eliminar cada caso con datos faltantes.

En el escenario del ***10%*** de datos faltantes:

- El nº de participantes disponibles para el análisis del nº de diseños únicos fue de 1815 (participantes con las 3 mediciones completas).

- El nº de participantes disponibles para el análisis del nº de errores perseverantes fue de 1846 (participantes con las 3 mediciones completas)

En el escenario del ***20%*** de datos faltantes:

- El nº de participantes disponibles para el análisis del nº de diseños únicos fue de 1245 (participantes con las 3 mediciones completas).

- El nº de participantes disponibles para el análisis del nº de errores perseverantes fue de 1297 (participantes con las 3 mediciones completas)

En el escenario del ***30%*** de datos faltantes:

- El nº de participantes disponibles para el análisis del nº de diseños únicos fue de 832 (participantes con las 3 mediciones completas).

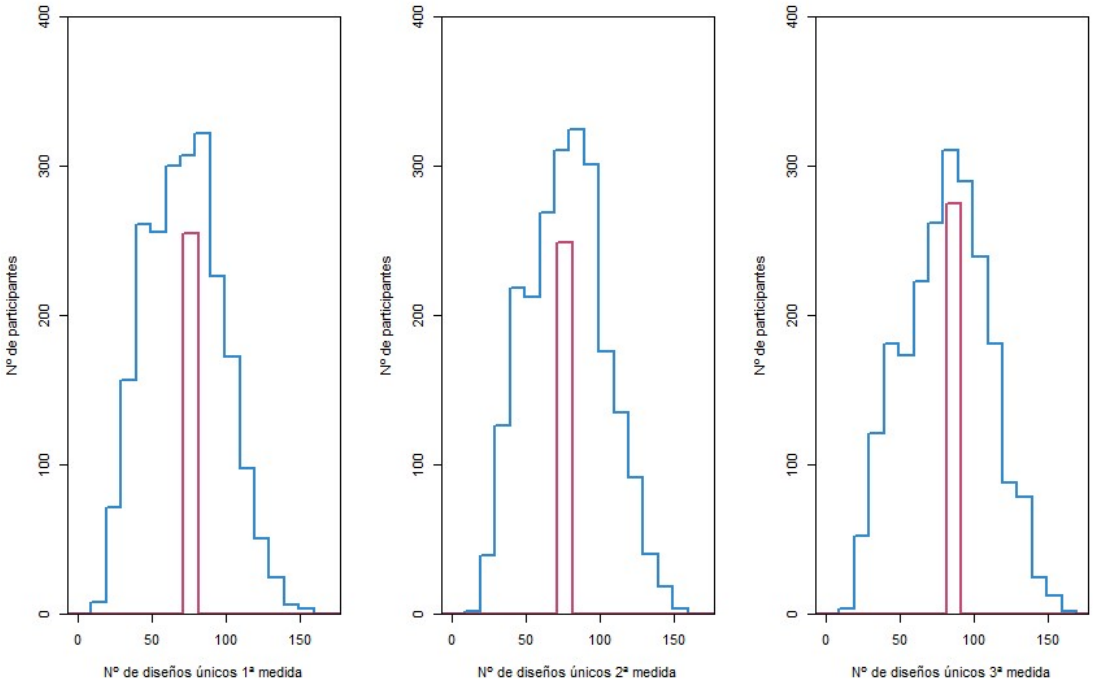
- El nº de participantes disponibles para el análisis del nº de errores perseverantes fue de 833 (participantes con las 3 mediciones completas)

5.5.2. Reemplazamiento por la media

Para cada uno de los 3 escenarios (10%, 20% y 30%) se generaron dos bases de datos (dataframes) por separado para los casos de Nº de diseños únicos y Nª de errores perseverantes, en los que se procedió a sustituir los datos faltantes por la media en cada variable que contenía datos faltantes.

La función *"mice"*[40] permite calcular la media en cada variable y reemplazar los datos faltantes de manera sencilla. A partir de varios ejemplos comentados por el Prof. Dr. Stef Van Buuren (http://www.stefvanbuuren.nl/mi/FIMD.html)[52] basados en la utilización de *mice,* se representa gráficamente la distribución de los datos imputados en cada una de las 3 variables de la medición del **nº de diseños únicos**, en el escenario de 10% de datos faltantes.

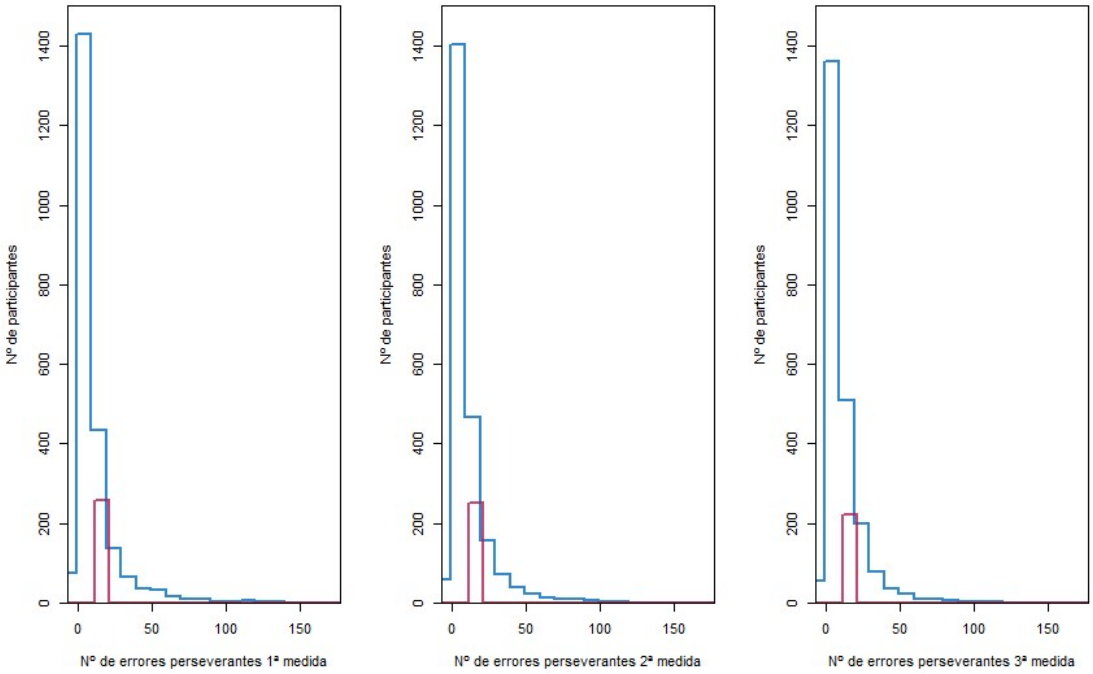
Figura 18. Imputación datos faltantes por media (MICE): Nº de diseños únicos (10%)



En esta gráfica se presenta a distribución de los datos sin datos faltantes (mediante la línea azul) y la media imputada (rojo), por ejemplo para el primer caso la media del Nº de diseños únicos en la 1ª medida fue 73.34 y se imputó en algo más de 250 casos (aproximado). En la 2ª medida la media fue 78.81 y en la 3ª medida fue 82.62.

De modo similar se imputan los datos faltantes para el ***nº de errores perseverantes*** en cada una de las 3 mediciones.

Figura 19. Imputación datos faltantes por media (MICE): Nº de errores perseverantes (10%)



La media del Nº de errores perseverantes en la 1ª medida fue 11.76, en la 2ª medida fue 11.74 y en la 3ª medida 11.99.

Analizando ambas gráficas, se puede apreciar como en al caso de variables que no presenten asimetría, especialmente si los datos siguen MCAR (como estos casos), los datos se verán menos afectados por esta imputación (aunque disminuirá la variabilidad). Pero en el caso del nº de errores perseverantes que presentan asimetría, como la media se ve afectada por la presencia de valores extremos, el resultado final puede verse muy afectado, especialmente en casos con un número alto de datos perdidos (por ejemplo el escenario del 30%).

Para los escenarios con 20% y 30% de datos faltantes, se reemplazaron los datos faltantes de manera similar (las gráficas se pueden ver en el informe):

- **20%**: La media del Nº de diseños únicos en la 1ª medida fue 73.3, en la 2ª medida fue 79.07 y en la 3ª medida 82.67. La media del Nº de errores perseverantes en la 1ª medida fue 11.93, en la 2ª medida fue 11.66 y en la 3ª medida 11.95.
- **30%**: La media del Nº de diseños únicos en la 1ª medida fue 73.46, en la 2ª medida fue 79.29 y en la 3ª medida 82.19. La media del Nº de errores perseverantes en la 1ª medida fue 11.76, en la 2ª medida fue 11.72 y en la 3ª medida 12.12.

5.5.3. Imputación por regresión simple

Para cada uno de los 3 escenarios (10%, 20% y 30%) se generaron dos bases de datos (dataframes) por separado para los casos de Nº de diseños únicos y Nª de errores perseverantes, en los que se procedió a sustituir los datos faltantes por el valor predicho a partir de la aplicación de la regresión simple sobre el resto de datos conocidos en el sujeto

Mediante la utilización de la opción *norm.predict* de la función *"mice"*[40] se reemplazan los datos faltantes de manera sencilla, de nuevo se pueden representar gráficamente la distribución de los datos imputados en cada una de las 3 variables de la medición del **nº de diseños únicos** y del **nº de errores perseverantes**, en el escenario de 10% de datos faltantes.

Figura 20. Imputación datos faltantes por regresión lineal (MICE): Nº de diseños únicos (10%)

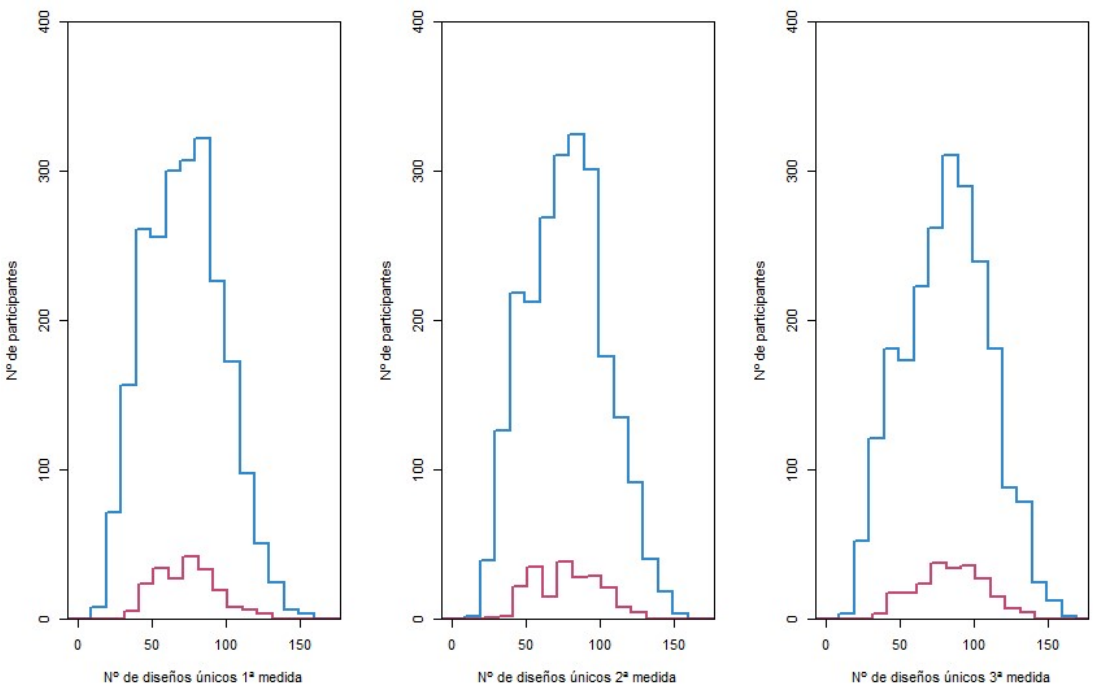
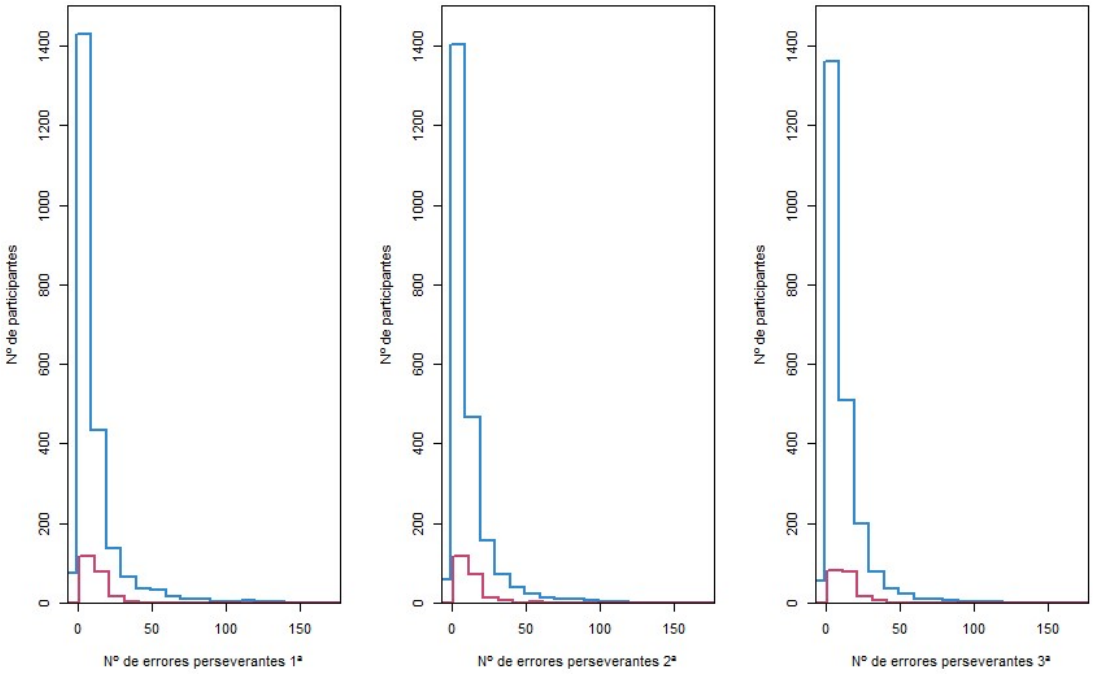


Figura 21. Imputación datos faltantes por regresión lineal (MICE): Nº de errores perseverantes (10%)



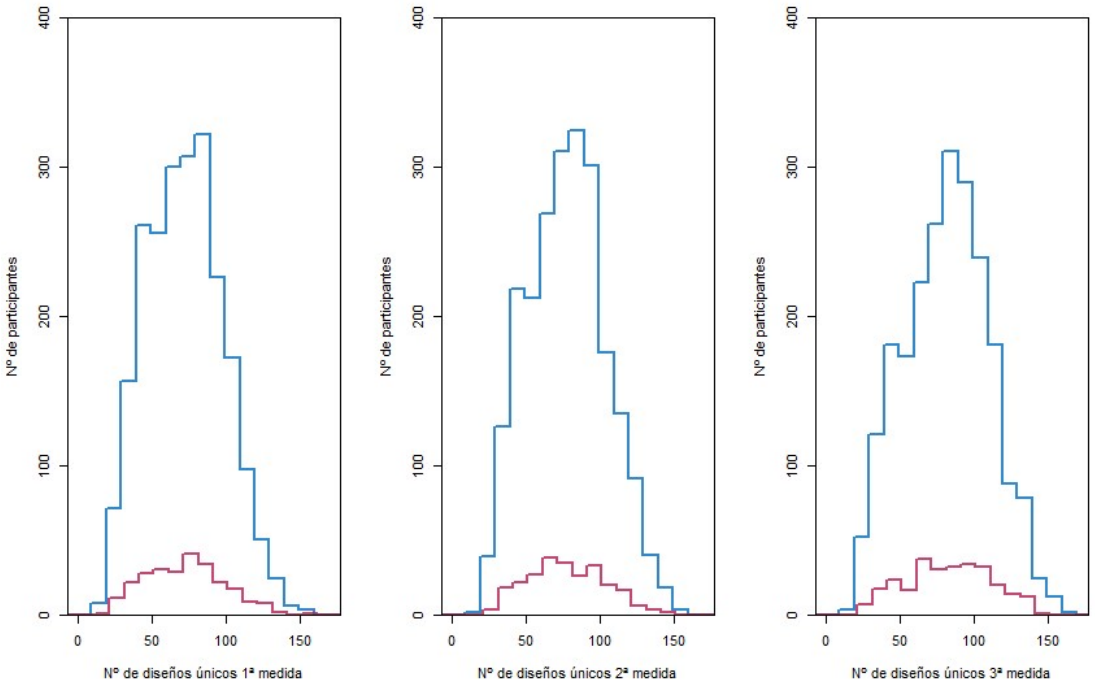
Las figuras anteriores, permiten apreciar que la distribución de los datos imputados presenta una variabilidad algo menor que aquellas de los datos reales, esto se explica porque con este método los casos imputados toman los valores con una mayor probabilidad de acuerdo al modelo de regresión. Para los escenarios con 20% y 30% de datos faltantes, se reemplazaron los datos faltantes del mismo modo (las gráficas se pueden ver en el informe) y los resultados son similares.

5.5.4. Imputación múltiple (método PMM)

Para cada uno de los 3 escenarios (10%, 20% y 30%) se generaron dos bases de datos (dataframes) por separado para los casos de Nº de diseños únicos y Nª de errores perseverantes, utilizando el algoritmo de *Múltiple Imputación utilizando el método PMM (Predictive mean matching)*. Este método obtiene buenos resultados tanto para variables continuas como categóricas (binarias o más categorías) sin necesidad de calcular los errores (residuals) ni ajustar por máxima verosimilitud, mantiene de manera muy eficaz la variabilidad original de los datos, pues el algoritmo "toma prestados" datos reales de casos para los que si se disponen de datos[28]. Mediante el método MI, cada una de las imputaciones genera un conjunto de datos diferentes los cuales se analizan por separado mediante técnicas estadísticas tradicionales, obteniéndose *m* estimaciones y sus errores estándar.

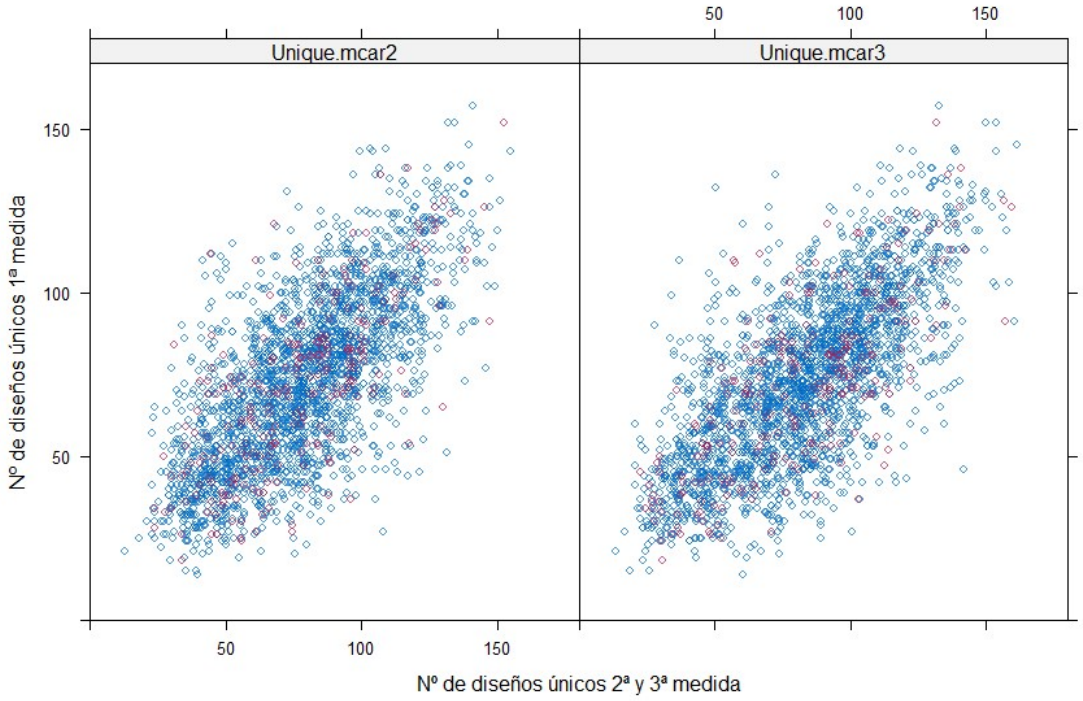
PMM es por defecto el algoritmo que utiliza la función *"mice"* para imputar datos faltantes, por lo que en la opción *meth* no sería necesario especificar que el método a utilizar es *pmm*. Se recomienda que la imputación múltiple (IM) se efectúe con al menos 5 imputaciones (también es el nº que *mice* emplea por defecto) y el nº de iteraciones también por defecto es de 5. Una vez imputados, mediante la función *pool* se obtiene el valor medio para todas las imputaciones Y se representa gráficamente la distribución de los datos imputados en cada una de las 3 variables de la medición del nº de diseños únicos.

Figura 22. Imputación datos faltantes por IM (MICE-PMM): Nº de diseños únicos (10%)



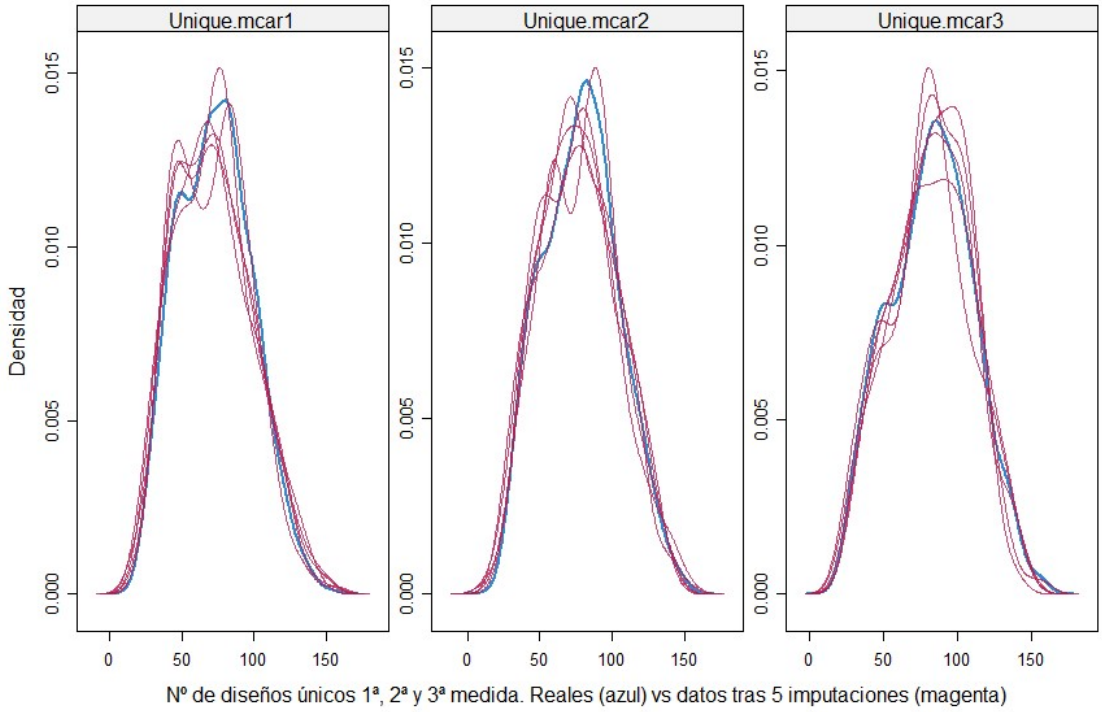
Se puede utilizar la función *xyplot* para representar los valores reales frente a los imputados en las 3 variables de medición del nº de diseños únicos (dos a dos)[53]. Se puede apreciar que no hay valores imputados (color magenta) que llamen la atención por presentar valores extremos respecto a los valores reales (azul), y que la variabilidad de los datos se mantiene de manera correcta respecto a los datos reales.

Figura 23. Valores reales vs imputados IM (MICE-PMM): Nº de diseños únicos (10%)



Para evaluar visualmente la eficacia de la imputación mediante el algoritmo PMM, utilizando la función ***densityplot*** existe la opción de obtener una figura comparativa de los gráficos de densidad de los datos contenidos en una variable determinada antes y en cada una de la imputaciones de los datos faltantes[53,54]. Tal y como se muestra en la figura a continuación, para las 3 mediciones del nº de diseños únicos, las distribuciones de los valores imputados (en color magenta) se solapan con aquella de los datos que no contenían datos faltantes (color azul), indicando que las imputaciones realizadas son eficientes y mantienen la variabilidad de los datos reales.

Figura 24. Densidad. Imputaciones IM (MICE-PMM) frente a datos reales: Nº de diseños únicos (10%)



De manera similar se imputaron los datos faltantes en las 3 mediciones del nº de errores perseverantes y se ha representado gráficamente la distribución de los datos imputados respecto a los datos reales:

Figura 25. Imputación datos faltantes por IM (MICE-PMM): Nº de errores perseverantes (10%)

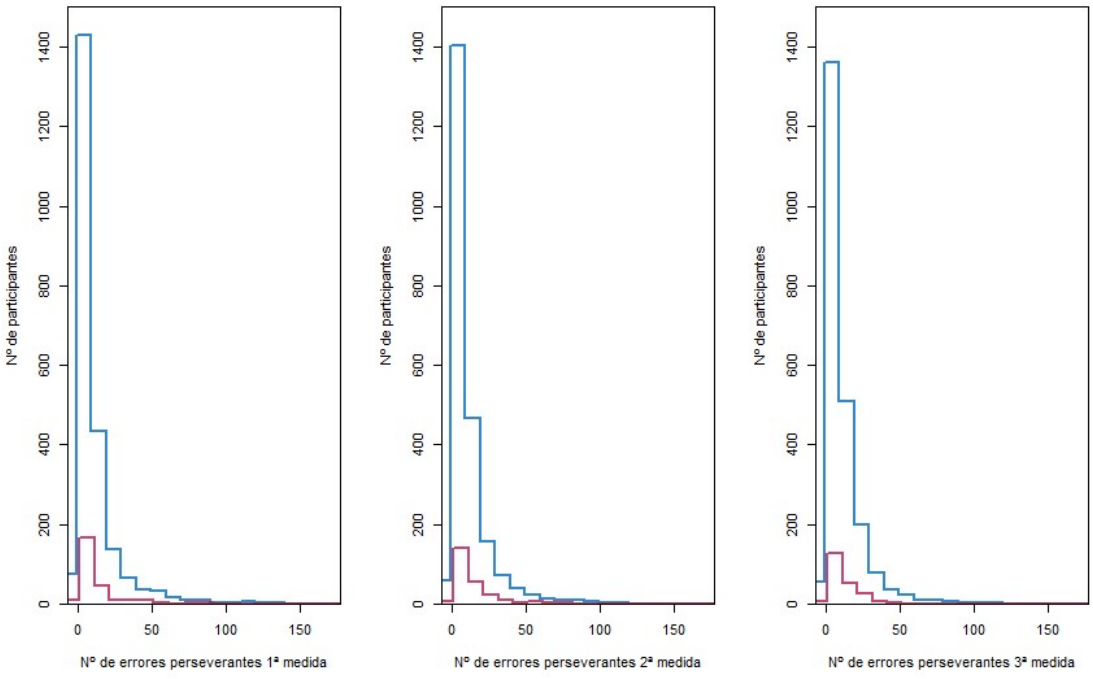


Figura 26. Valores reales vs imputados IM (MICE-PMM): Nº de errores perseverantes (10%)

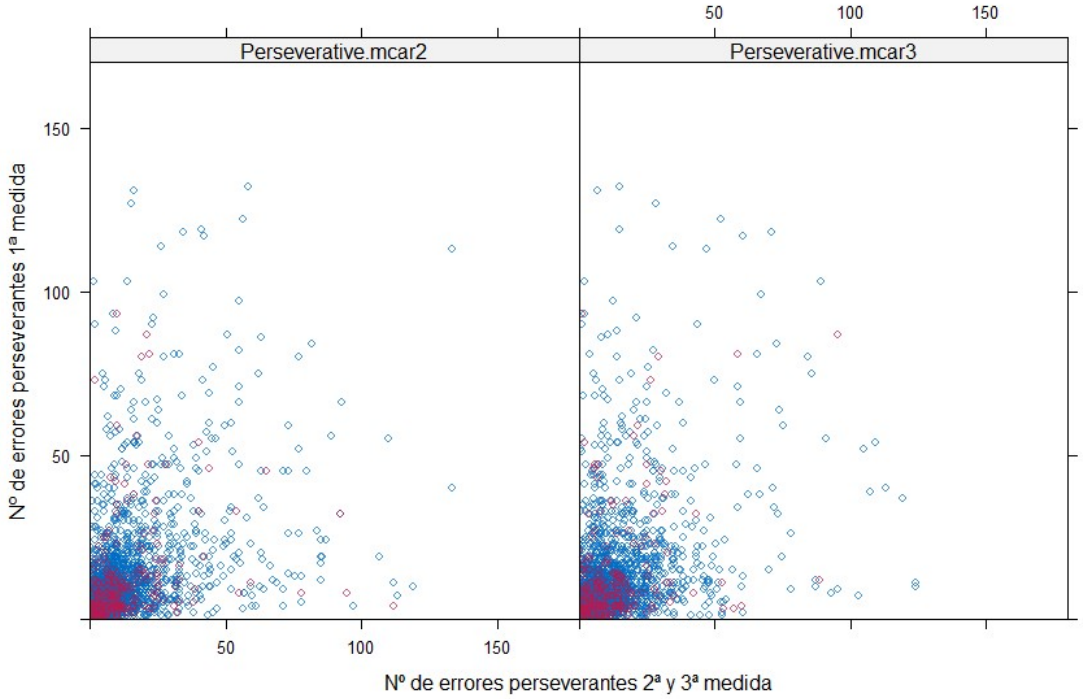
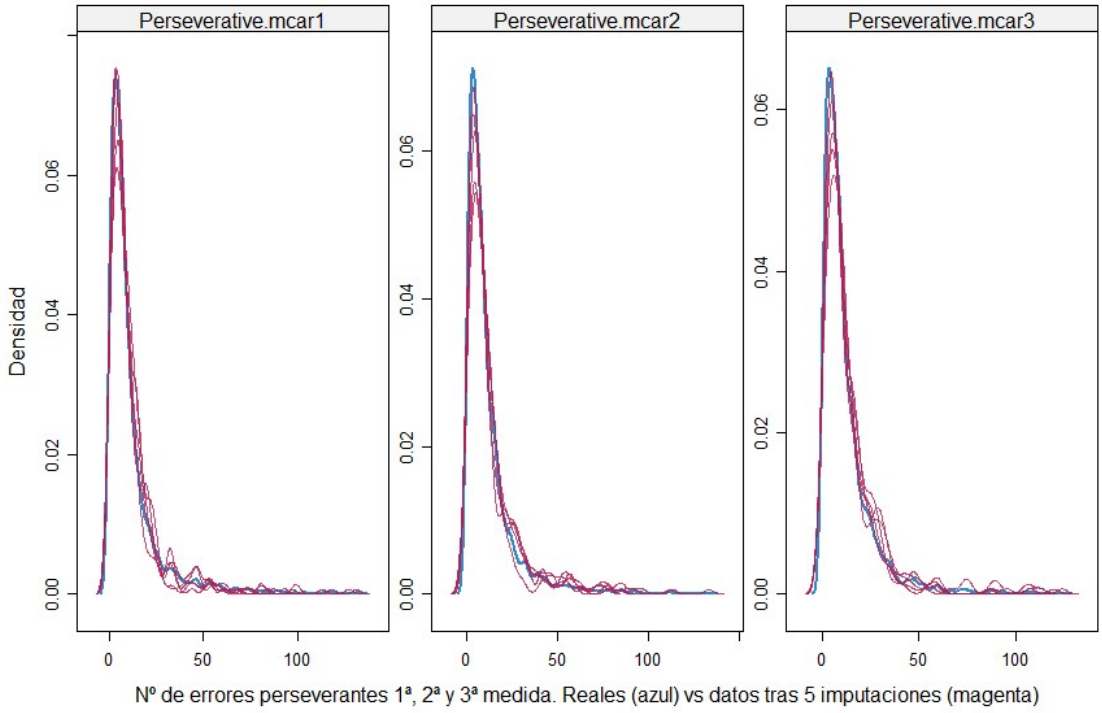


Figura 27. Densidad. Imputaciones IM (MICE-PMM) frente a datos reales: Nº de errores perseverantes (10%)



Tal y como se muestra en las figuras previas, para las 3 mediciones del nº de errores perseverantes, las distribuciones de los valores imputados (en color

magenta) se solapan con aquellas de los datos que no contenían datos faltantes (color azul), y apenas se perciben casos imputados de manera anómala, indicando que las imputaciones realizadas son eficientes y mantienen la variabilidad de los datos reales.

Para los escenarios con 20% y 30% de datos faltantes, pese al aumento sustancial de datos faltantes se obtuvieron resultados similares. A modo de ejemplo presentamos las funciones de densidad de los datos imputados para el nº de diseños únicos en ambos escenarios (el resto de gráficas está disponible en el informe estadístico).

Figura 28. Densidad. Imputaciones IM (MICE-PMM) frente a datos reales: Nº de diseños únicos (20%)

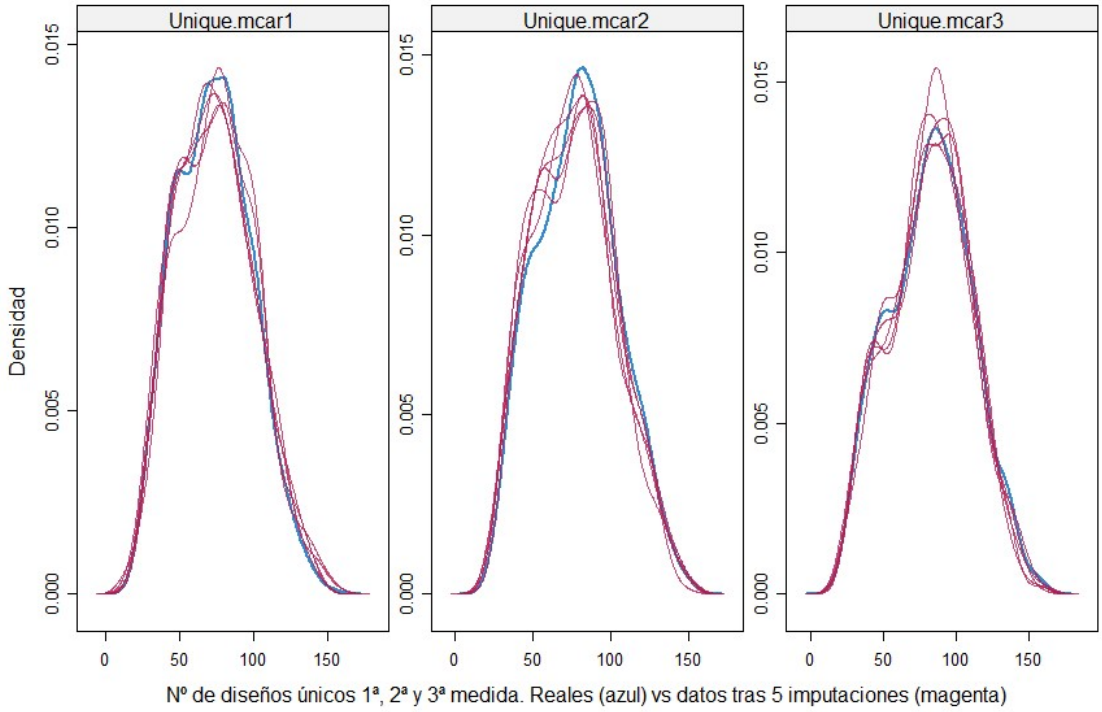
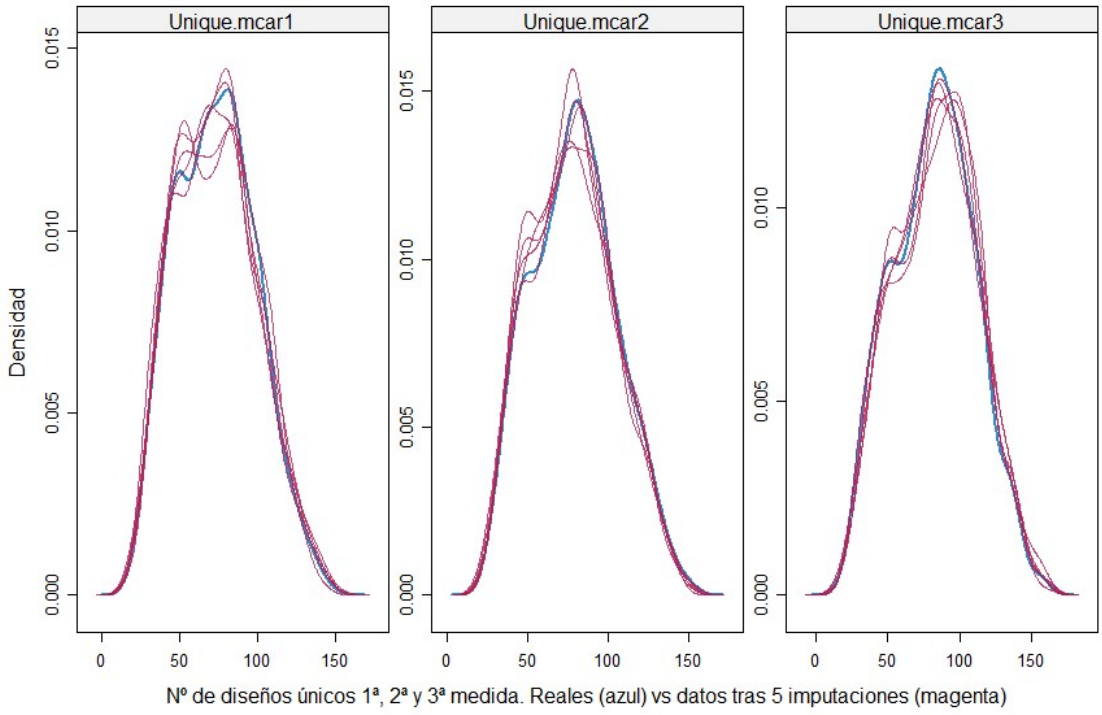


Figura 29. Densidad. Imputaciones IM (MICE-PMM) frente a datos reales: Nº de diseños únicos (30%)



5.6. Reproducción de los análisis originales publicados por los autores (base de datos completa).

El objetivo de las secciones finales del TFM, es reproducir los análisis originales utilizando las bases de datos obtenidas mediante cada uno de los cuatro métodos presentados en las secciones previas en cada uno de los tres escenarios (10%, 20% y 30% de datos faltantes) y compararlos con los resultados originales, para tratar de determinar cuál de los métodos obtiene resultados que se asemejen más a los originales. Es decir, no se entró a analizar en profundidad los métodos de análisis por los autores, se focalizó en la reproducibilidad de sus resultados tras el proceso de imputación de los datos faltantes.

A partir de la base de datos original se reprodujeron los resultados publicados:

Perfiles de los valores medios

En primer lugar se analizaron los valores medios y medianos del nº de diseños únicos y de errores perseverantes en cada visita, presentándolos junto con el nº de participantes analizables en cada visita. Para reproducir los análisis del artículo original (Marlise E. A. van Eersel et al.)[47], se utilizó el análisis de varianza (ANOVA) y el test de Kruskal Wallis, para analizar el cambio a lo largo de las 3 mediciones en el nº de diseños únicos y el nº de errores perseverantes, respectivamente. Los resultados coinciden con los publicados[47].

Tabla 2. Resumen resultados medidas consecutivas (datos originales)

Nº de diseños únicos, media (DE)	**1ª (2003-06)**	**2ª (2006-08)**	**3ª (2008-12)**	**p-valor (anova)**
Todos los participantes	73 (26)	79 (27)	83 (28)	p<0.001
Nº errores perseverantes, mediana (RIQ)	**1ª (2003-06)**	**2ª (2006-08)**	**3ª (2008-12)**	**p-valor (K-W)**
Todos los participantes	7 (3-13)	7 (4-14)	8 (4-15)	0.002

Modelos multivariados: Nº de diseños únicos

Para reproducir los resultados obtenidos por los autores se utiliza la función *lme* (paquete *nlme*)[55-57] que permite ajustar los efectos de ***modelos lineales mixtos*** y analizar los datos longitudinales. Los resultados obtenidos son muy similares a los que se presentan en el artículo (diferencias mínimas en algunos estimadores).

Se ajustan los 3 modelos especificados en el artículo tomando la variable nº de diseños únicos como variable dependiente:

- En el primer modelo se incluyeron *Edad, Género, Nivel de educación,*
- en el 2º se añade *Medida (nº consecutivo)*
- y en el 3º la interacción entre *edad y medida (nº consecutivo)*.

Tabla 3. Modelos lineales mixtos de los resultados RFFT: nº de diseños únicos

Variables	Modelo 1	Modelo 2	Modelo 3
	B, IC 95% (p-valor)	*B*, IC 95% (p-valor)	*B*, IC 95% (p-valor)
Edad	-1.12, -1.2 to -1.05 (<0.001)	-1.12, -1.2 to -1.05 (<0.001)	-0.66, -0.76 to -0.55 (<0.001)
Sexo: Hombre	Ref.	Ref.	Ref.
Sexo: Mujer	1.36, -0.06 to 2.79 (0.061)	1.38, -0.05 to 2.8 (0.058)	1.38, -0.04 to 2.8 (0.058)
NE: Escuela primaria	Ref.	Ref.	Ref.
NE: Secundaria inicial	6.69, 3.6 to 9.77 (<0.001)	6.64, 3.56 to 9.72 (<0.001)	6.63, 3.55 to 9.7 (<0.001)
NE: Secundaria superior	13.11, 9.97 to 16.25 (<0.001)	13.06, 9.93 to 16.19 (<0.001)	13.04, 9.91 to 16.17 (<0.001)
NE: Universitaria	25.41, 22.34 to 28.48 (<0.001)	25.36, 22.3 to 28.42 (<0.001)	22.28, 25.35 to 28.41 (<0.001)
Medida (nº consecutivo)		4.88, 4.48 to 5.29 (<0.001)	17.11, 15.09 to 19.14 (<0.001)
Edad x Medida (nº consecutivo)			-0.23, -0.27 to -0.19 (<0.001)

Se inspeccionó la información numérica que aporta la opción *"summary"* de la función *"anova"* acerca de la comparación de estos modelos. Dicha función proporciona los valores del criterio Akaike (AIC) que permite la comparación entre modelos, donde aquellos modelos con un AIC menor presentarían un mejor ajuste.

- El AIC del modelo 1 es 65749.1.
- El AIC del modelo 2 es 65238.
- El AIC del modelo 3 es 65103.7.
- la varianza residual del modelo 1 es 16.646 (la salida da 'sigma'), que es la estimación de la varianza de los residuos (como std.dev), por lo tanto reporta la variabilidad entre observaciones ("within-participante variability")[56,57], para el modelo 2 toma el valor 14.719 y para el modelo 3: 14.273.

Al comparar el modelo 1 con el 2, mediante los valores del criterio Akaike (AIC) se obtiene una diferencia de 511.18, que resulta estadísticamente significativa en beneficio del modelo 2, (p<0.001). Al comparar el modelo 2 con el 3, mediante los valores del criterio Akaike (AIC) se obtiene una diferencia de 134.25, que resulta estadísticamente significativa en beneficio del modelo 3, (p<0.001). ***En este caso el modelo que se selecciona para comparar con las bases de datos generadas con datos faltantes es el modelo 3.***

Como habían reportado los autores, en el modelo 3, el aumento en el nº de diseños únicos se asoció negativamente con la edad (B=-0.66, p<0.001) y en -0.23 por incremento de año, (la interacción de edad con medida, p<0.001). El resto de estimadores y p-valores coinciden, salvo pequeñas diferencias en algunos decimales, con los resultados publicados (ver tabla 4 de la publicación[47]).

Modelos multivariados: Nº de errores perseverantes

De igual manera se reprodujo el análisis de la evolución del nº de errores perseverantes. También Se ajustan los 3 modelos especificados en el artículo pero tomando la variable nº de errores perseverantes como variable dependiente. De nuevo se utilizó la función lme del paquete nlme para ajustar un modelo lineal mixto:

- En el primer modelo se incluyeron *Edad, Género, Nivel de educación,*
- en el 2º se añade *Medida (nº consecutivo)*
- y en el 3º la interacción entre *edad y medida (nº consecutivo)*.

Tabla 4. Modelos lineales mixtos de los resultados RFFT: nº de errores perseverantes

Variables	Modelo 1	Modelo 2	Modelo 3
	***B*, IC 95% (p-valor)**	***B*, IC 95% (p-valor)**	***B*, IC 95% (p-valor)**
Edad	-0.02, -0.06 to 0.03 (0.491)	-0.02, -0.06 to 0.03 (0.491)	0.08, 0 to 0.15 (0.046)
Sexo: Hombre	Ref.	Ref.	Ref.
Sexo: Mujer	2.64, 1.76 to 3.53 (<0.001)	2.64, 1.76 to 3.53 (<0.001)	2.64, 1.76 to 3.53 (<0.001)
NE: Escuela primaria	Ref.	Ref.	Ref.
NE: Secundaria inicial	1.33, -0.58 to 3.24 (0.171)	1.33, -0.58 to 3.24 (0.171)	1.33, -0.57 to 3.24 (0.171)
NE: Secundaria superior	-0.26, -2.21 to 1.68 (0.792)	-0.26, -2.21 to 1.68 (0.792)	-0.26, -2.21 to 1.68 (0.792)
NE: Universitaria	-1.25, -3.15 to 0.65 (0.196)	-1.25, -3.15 to 0.65 (0.196)	-3.15, -1.25 to 0.65 (0.196)
Medida (nº consecutivo)		0.09, -0.23 to 0.41 (0.562)	2.54, 0.89 to 4.19 (0.002)
Edad x Medida (nº consecutivo)			-0.05, -0.08 to -0.02 (0.003)

B: estimador; IC: Intervalo de confianza;

Se inspeccionó la información numérica que aporta la opción *"summary"* de la función *"anova"* acerca de la comparación de estos modelos. Dicha función proporciona los valores del criterio Akaike (AIC) que permite la comparación entre modelos, donde aquellos modelos con un AIC menor presentarían un mejor ajuste.

- El AIC del modelo 1 es 60778.2.
- El AIC del modelo 2 es 60781.6.
- El AIC del modelo 3 es 60781.3.
- la varianza residual del modelo 1 es 11.589 (la salida da 'sigma'), que es la estimación de la varianza de los residuos (como std.dev), por lo tanto reporta la variabilidad entre observaciones ("within-participante variability")[56,57], para el modelo 2 toma el valor 11.591 y para el modelo 3: 11.573.

Al comparar el modelo 1 con el 2, mediante los valores del criterio Akaike (AIC) se obtiene una diferencia de -3.45, que no resultó estadísticamente significativa en beneficio del modelo 2, p=0.228. Al comparar el modelo 2 con el 3, mediante los valores del criterio Akaike (AIC) se obtiene una diferencia de 0.33, que no resultó estadísticamente significativa en beneficio del modelo 3, p=0.127. Al comparar el modelo 1 con el 3, mediante los valores del criterio Akaike (AIC) se obtiene una diferencia de -3.12, que no resultó estadísticamente significativa en beneficio del modelo 3, p=0.643.

En este caso el modelo que se selecciona para comparar con las bases de datos generadas con datos faltantes es el modelo 1.

Como habían reportado los autores (ver tabla 5 de la publicación[47]), en el modelo 1, el aumento en el nº de errores perseverantes no se asoció negativamente con la edad (B=-0.02, p=0.491) y solo la variable Género presentó diferencias entre hombre y mujer (B_{mujer}=2.64, p<0.001). El resto de estimadores y p-valores coinciden, salvo pequeñas diferencias en algunos decimales, con los resultados publicados (ver tabla 4 de la publicación[47]).

5.7.Resumen de resultados: Datos originales frente a imputaciones con 10% de datos faltantes.

En el informe estadístico está detallada paso a paso la reproducción de los resultados originales en los distintos escenarios de datos faltantes (10, 20 y 30%) para ambas variables dependientes. A continuación se presentan varias tablas que resumen dichos resultados.

En el escenario de **10% de datos faltantes**, al analizar mediante el análisis de varianza la evolución del nº de diseños únicos en cada una de las tres medidas, los resultados fueron muy similares a los datos originales. Quizás lo más destacable, pero esperable es que en el caso de la imputación por la media se obtienen menores desviaciones estándar, menor variabilidad.

En cuanto a la evolución del nº de errores perseverantes (se utilizó el test no paramétrico de Kruskal-Wallis), los valores medianos y el RIQ, tampoco presenta diferencias muy notables, quizás a excepción de nuevo de la imputación por la media que presenta diferencias de una unidad o incluso 1.5 en alguna de las medidas.

Tabla 5. Resumen resultados medidas consecutivas del nº de diseños únicos y nº de errores perseverantes (10%)

Análisis: Nº de diseños únicos, media(DE)	**1ª (2003-06)**	**2ª (2006-08)**	**3ª (2008-12)**	**p-valor (anova)**
Datos originales	72.9 (25.6)	78.7 (26.6)	82.7 (28.4)	p<0.001
Eliminación de los casos (listwise)	73.6 (25.7)	79 (26.5)	82.9 (28.3)	p<0.001
Reemplazamiento por la media	73.3 (24.3)	78.8 (25.1)	82.6 (26.9)	p<0.001
Regresión	73.2 (25.1)	78.6 (26)	82.6 (27.9)	p<0.001
Imputación múltiple (*PMM*)	73.2 (25.7)	78.6 (26.3)	82.6 (28.5)	p<0.001
Análisis: Nº errores perseverantes, mediana(RIQ)	**1ª (2003-06)**	**2ª (2006-08)**	**3ª (2008-12)**	**p-valor (K-W)**
Datos originales	7 (3-13)	7 (3.5-14)	8 (4-15)	0.002
Eliminación de los casos (listwise)	7 (3-13)	7 (3.5-14)	8 (4-15)	0.002
Reemplazamiento por la media	8 (4-12)	8.5 (4-13)	9 (4-14)	0.003
Regresión	7 (4-13.2)	8 (4-14)	8.5 (4-15)	0.001
Imputación múltiple (*PMM*)	7 (3-13)	7.5 (3.5-14.5)	8 (4-15)	<0.001

DE: Desviación estándard; RIQ: rango inter-cuartílico; K-W: Kruskal-Wallis; PMM: Predictive Mean Matching o Equiparación de media predictiva

A modo de análisis de sensibilidad se compararon los resultados de los modelos con los 4 métodos de tratamiento de los datos faltantes, con los resultados originales. En el caso del nº de diseños únicos el 3er modelo, que incluye Edad, Género, Nivel de educación, Medida (nº consecutivo) y la interacción entre edad y medida (nº consecutivo), fue el que presentó un menor valor del criterio Akaike-AIC (hallándose diferencias estadísticamente significativas respecto a los modelos más parsimoniosos), y es el utilizado para hacer la comparación con las 4 simulaciones.

Tabla 6. Resumen comparativo resultados de los modelos: nº de diseños únicos (10%)

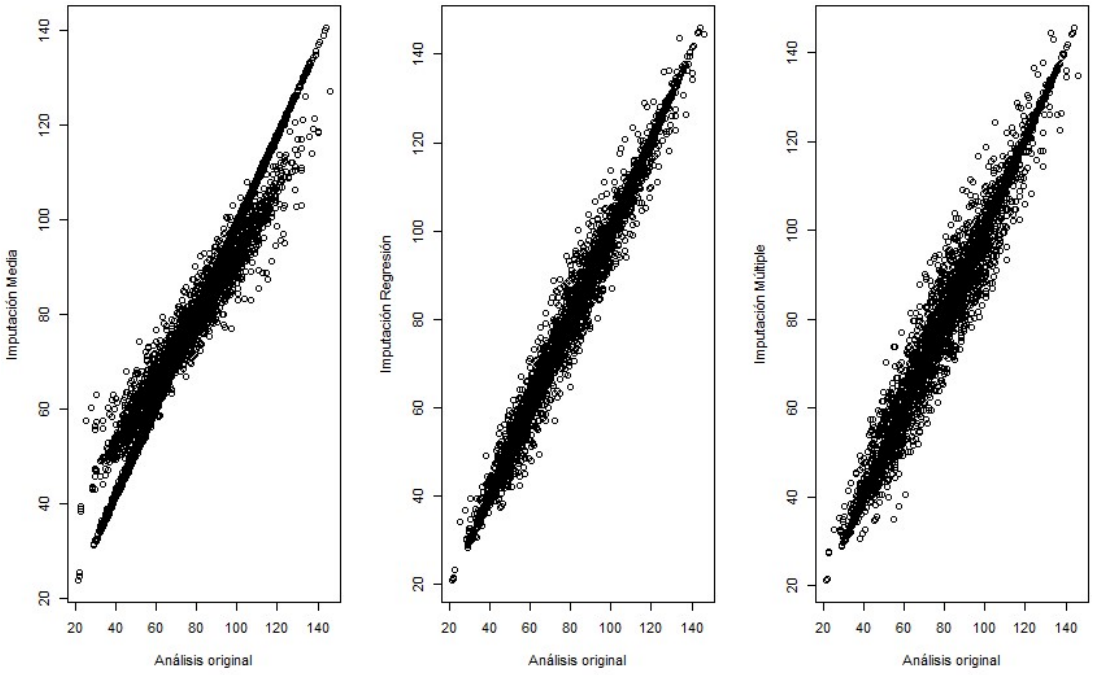
Tipo de análisis	Análisis original	Listwise	Media	Regresión	IM (5, PMM)
Modelo	**Modelo 3**	**Modelo 3**	**Modelo 3**	**Modelo 3**	**Modelo 3**
Variables	***B* (DE), IC95% (p-valor)**	***B* (DE), IC95% (p-valor)**	***B* (DE), IC95% (p-valor)**	***B* (DE), IC95% (p-valor)**	***B* (DE), IC95% (p-valor)**
Edad	-0.66 (0.053), -0.76 to -0.55 (<0.001)	-0.69 (0.062) , -0.81 to -0.57 (<0.001)	-0.61 (0.053), -0.72 to -0.51 (<0.001)	-0.67 (0.051), -0.76 to -0.57 (<0.001)	-0.66 (0.053), -0.76 to -0.56 (<0.001)
Sexo: Hombre	Ref.	Ref.	Ref.	Ref.	Ref.
Sexo: Mujer	1.38 (0.725), -0.04 to 2.8 (0.058)	1.14 (0.853) , -0.53 to 2.82 (0.18)	1.01 (0.68), -0.33 to 2.34 (0.139)	1.27 (0.719), -0.14 to 2.68 (0.077)	1.16 (0.723), -0.25 to 2.58 (0.108)
NE: Escuela primaria	Ref.	Ref.	Ref.	Ref.	Ref.
NE: Secundaria inicial	6.63 (1.568), 3.55 to 9.7 (<0.001)	6.66 (1.866), 3 to 10.32 (<0.001)	5.63 (1.469), 2.75 to 8.51 (<0.001)	6.57 (1.555), 3.52 to 9.62 (<0.001)	6.7 (1.562), 3.64 to 9.77 (<0.001)
NE: Secundaria superior	13.04 (1.597), 9.91 to 16.17 (<0.001)	12.24 (1.896), 8.52 to 15.96 (<0.001)	10.94 (1.496), 8.01 to 13.87 (<0.001)	12.76 (1.583), 9.66 to 15.86 (<0.001)	12.84 (1.59), 9.72 to 15.95 (<0.001)
NE: Universitaria	25.35 (1.561), 22.28 to 28.41 (<0.001)	24.56 (1.846), 20.94 to 28.18 (<0.001)	21.97 (1.462), 19.1 to 24.84 (<0.001)	25.04 (1.547), 22 to 28.07 (<0.001)	25.14 (1.555), 22.09 to 28.19 (<0.001)
Medida (nº consecutivo)	17.11 (1.032), 15.09 to 19.14 (<0.001)	24.56 (1.207), 13.97 to 18.71(0.18)	15.2 (1.098), 13.04 to 17.35(<0.001)	16.64 (0.952), 14.77 to 18.51(<0.001)	16.9 (1.021), 14.9 to 18.9(<0.001)
Edad x Medida (nº consecutivo)	-0.23 (0.019), -0.27 to -0.19 (<0.001)	-0.22 (0.023), -0.27 to -0.18 (<0.001)	-0.2 (0.02), -0.24 to -0.16 (<0.001)	-0.23 (0.018), -0.26 to -0.19 (<0.001)	-0.23 (0.019), -0.27 to -0.19 (<0.001)
AIC	65103.7	47020.5*	65528.5	64291.2	65187.4
Varianza residual	14.3	14.2*	15.1	13.2	14.1

DE: Desviación estándard; IM: Imputación Múltiple; PMM: Predictive Mean Matching o Equiparación de media predictiva; AIC: Criterio Akaike.
*: Se *ha de tener en cuenta que este caso presenta una **n** distinta al resto por lo que el valor del AIC no es comparable al resto de modelos.*

En cuanto al valor del AIC, el que más se aproxima al original, es el obtenido mediante la IM (PMM), aunque la varianza residual es ligeramente más próxima la obtenida mediante el método listwise. Por lo que se refiere a los estimadores *B* y sus *DE*, es de nuevo el modelo IM (PMM) el que proporciona en general estimadores más cercanos a los originales, seguido en esta ocasión por la regresión lineal. En todos los casos los p-valores se mantienen significativos cuando lo son en el modelo original.

También se han analizado las correlaciones de los valores ajustados para para los modelos que contienen el mismo tamaño muestral respecto del análisis original (los resultados de las correlaciones se pueden hallar en el informe estadísticos adjunto) y se representan gráficamente en la siguiente figura. Como se puede observar de nuevo los modelos resultantes de la imputación por regresión e IM presentan una correlación mayor (aunque hay mayor variabilidad en la IM).

Figura 30. Correlaciones valores ajustados diseños únicos (Modelo 3) (10%)



En el caso del nº de errores perseverantes el 1er modelo, que incluye Edad, Género y Nivel de educación, es el utilizado para hacer la comparación con las 4 simulaciones.

Tabla 7. Resumen comparativo resultados de los modelos: nº de errores perseverantes (10%)

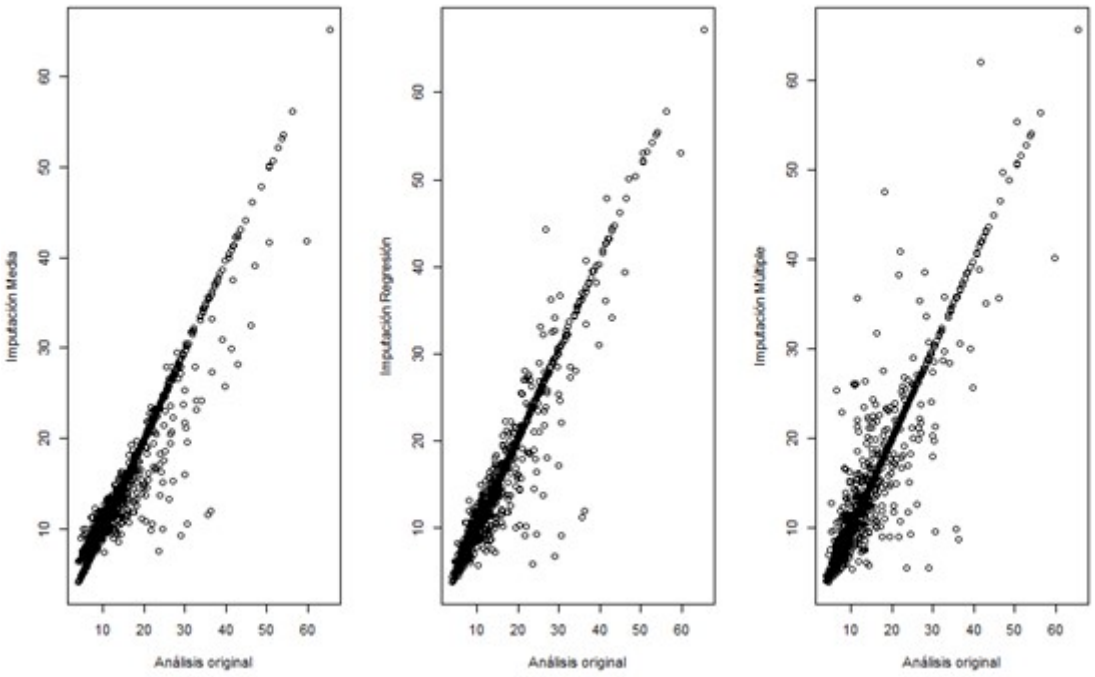
Tipo de análisis	**Análisis original**	**Listwise**	**Media**	**Regresión**	**IM (5, PMM)**
Modelo	**Modelo 1**	**Modelo 1**	**Modelo 1**	**Modelo 1**	**Modelo 1**
Variables	***B* (DE), IC95% (p-valor)**	***B* (DE), IC95% (p-valor)**	***B* (DE), IC95% (p-valor)**	***B* (DE), IC95% (p-valor)**	***B* (DE), IC95% (p-valor)**
Edad	-0.02 (0.023), -0.06 to 0.03 (0.491)	-0.02 (0.026), -0.07 to 0.03(0.398)	-0.015 (0.021), – to – (0.489)	-0.02 (0.022), -0.06 to 0.03 (0.47)	-0.01 (0.023), -0.06 to 0.03 (0.58)
Sexo: Hombre	Ref.	Ref.	Ref.	Ref.	Ref.
Sexo: Mujer	2.64 (0.451), 1.76 to 3.53 (<0.001)	2.33 (0.526), 1.3 to 3.36 (<0.001)	2.261 (0.417), – to – (<0.001)	2.52 (0.44), 1.65 to 3.38 (<0.001)	2.56 (0.452), 1.68 to 3.45 (<0.001)
NE: Escuela primaria	Ref.	Ref.	Ref.	Ref.	Ref.
NE: Secundaria inicial	1.33 (0.974), -0.58 to 3.24 (0.171)	0.75 (1.133), -1.47 to 2.98 (0.506)	1.272 (0.902), – to – (0.158)	1.54 (0.952), -0.33 to 3.41 (0.106)	1.63 (0.976), -0.29 to 3.54 (0.095)
NE: Secundaria superior	-0.26 (0.992), -2.21 to 1.68 (0.792)	-0.37 (1.154), -2.63 to 1.9 (0.751)	-0.156 (0.918), – to – (0.865)	-0.14 (0.969), -2.04 to 1.76 (0.883)	-0.04 (0.994), -1.99 to 1.91 (0.968)
NE: Universitaria	-1.25 (0.97), -3.15 to 0.65 (0.196)	-1.54 (1.129), -3.75 to 0.68 (0.174)	0.898 (0.898), – to – (0.231)	-1.16 (0.948), -3.01 to 0.7 (0.223)	-1.02 (0.972), -2.93 to 0.88 (0.293)
AIC	60778.2	44393.4*	60020.2	59790	60753.2
Varianza residual	11.6	11.3*	11	10.8	11.6

DE: Desviación estándard; IM: Imputación Múltiple; PMM: Predictive Mean Matching o Equiparación de media predictiva; AIC: Criterio Akaike.
*: Se *ha de tener en cuenta que este caso presenta una **n** distinta al resto por lo que el valor del AIC no es comparable al resto de modelos.*

En cuanto al valor del AIC, el que más se aproxima al original, esta vez con diferencia, es el obtenido mediante la IM (PMM), y también en cuanto al valor de la varianza residual. Por lo que se refiere a los estimadores *B* y sus *DE*, ninguna de las 4 opciones proporciona en general estimadores más cercanos a los originales respecto al resto, aunque el modelo IM (PMM) proporciona DE que más ajustadas a los valores del análisis original. En todos los casos los p-valores se mantienen significativos cuando lo son en el modelo original (solo para la variable género).

También se han analizado las correlaciones de los valores ajustados para para los modelos que contienen el mismo tamaño muestral respecto del análisis original (los resultados de las correlaciones se pueden hallar en el informe estadísticos adjunto) y se representan gráficamente en la siguiente figura. Como se puede observar el modelo resultante de la imputación por regresión muestra una correlación mayor, presentándose mayor variabilidad en la IM.

Figura 31. Correlaciones valores ajustados errores perseverantes (Modelo 1) (10%)



5.8.Resumen de resultados: Datos originales frente a imputaciones con 20% de datos faltantes.

En el escenario de **30% de datos faltantes**, a modo de análisis de sensibilidad se compararon los resultados de los modelos con los 4 métodos de tratamiento de los datos faltantes, con los resultados originales. En el caso del nº de diseños únicos el 3er modelo, que incluye Edad, Género, Nivel de educación, Medida (nº consecutivo) y la interacción entre edad y medida (nº consecutivo), fue el que presentó un menor valor del criterio Akaike-AIC (hallándose diferencias estadísticamente significativas respecto a los modelos más parsimoniosos), y es el utilizado para hacer la comparación con las 4 simulaciones.

En el escenario de **20% de datos faltantes**, a modo de análisis de sensibilidad se compararon los resultados de los modelos con los 4 métodos de tratamiento de los datos faltantes, con los resultados originales. En el caso del nº de diseños únicos el 3er modelo, que incluye Edad, Género, Nivel de educación, Medida (nº consecutivo) y la interacción entre edad y medida (nº consecutivo), fue el que presentó un menor valor del criterio Akaike-AIC (hallándose diferencias estadísticamente significativas respecto a los modelos más parsimoniosos), y es el utilizado para hacer la comparación con las 4 simulaciones.

Tabla 8. Resumen comparativo resultados de los modelos: nº de diseños únicos (20%)

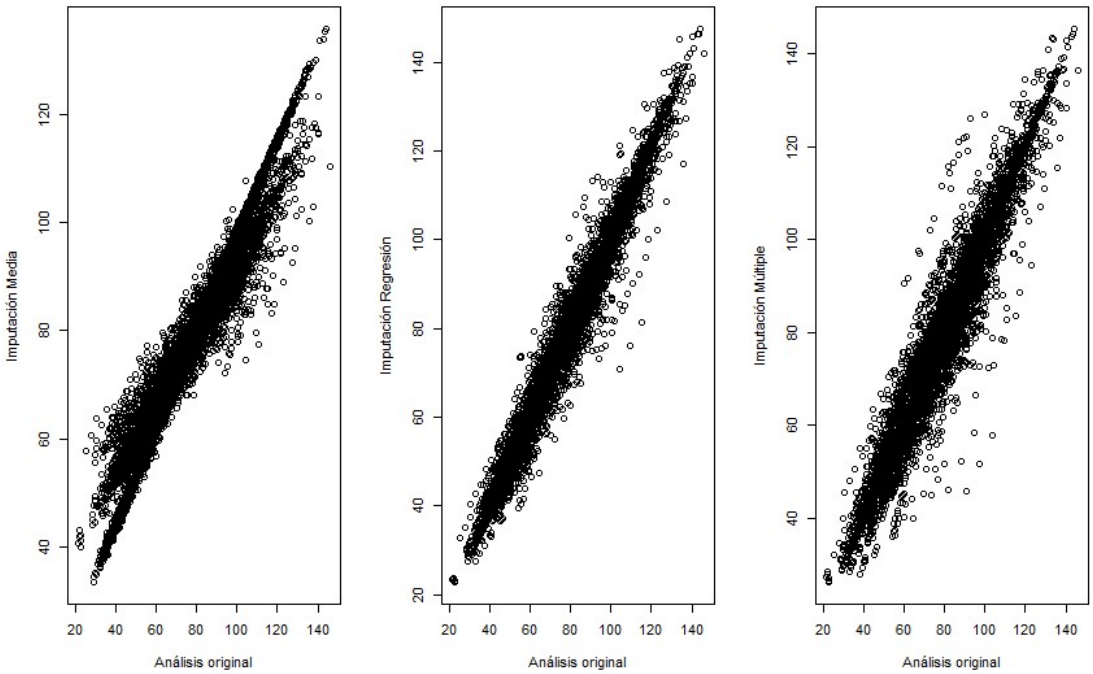
Tipo de análisis	Análisis original	Listwise	Media	Regresión	IM (5, PMM)
Modelo	**Modelo 3**	**Modelo 3**	**Modelo 3**	**Modelo 3**	**Modelo 3**
Variables	***B* (DE), IC95% (p-valor)**	***B* (DE), IC95% (p-valor)**	***B* (DE), IC95% (p-valor)**	***B* (DE), IC95% (p-valor)**	***B* (DE), IC95% (p-valor)**
Edad	-0.66 (0.053), -0.76 to -0.55 (<0.001)	-0.69 (0.075) , -0.83 to -0.54 (<0.001)	-0.54 (0.053), -0.65 to -0.44 (<0.001)	-0.64 (0.049), -0.74 to -0.55 (<0.001)	-0.68 (0.053), -0.78 to -0.57 (<0.001)
Sexo: Hombre	Ref.	Ref.	Ref.	Ref.	Ref.
Sexo: Mujer	1.38 (0.725), -0.04 to 2.8 (0.058)	0.68 (1.023) , -1.33 to 2.69 (0.506)	0.97 (0.629), -0.26 to 2.2 (0.123)	1.39 (0.712), -0.01 to 2.79 (0.051)	1.25 (0.733), -0.19 to 2.69 (0.088)
NE: Escuela primaria	Ref.	Ref.	Ref.	Ref.	Ref.
NE: Secundaria inicial	6.63 (1.568), 3.55 to 9.7 (<0.001)	6.87 (2.302), 2.35 to 11.38 (0.003)	4.95 (1.36), 2.28 to 7.62 (<0.001)	6.48 (1.54), 3.46 to 9.5 (<0.001)	6.74 (1.585), 3.64 to 9.85 (<0.001)
NE: Secundaria superior	13.04 (1.597), 9.91 to 16.17 (<0.001)	11.43 (2.344), 6.84 to 16.03 (<0.001)	9.6 (1.385), 6.88 to 12.31 (<0.001)	12.81 (1.568), 9.73 to 15.88 (<0.001)	12.62 (1.614), 9.46 to 15.79 (<0.001)
NE: Universitaria	25.35 (1.561), 22.28 to 28.41 (<0.001)	23.44 (2.278), 18.97 to 27.91 (<0.001)	19.38 (1.354), 16.72 to 22.03 (<0.001)	25.11 (1.533), 22.11 to 28.12 (<0.001)	25.6 (1.578), 22.51 to 28.69 (<0.001)
Medida (nº consecutivo)	17.11 (1.032), 15.09 to 19.14 (<0.001)	23.44 (1.465), 14.6 to 20.35(0.506)	13.92 (1.149), 11.67 to 16.18(<0.001)	16.8 (0.88), 15.07 to 18.52(<0.001)	15.84 (1.02), 13.85 to 17.84(<0.001)
Edad x Medida (nº consecutivo)	-0.23 (0.019), -0.27 to -0.19 (<0.001)	-0.24 (0.027), -0.3 to -0.19 (<0.001)	-0.18 (0.021), -0.22 to -0.13 (<0.001)	-0.23 (0.016), -0.26 to -0.2 (<0.001)	-0.21 (0.019), -0.25 to -0.18 (<0.001)
AIC	65103.7	32348.2*	65665	63483.5	65193.4
Varianza residual	14.3	14.3*	15.8	12.1	14.1

DE: Desviación estándard; IM: Imputación Múltiple; PMM: Predictive Mean Matching o Equiparación de media predictiva; AIC: Criterio Akaike.
*: Se *ha de tener en cuenta que este caso presenta una **n** distinta al resto por lo que el valor del AIC no es comparable al resto de modelos.*

En cuanto al valor del AIC, el que más se aproxima al original, es el obtenido mediante la IM (PMM), aunque la varianza residual es más próxima la obtenida mediante el método listwise. Por lo que se refiere a los estimadores *B* y sus *DE*, es la imputación por regresión lineal, la que proporciona en general estimadores *B* más cercanos a los originales, seguido en esta ocasión el modelo IM (PMM), que también proporciona *DE* más cercanas. En todos los casos los p-valores se mantienen significativos cuando lo son en el modelo original, excepto en algún caso del modelo *listwise*.

También se han analizado las correlaciones de los valores ajustados para para los modelos que contienen el mismo tamaño muestral respecto del análisis original (los resultados de las correlaciones se pueden hallar en el informe estadístico adjunto) y se representan gráficamente en la siguiente figura. Como se puede observar de nuevo los modelos resultantes de la imputación por regresión e IM presentan una correlación mayor (aunque hay mayor variabilidad en la IM).

Figura 32. Correlaciones valores ajustados diseños únicos (Modelo 3) (20%)



En el caso del nº de errores perseverantes el 1er modelo, que incluye Edad, Género y Nivel de educación, es el utilizado para hacer la comparación con las 4 simulaciones.

Tabla 9. Resumen comparativo resultados de los modelos: nº de errores perseverantes (20%)

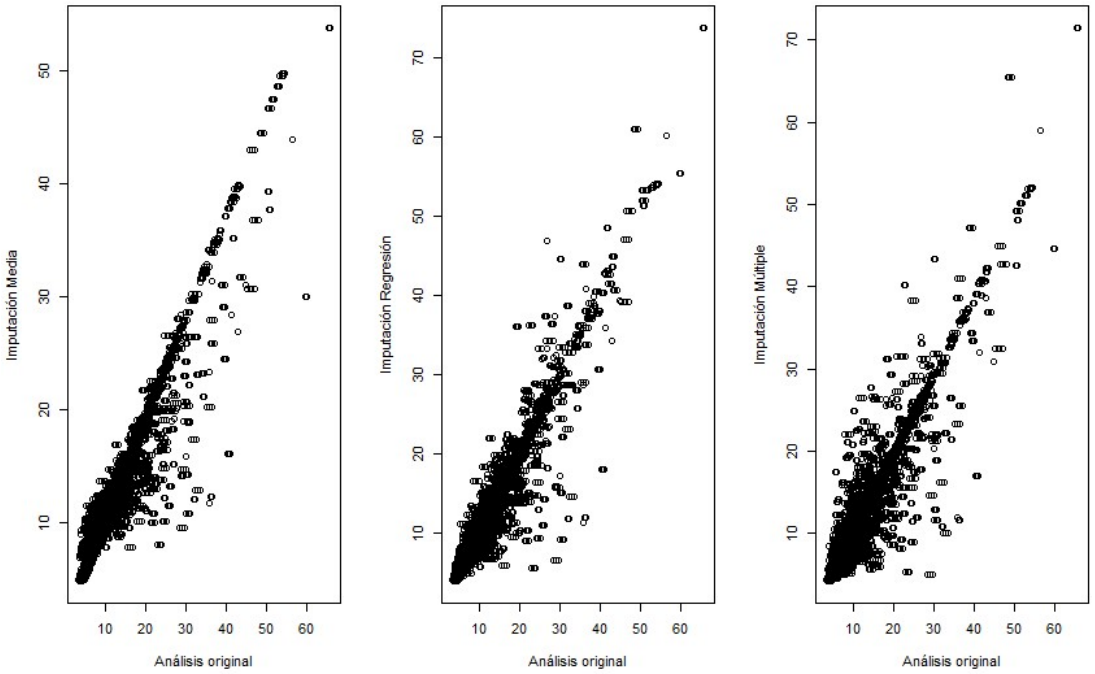
Tipo de análisis	Análisis original	Listwise	Media	Regresión	IM (5, PMM)
Modelo	**Modelo 1**	**Modelo 1**	**Modelo 1**	**Modelo 1**	**Modelo 1**
Variables	***B* (DE), IC95% (p-valor)**	***B* (DE), IC95% (p-valor)**	***B* (DE), IC95% (p-valor)**	***B* (DE), IC95% (p-valor)**	***B* (DE), IC95% (p-valor)**
Edad	-0.02 (0.023), -0.06 to 0.03 (0.491)	-0.02 (0.031), -0.08 to 0.04(0.513)	-0.01 (0.019), – to – (0.617)	-0.01 (0.022), -0.06 to 0.03 (0.595)	-0.01 (0.022), -0.06 to 0.03 (0.572)
Sexo: Hombre	Ref.	Ref.	Ref.	Ref.	Ref.
Sexo: Mujer	2.64 (0.451), 1.76 to 3.53 (<0.001)	2.44 (0.615), 1.23 to 3.64 (<0.001)	2.089 (0.385), – to – (<0.001)	2.66 (0.445), 1.78 to 3.53 (<0.001)	2.61 (0.446), 1.74 to 3.49 (<0.001)
NE: Escuela primaria	Ref.	Ref.	Ref.	Ref.	Ref.
NE: Secundaria inicial	1.33 (0.974), -0.58 to 3.24 (0.171)	-0.44 (1.33), -3.05 to 2.17 (0.739)	1.146 (0.832), – to – (0.169)	1.79 (0.963), -0.1 to 3.68 (0.063)	1.88 (0.963), -0.01 to 3.76 (0.052)
NE: Secundaria superior	-0.26 (0.992), -2.21 to 1.68 (0.792)	-1.5 (1.347), -4.14 to 1.14 (0.265)	-0.078 (0.847), – to – (0.927)	0.13 (0.98), -1.79 to 2.06 (0.892)	0.15 (0.981), -1.78 to 2.07 (0.882)
NE: Universitaria	-1.25 (0.97), -3.15 to 0.65 (0.196)	-2.87 (1.319), -5.46 to -0.28 (0.03)	0.829 (0.829), – to – (0.262)	-0.88 (0.959), -2.76 to 1 (0.358)	-0.72 (0.959), -2.6 to 1.16 (0.454)
AIC	60778.2	31082.9*	59336.4	58776	60347.5
Varianza residual	11.6	11.6*	10.8	10.3	11.4

DE: Desviación estándard; IM: Imputación Múltiple; PMM: Predictive Mean Matching o Equiparación de media predictiva; AIC: Criterio Akaike.
*: Se *ha de tener en cuenta que este caso presenta una* ***n*** *distinta al resto por lo que el valor del AIC no es comparable al resto de modelos.*

En cuanto al valor del AIC, el que más se aproxima al original, esta vez con diferencia, es el obtenido mediante la IM (PMM), aunque la varianza residual es más próxima la obtenida mediante el método listwise. Por lo que se refiere a los estimadores *B* y sus *DE*, el modelo IM (PMM) y el de regresión lineal son los que proporcionan *B* más ajustadas a los valores del análisis original con DE también muy parecidas y relativamente cercanas a las originales. En todos los casos los p-valores se mantienen significativos cuando lo son en el modelo original (solo para la variable género).

También se han analizado las correlaciones de los valores ajustados para para los modelos que contienen el mismo tamaño muestral respecto del análisis original (los resultados de las correlaciones se pueden hallar en el informe estadísticos adjunto) y se representan gráficamente en la siguiente figura. Como se puede observar el modelo resultante de la imputación por regresión muestra una correlación mayor, presentándose mayor variabilidad en la IM.

Figura 33. Correlaciones valores ajustados errores perseverantes (Modelo 1) (20%)


Imputación Media
Análisis original
Imputación Regresión
Análisis original
Imputación Múltiple
Análisis original

5.9.Resumen de resultados: Datos originales frente a imputaciones con 30% de datos faltantes.

En el escenario de **30% de datos faltantes**, a modo de análisis de sensibilidad se compararon los resultados de los modelos con los 4 métodos de tratamiento de los datos faltantes, con los resultados originales. En el caso del nº de diseños únicos el 3er modelo, que incluye Edad, Género, Nivel de educación, Medida (nº consecutivo) y la interacción entre edad y medida (nº consecutivo), fue el que presentó un menor valor del criterio Akaike-AIC (hallándose diferencias estadísticamente significativas respecto a los modelos más parsimoniosos), y es el utilizado para hacer la comparación con las 4 simulaciones.

Tabla 10. Resumen comparativo resultados de los modelos: nº de diseños únicos (30%)

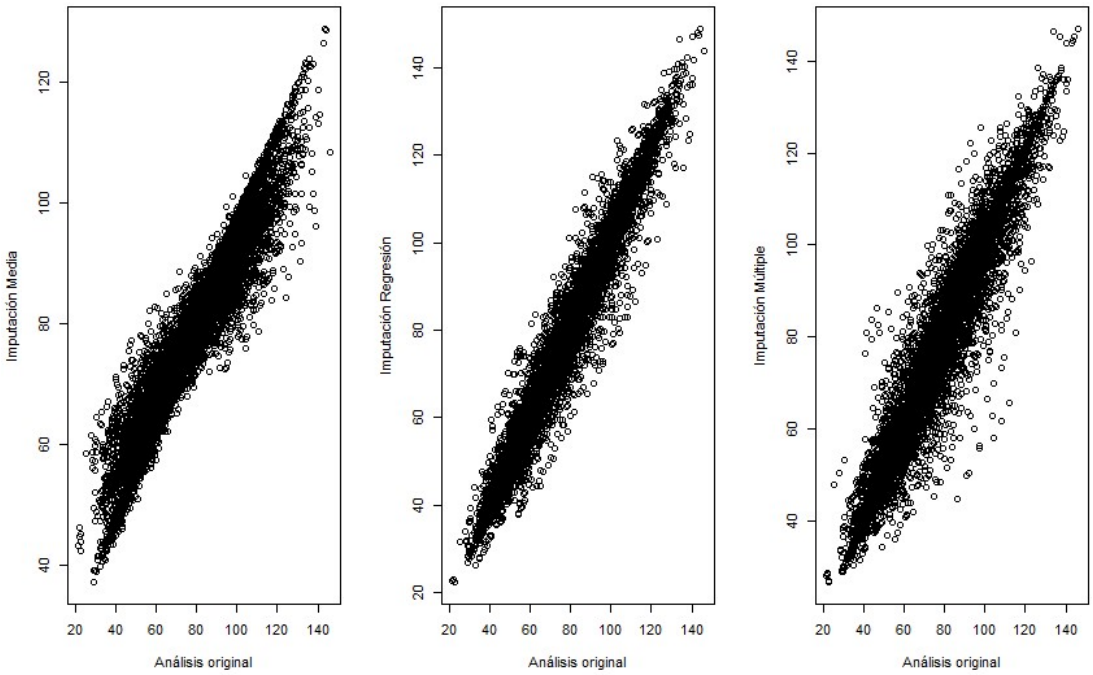
Tipo de análisis	Análisis original	Listwise	Media	Regresión	IM (5, PMM)
Modelo	**Modelo 3**	**Modelo 3**	**Modelo 3**	**Modelo 3**	**Modelo 3**
Variables	***B* (DE), IC95% (p-valor)**	***B* (DE), IC95% (p-valor)**	***B* (DE), IC95% (p-valor)**	***B* (DE), IC95% (p-valor)**	***B* (DE), IC95% (p-valor)**
Edad	-0.66 (0.053), -0.76 to -0.55 (<0.001)	-0.58 (0.093) , -0.76 to -0.4 (<0.001)	-0.5 (0.053), -0.6 to -0.4 (<0.001)	-0.66 (0.047), -0.75 to -0.57 (<0.001)	-0.64 (0.053), -0.75 to -0.54 (<0.001)
Sexo: Hombre	Ref.	Ref.	Ref.	Ref.	Ref.
Sexo: Mujer	1.38 (0.725), -0.04 to 2.8 (0.058)	1.24 (1.245) , -1.2 to 3.69 (0.318)	1 (0.575), -0.12 to 2.13 (0.081)	1.6 (0.708), 0.21 to 2.99 (0.024)	1.76 (0.726), 0.34 to 3.18 (0.015)
NE: Escuela primaria	Ref.	Ref.	Ref.	Ref.	Ref.
NE: Secundaria inicial	6.63 (1.568), 3.55 to 9.7 (<0.001)	8.09 (2.975), 2.26 to 13.93 (0.007)	4.21 (1.242), 1.77 to 6.64 (0.001)	6.23 (1.531), 3.23 to 9.23 (<0.001)	6.05 (1.569), 2.98 to 9.13 (<0.001)
NE: Secundaria superior	13.04 (1.597), 9.91 to 16.17 (<0.001)	11.43 (3.021), 5.5 to 17.36 (<0.001)	7.91 (1.265), 5.43 to 10.39 (<0.001)	12.61 (1.559), 9.55 to 15.67 (<0.001)	12.15 (1.597), 9.02 to 15.28 (<0.001)
NE: Universitaria	25.35 (1.561), 22.28 to 28.41 (<0.001)	25.03 (2.951), 19.23 to 30.82 (<0.001)	16.88 (1.237), 14.45 to 19.3 (<0.001)	25.06 (1.524), 22.07 to 28.04 (<0.001)	24.48 (1.562), 21.42 to 27.54 (<0.001)
Medida (nº consecutivo)	17.11 (1.032), 15.09 to 19.14 (<0.001)	25.03 (1.819), 14.99 to 22.13(0.318)	12.03 (1.183), 9.71 to 14.34(<0.001)	16.47 (0.805), 14.9 to 18.05(<0.001)	17.26 (1.038), 15.22 to 19.29(<0.001)
Edad x Medida (nº consecutivo)	-0.23 (0.019), -0.27 to -0.19 (<0.001)	-0.27 (0.034), -0.33 to -0.2 (<0.001)	-0.15 (0.022), -0.19 to -0.1 (<0.001)	-0.23 (0.015), -0.26 to -0.2 (<0.001)	-0.23 (0.019), -0.27 to -0.2 (<0.001)
AIC	65103.7	21661.3*	65482.6	62548.6	65369.7
Varianza residual	14.3	14.5*	16.3	11.1	14.3

DE: Desviación estándard; IM: Imputación Múltiple; PMM: Predictive Mean Matching o Equiparación de media predictiva; AIC: Criterio Akaike.
*: Se *ha de tener en cuenta que este caso presenta una* ***n*** *distinta al resto por lo que el valor del AIC no es comparable al resto de modelos.*

En cuanto al valor del AIC, el que más se aproxima al original, es el obtenido mediante la IM (PMM), y también para la varianza residual. Por lo que se refiere a los estimadores *B* y sus *DE*, es la imputación por regresión lineal, la que proporciona en general estimadores *B* más cercanos a los originales, seguido en esta ocasión el modelo IM (PMM), que también proporciona *DE* más cercanas. En todos los casos los p-valores se mantienen significativos cuando lo son en el modelo original, excepto en algún caso del modelo *listwise*.

También se han analizado las correlaciones de los valores ajustados para para los modelos que contienen el mismo tamaño muestral respecto del análisis original (los resultados de las correlaciones se pueden hallar en el informe estadístico adjunto) y se representan gráficamente en la siguiente figura. Como se puede observar de nuevo los modelos resultantes de la imputación por regresión e IM presentan una correlación mayor (aunque hay mayor variabilidad en la IM).

Figura 34. Correlaciones valores ajustados diseños únicos (Modelo 3) (30%)



En el caso del nº de errores perseverantes el 1er modelo, que incluye Edad, Género y Nivel de educación, es el utilizado para hacer la comparación con las 4 simulaciones.

Tabla 11. Resumen comparativo resultados de los modelos: nº de errores perseverantes (30%)

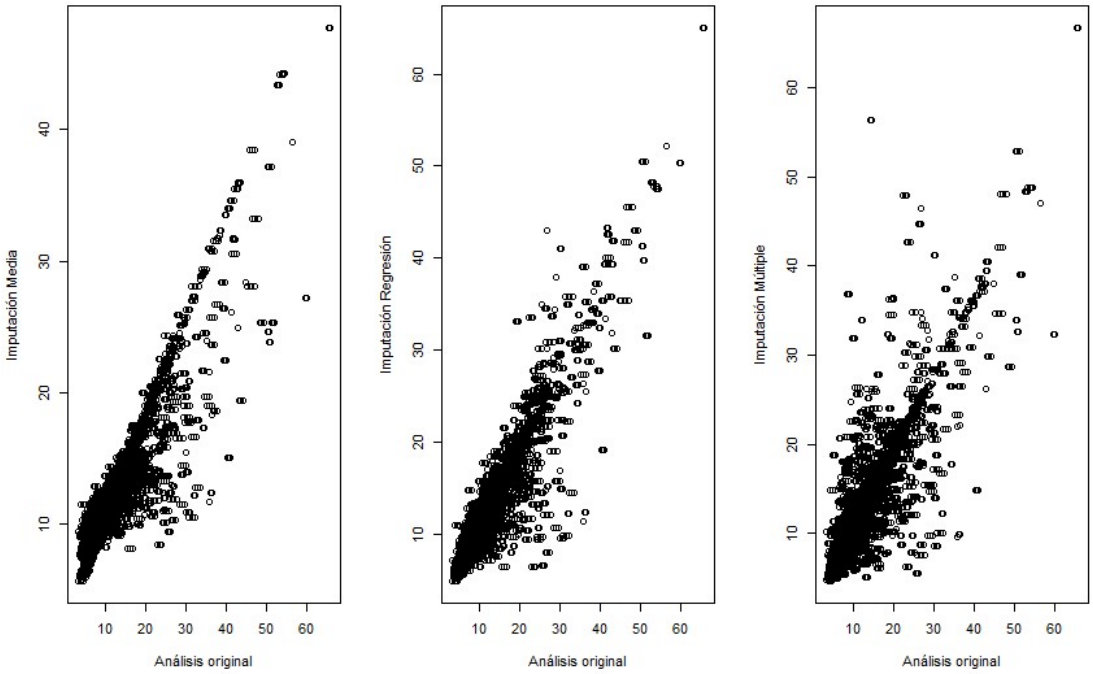
Tipo de análisis	Análisis original	Listwise	Media	Regresión	IM (5, PMM)
Modelo	**Modelo 1**	**Modelo 1**	**Modelo 1**	**Modelo 1**	**Modelo 1**
Variables	***B* (DE), IC95% (p-valor)**	***B* (DE), IC95% (p-valor)**	***B* (DE), IC95% (p-valor)**	***B* (DE), IC95% (p-valor)**	***B* (DE), IC95% (p-valor)**
Edad	-0.02 (0.023), -0.06 to 0.03 (0.491)	-0.06 (0.039), -0.13 to 0.02(0.153)	-0.013 (0.018), – to – (0.449)	-0.02 (0.022), -0.06 to 0.03 (0.421)	-0.01 (0.023), -0.05 to 0.04 (0.774)
Sexo: Hombre	Ref.	Ref.	Ref.	Ref.	Ref.
Sexo: Mujer	2.64 (0.451), 1.76 to 3.53 (<0.001)	2.13 (0.773), 0.61 to 3.65 (0.006)	1.755 (0.348), – to – (<0.001)	2.65 (0.44), 1.79 to 3.51 (<0.001)	2.82 (0.454), 1.93 to 3.71 (<0.001)
NE: Escuela primaria	Ref.	Ref.	Ref.	Ref.	Ref.
NE: Secundaria inicial	1.33 (0.974), -0.58 to 3.24 (0.171)	0.07 (1.649), -3.17 to 3.3 (0.969)	0.956 (0.752), – to – (0.204)	1.38 (0.952), -0.49 to 3.25 (0.147)	1.28 (0.982), -0.65 to 3.2 (0.193)
NE: Secundaria superior	-0.26 (0.992), -2.21 to 1.68 (0.792)	-1.8 (1.669), -5.08 to 1.47 (0.28)	-0.152 (0.766), – to – (0.843)	-0.33 (0.97), -2.23 to 1.57 (0.731)	-0.42 (1), -2.38 to 1.54 (0.675)
NE: Universitaria	-1.25 (0.97), -3.15 to 0.65 (0.196)	-3.06 (1.631), -6.26 to 0.14 (0.061)	0.749 (0.749), – to – (0.212)	-1.32 (0.948), -3.18 to 0.54 (0.164)	-1.42 (0.977), -3.33 to 0.5 (0.148)
AIC	60778.2	19935.6*	58464.5	57633.1	60576.1
Varianza residual	11.6	11.7*	10.5	10.3	11.9

DE: Desviación estándard; IM: Imputación Múltiple; PMM: Predictive Mean Matching o Equiparación de media predictiva; AIC: Criterio Akaike.
*: Se *ha de tener en cuenta que este caso presenta una **n** distinta al resto por lo que el valor del AIC no es comparable al resto de modelos.*

En cuanto al valor del AIC, el que más se aproxima al original, esta vez con diferencia, es el obtenido mediante la IM (PMM), aunque la varianza residual es más próxima la obtenida mediante el método listwise. Por lo que se refiere a los estimadores *B* y sus *DE*, el modelo IM (PMM) y el de regresión lineal son los que proporcionan *B* más ajustadas a los valores del análisis original con DE también muy parecidas y relativamente cercanas a las originales. En todos los casos los p-valores se mantienen significativos cuando lo son en el modelo original (solo para la variable género).

También se han analizado las correlaciones de los valores ajustados para para los modelos que contienen el mismo tamaño muestral respecto del análisis original (los resultados de las correlaciones se pueden hallar en el informe estadísticos adjunto) y se representan gráficamente en la siguiente figura. Como se puede observar el modelo resultante de la imputación por regresión muestra una correlación mayor, presentándose mayor variabilidad en la IM.

Figura 35. Correlaciones valores ajustados errores perseverantes (Modelo 1) (30%)



6.Conclusiones

Mediante este TFM se ha pretendido caracterizar los estudios con datos longitudinales y los problemas derivados de los análisis en los que se presentan datos faltantes. Precisamente, uno de los mayores focos de dificultad en el análisis de datos longitudinales es la presencia de dichos valores perdidos o datos faltantes (missing data). Ya hemos comentado con anterioridad que los valores perdidos pueden afectar desde algunas de las variables del estudio y algunos de los individuos de la muestra hasta la totalidad de los datos de algunos de los individuos en la muestra y los motivos por los que se producen son muy diversos, desde la falta de consentimiento por parte del paciente que permita al investigador acceder a un determinado resultado analítico, pasando por un problema técnico a la hora de analizar una determinada muestra o la no asistencia del paciente a una visita programada. Se han caracterizado los tipos de datos perdidos en función de la definición de Rubin en su artículo de 1976[12]: MCAR, MAR y MNAR.

En los últimos años la capacidad computacional ha crecido de manera exponencial y derivado de dicha situación se han ido desarrollado nuevas técnicas que permiten la aplicación de algoritmos más complejos, y de este modo se han desarrollado nuevos métodos de tratamiento de datos faltantes en el contexto del análisis de datos longitudinales. A pesar de estos desarrollos es bastante habitual encontrarse muchos trabajos publicados en los que no se presenta un análisis de los datos faltantes, y hay que hacer hincapié en que para proporcionar análisis más rigurosos, los estadísticos deben incorporar esta metodología de manera más frecuente.

Tras detallar los distintos métodos de tratamiento de datos perdidos adecuados a los análisis de datos longitudinales, y en la fase final del trabajo se ha presentado una ejemplificación basada en una base de datos biomédica que recoge datos en 3 mediciones temporales distintas del test RFFT. Como la base de datos original no presentaba datos perdidos, se generaron tres escenarios distintos de manera aleatoria, 10%, 20% y 30% de datos perdidos en cada medida de las variables ***número total de diseños únicos*** y ***número total de errores perseverantes***. Tras reproducir los análisis originales publicados y confirmar que los datos faltantes generados cumplen las premisas de MCAR, se llevó a cabo la comparación de los resultados originales con los resultados obtenidos mediante cuatro métodos de tratamiento de datos perdidos. A partir de los resultados que se han presentado anteriormente, en los 3 escenarios, el método de imputación múltiple (basada en el algoritmo PMM) de la función *"mice",* presenta resultados robustos y en general es el método que presenta las estimaciones más cercanas a los resultados originales. También cabe destacar que utilizando el método de regresión basado en el método *norm.predict* de la función *"mice",* la imputación también resultó eficaz en la mayoría de las ocasiones tanto para los resultados obtenidos con la variable número total de diseños únicos como con el número total de errores perseverantes. Por lo que es un método que también puede tenerse en cuenta cuando los datos faltantes cumplen las premisas MCAR. Utilizar métodos de eliminación cuando el mecanismo de datos perdidos MCAR, conlleva pérdida de eficacia aunque la muestra permanezca insesgada.

Son métodos que junto al de sustitución por la media deberían tratar de evitarse, en vista de la facilidad de aplicación y los resultados tan robustos que se obtienen con métodos como la regresión o la imputación múltiple.

El presente trabajo cuenta con la limitación de haberse circunscrito al caso de datos MCAR, por lo que las conclusiones no son extrapolables a todos los tipos de datos perdidos.

7.Glosario

AIC	Criterio Akaike
ANOVA	Análisis de varianza
DE	Desviación Estándar
EM	Expectación-maximización
FCS	Fully Conditional Specification
FIML	Full Information Maximum Likelihood
GEE	Ecuaciones de Estimación Generalizada
IQR	Rango intercuartílico
K-W	Kruskal Wallis
LOCF	Last Observation Carried Forward
MAR	Missing at random
MCAR	Missing completely at random
MCMC	Markov Chain Monte Carlo
MI o IM	Imputación múltiple
MICE	Multivariate Imputation via Chained Equations
MNAR	Missing not at random
PMM	Equiparación de media predictiva

8.Bibliografía

1.Montenegro-Montenegro E., Oh Y., Chesnut S. No le tema a los datos perdidos: enfoques modernos para el manejo de datos perdidos. Actualidades en Psicología, 29(119), 2015, 29-42. http://revistas.ucr.ac.cr/index.php/actualidades/article/view/18812

2.Gómez-García J., Palarea Albaladejo J., Martín-Fernández J.A. Métodos de inferencia estadística con datos faltantes. Estudio de simulación sobre los efectos en las estimaciones. Estadística Española. 48 (162), 2006, 241-70

3.Armitage P., Berry G., Matthews J.N.S. Statistical Methods in Medical Research. Blackwell Publishing. 4the Edition, 2005.

4.Delgado M., Llorca J. Estudios longitudinales: concepto y particularidades. Rev. Esp. Salud Pública, mar.-abr. 2004, vol.78, no.2, p.141-148. ISSN 1135-5727. http://www.monografias.com/trabajos902/estudios-longitudinales/estudios-longitudinales.shtml#ixzz4bCQlxJyU

5.Goldstein H. The design and analysis of longitudinal studies. Londres: Academic Press, 1979

6.Rosner B. The analysis of longitudinal data in epidemiologic studies. J Chron Dis 1979; 32: 163-73.

7.Pérez Andrés Cristina, Martín Moreno José María. Sobre los estudios longitudinales en epidemiología. Rev. Esp. Salud Publica [Internet]. 2004 Apr [cited 2017 Apr 03]; 78(2): 135-140. Available from: http://www.scielosp.org/scielo.php?script=sci_arttext&pid=S1135-57272004000200001&lng=en. http://dx.doi.org/10.1590/S1135-57272004000200001

8.Durán P. Los datos perdidos en estudios de investigación ¿son realmente datos perdidos? Arch. argent. pediatr. [Internet]. 2005 Dic [citado 2017 Mar 10]; 103(6): 566-568. Disponible en: http://www.scielo.org.ar/scielo.php?script=sci_arttext&pid=S0325-00752005000600015&lng=es.

9.Albert PS. (1999). Longitudinal data analysis (repeated measures) in clinical trials. Statistics in Medicine, 18, 1707-1732.

10. Bock RD. (1975). Multivariate statistical methods in behavioural research. New York: McGraw-Hill.

11.Arnau J, Bono R. Estudios longitudinales de medidas repetidas: Modelos de diseño y análisis. Escritos de Psicología [Internet]. 2008 Dic [citado 2017 Mar 10]; 2(1): 32-41. Disponible en: http://scielo.isciii.es/scielo.php?script=sci_arttext&pid=S1989-38092008000300005&lng=es.

12. Rubin DB. Inference and missing data. Biometrika, 1976, 63, 581-592.

13. Jeličić H, Phelps E, Lerner RM. Use of missing data methods in longitudinal studies: The persistence of bad practices in developmental psychology. Developmental Psychology, 2009, 45(4), 1195-199. http://dx.doi.org.sire.ub.edu/10.1037/a0015665

14. Samet, J. M., & Muñoz, A. (1998). Evolution of the cohort study. Epidemiologic reviews, 20(1), 1-14.

15. Nesselroade, J. R. y Baltes, P. B. (Eds.) (1979). Longitudinal research in the study of behaviour and development. New York: Academic Press.

16. Wu, Y. B., Clopper, R. y Wooldridge, P. J. (1999). A comparison of traditional approaches to hierarchical lineal modeling when analyzing longitudinal data. Research in Nursing & Healt, 22, 421-432.

17. Useche Castro L M, Mesa Ávila D M, Una introducción a la Imputación de Valores Perdidos. Terra Nueva Etapa 2006XXII127-151. Disponible en: http://www.redalyc.org/articulo.oa?id=72103106. Fecha de consulta: 1 de abril de 2017.

18. Tseng CH, Elashoff R, Li N, Li G. Longitudinal data analysis with non-ignorable missing data. Stat Methods Med Res. 2016 Feb; 25(1):205-20.

19. Longford NT. Missing Data and Small-Area Estimation (libro). Modern Analytical Equipment for the Survey Statistician. Sección: 4.6.3 Coarse data and rounding (p84). Springer Link. 2005.

20. Wilks S. (1932). Moments and distributions of estimates of population parameters from fragmentary simple. Annals of Mathematical Statistics, B, 163-195.

21. Graham JW. Missing data analysis: making it work in the real world. Annu Rev Psychol. 2009; 60:549-76.

22. Enders CK. Applied Missing Data Analysis. Guilford Press, New York, 2010.

23. Little RJA., Rubin DB. (2002). Statistical analysis with missing data (2nd ed.). Hoboken, NJ: Wiley.

24. Jöreskog KG, Sorbom D. (1993). PRELIS 2 user's reference guide [Computer software]. Chicago: Scientific software.

25. Dempster AP, Laird NM, Rubin DB. Maximum likelihood estimation from incomplete data via the EM algorithm (with discussion) 1977. Journal of the Royal Statistical Association 1977; B39: 1-38.

26. Ferro MA. Missing data in longitudinal studies: cross-sectional multiple imputation provides similar estimates to full-information maximum likelihood. Ann Epidemiol. 2014 Jan; 24(1):75-7.

27. Rubin DB. Multiple Imputation for Nonresponse in Surveys. New York: John Wiley and Sons; 1987.

28. Allison P. Imputation by Predictive Mean Matching: Promise & Peril | Statistical Horizons. Statisticalhorizons.com. (2015). [online] Available at: http://statisticalhorizons.com/predictive-mean-matching (último acceso 11-12-2017)

29. Vink, G., Frank, L. E., Pannekoek, J. and van Buuren, S. (2014), Predictive mean matching imputation of semicontinuous variables. Statistica Neerlandica, 68: 61–90. doi:10.1111/stan.12023 http://www.stefvanbuuren.nl/publications/2014%20Semicontinuous%20-%20Stat%20Neerl.pdf (último acceso 11-12-2017)

30. Roberts MB, Sullivan MC, Winchester SB. Examining solutions to missing data in longitudinal nursing research. J Spec Pediatr Nurs. 2017 Apr; 22(2).

31. Van Buure S, Brand JPL, Groothuis-Oudshoorn K, Rubin DB. Fully conditional specification in multivariate imputation. Journal of Statistical Computation and Simulation (2006) 76(12):1049-64. https://dspace.library.uu.nl/handle/1874/19951 (último acceso 28-12-2017)

32. Little TD, Jorgensen TD, Lang KM, Moore WEG. On the Joys of Missing Data. J Pediatr Psychol 2014; 39 (2): 151-162.

33. Bell ML, Fairclough DL. Practical and statistical isssues in missing data for longitudinal patient-reported outcomes. Statistical Methods in Medical Research, 2014; 23(5): 440–459.

34. Montenegro- Montenegro E, Oh Y, Chestnut S. No le tema a los datos perdidos: enfoques modernos para el manejo de datos perdidos. Actualidades en Psicología, [S.l.], v. 29, n. 119, p. 29-42, nov. 2015. https://revistas.ucr.ac.cr/index.php/actualidades/article/view/18812 (último acceso 28-12-2017)

35. Allison PD. Missing data techniques for structural equation modeling. J Abnorm Psychol 2003; 112(4):545-57.

36. Dong Y, Peng CYJ. Principled missing data methods for researchers. SpringerPlus 2013; 2(222), 1–17. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3701793/ (último acceso 28-12-2017)

37. Xiaoyue Cheng, Dianne Cook, Heike Hofmann (2015). Visually Exploring Missing Values in Multivariable Data Using a Graphical User Interface. Journal of Statistical Software, 68(6), 1-23.<doi:10.18637/jss.v068.i06> (MissingDataGUI)

38. Alexander Kowarik, Matthias Templ (2016). Imputation with the R Package VIM. Journal of Statistical Software, 74(7), 1-16. <doi:10.18637/jss.v074.i07> URL https://www.jstatsoft.org/article/view/v074i07. (último acceso 28-12-2017) (VIM y VIMGUI)

39. R in Action: Chapter 15 Advanced methods for missing data. http://rstudio-pubs-

static.s3.amazonaws.com/4625_fa990d611f024ea69e7e2b10dd228fe7.html (último acceso 28-12-2017)

40. Stef van Buuren, Karin Groothuis-Oudshoorn (2011). Mice: Multivariate Imputation by Chained Equations in R. Journal of Statistical Software, 45(3), 1-67. URL http://www.jstatsoft.org/v45/i03/. (último acceso 28-12-2017) (MICE)

41. James Honaker, Gary King, Matthew Blackwell (2011). Amelia II: A Program for Missing Data. Journal of Statistical Software, 45(7), 1-47. URL http://www.jstatsoft.org/v45/i07/. (último acceso 28-12-2017) (Amelia II)

42. Gelman A and Hill J (2011). "Opening Windows to the Black Box." Journal of Statistical Software, 40. (mi)

43. Mangiafico, S.S. 2016. Summary and Analysis of Extension Program Evaluation in R, version 1.6.19 rcompanion.org/handbook/. (Pdf versión: rcompanion.org/documents/RHandbookProgramEvaluation.pdf.) (último acceso 28-12-2017)

44. RStudio documentation. https://support.rstudio.com/hc/en-us/categories/200035113-Documentation?version=1.1.383&mode=desktop (último acceso 28-12-2017)

45. Knitr: Elegant, flexible, and fast dynamic report generation with R. https://yihui.name/knitr/ (último acceso 28-12-2017)

46. Introduction to R Markdown. http://rmarkdown.rstudio.com/articles_intro.html (último acceso 28-12-2017)

47. Van Eersel MEA, Joosten H, Koerts J, Gansevoort RT, Slaets JPJ, et al. (2015) Longitudinal Study of Performance on the Ruff Figural Fluency Test in Persons Aged 35 Years or Older. PLOS ONE 10(3): e0121411. https://doi.org/10.1371/journal.pone.0121411 (último acceso 28-12-2017)

48. Acosta, M. R., Avendaño, B. L., Martínez, M. y Romero, L. M. (2014). Análisis psicométrico del test de "fluidez de diseños de Ruff" en población universitaria de Bogotá. Acta Colombiana de Psicología, 17(1), 45-52. doi: 10.14718/ACP.2014.17.1.5 (http://editorial.ucatolica.edu.co/ojsucatolica/revistas_ucatolica/index.php/acta-colombiana-psicologia/article/view/16/html_4) (último acceso 28-12-2017)

49. Delgado-Mejía ID, Etchepareborda MC. https://cran.r-project.org/web/packages/BaylorEdPsych/BaylorEdPsych.pdf

50. Beaujean AA. R Package for Baylor University Educational Psychology. Package 'BaylorEdPsych'. February 2015. CRAN R Project. https://cran.r-project.org/web/packages/BaylorEdPsych/BaylorEdPsych.pdf(último acceso 28-12-2017)

51. Ohlsen N. Multiple Imputation in R. How to impute data with MICE for lavaan, January 2017. http://statistics.ohlsen-web.de/multiple-imputation-with-mice/ (último acceso 28-12-2017)

52. Van Buuren, S. (2012), Flexible Imputation of Missing Data. Chapman & Hall/CRC, Boca Raton, FL. ISBN 9781439868249. http://www.stefvanbuuren.nl/mi/FIMD.html (último acceso 28-12-2017)

53. Alice M. Imputing Missing Data with R; MICE package. October 2015 (Updated April 2017) https://datascienceplus.com/imputing-missing-data-with-r-mice-package/ (último acceso 28-12-2017)

54. Noghrehchi F. (May 2015). Missing Data Analysis with mice. http://web.maths.unsw.edu.au/~dwarton/missingDataLab.html (último acceso 28-12-2017)

55. Starkweather, J. Linear Mixed Effects Modeling using R. (2010). https://it.unt.edu/sites/default/files/linearmixedmodels_jds_dec2010.pdf (último acceso 28-12-2017)

56. Seoane J. Análisis bioestadístico con modelos de regresión en R. Universidad Autónoma de Madrid (2013). https://www.uam.es/personal_pdi/ciencias/jspinill/CFCUAM2013/ModelosMixtos-01-Anidados_CFCUAM2013.html (último acceso 28-12-2017)

57. Pinheiro J. et al. Linear and Nonlinear Mixed Effects Models. Package 'nlme' (Version 3.1-131, June 2017) https://cran.r-project.org/web/packages/nlme/nlme.pdf (último acceso 28-12-2017)