IMPORTANCIA DEL TRATAMIENTO DE DATOS PERDIDOS. APLICACIÓN EN ESTUDIOS LONGITUDINALES PEQUEÑOS

Pedro J. Mallol Roselló
Máster Bioinformática y Bioestadística
Departamento Estadística Universidad de Barcelona

Directora: Nuria Pérez Álvarez
Tutor: Alexandre Sánchez Pla

Entrega: 31 Mayo 2017

Esta obra está sujeta a una licencia de Reconocimiento-NoComercial-SinObraDerivada 3.0 España de Creative Commons

Título del trabajo:	Importancia del tratamiento de datos perdidos. aplicación en estudios longitudinales pequeños
Nombre del autor:	Pedro J. Mallol Roselló
Nombre del consultor/a:	Nuria Pérez Álvarez
Nombre del PRA:	Alexandre Sánchez Pla
Fecha de entrega (mm/aaaa):	Mayo 2017
Titulación:	Máster Bioinformática y Bioestadística
Área del Trabajo Final:	Strategies for dealing with Missing data in longitudinal data
Idioma del trabajo:	Español
Palabras clave	EQ5D, calidad vida, MICE, datos perdidos, datos longitudinales, MCAR, MAR, MNAR

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

Contexto: Pese a que se suele hacer un gran esfuerzo a la hora de la planificación y realización de las investigaciones clínicas, la pérdida de datos es un fenómeno que siempre va a estar presente y está demostrado que una pérdida de datos puede conllevar unos resultados erróneos. Aunque en las grandes investigaciones este punto suele estar controlado para así evitar una pérdida innecesaria de tiempo y dinero, en las que se realizan más a pequeña escala no se suele contemplar. Creemos pues conveniente realizar una pequeña demostración de la importancia que tienen los datos perdidos en nuestras investigaciones y las posibles consecuencias que pueden darse al no tenerlos en cuenta.

Métodos: Partiremos de una base de datos sencilla y bastante limitada tanto en tamaño como en variabilidad de datos. Para ello nos basaremos en los cuestionarios de calidad de vida EQ5D, en los cuales hemos simulado una perdida según diferentes mecanismos (MCAR, MAR, MNAR) y en diferentes porcentajes (10%, 20%, 30%).

Resultados: Cuanto mayor porcentaje de pérdida, mayor desviación respecto a la realidad presentarán nuestros resultados. Para pérdidas de datos MCAR y MAR será mejor realizar la técnica de *Imputación Múltiple*, en cambio para MNAR es mejor optar por *Pairwise*

Conclusiones: Es muy importante conocer nuestra Base de Datos. Saber qué porcentaje de pérdida tenemos y qué tipo de pérdida está afectando a nuestros datos, así como conocer los diferentes métodos que existen para tratar los datos perdidos y poder aplicar el más idóneo a nuestro caso particular.

Abstract (in English, 250 words or less):

Background: Although a great effort is usually made when planning and conducting clinical research, missing data will always be present and it is demonstrated that this may bias the results obtained. This situation is usually under control in large and well structured clinical research in order to avoid loss of time and money, but in a small scale, this situation is not usually contemplated. For this reason, a demonstration of the importance of dealing with missing data is mandatory.

Methods: A small and limited dataset is created in order to simulate a small clinical research. We perform missing data according to different types of loss mechanisms (MCAR, MAR, MNAR) and in different percentages of missing (10%, 20%, 30%).

Results: The greater percentage of loss, the greater deviation from reality will present our results. For MCAR and MAR mechanisms, *Multiple imputation* must be the method, whereas for MNAR it is better to choose Pairwise.

Conclusions: It is very important to understand our dataset. To know what percentage of loss we have and which mechanism of missing data we have, as well as to know the different methods that exist to deal with the missing data and to be able to apply the most appropriate to our particular research.

Índice

Resumen

Pese a que se suele hacer un gran esfuerzo a la hora de la planificación y realización de las investigaciones clínicas, la pérdida de datos es un fenómeno que siempre va a estar presente y está demostrado que una pérdida de datos puede conllevar unos resultados erróneos.

En las grandes investigaciones que tienen una gran financiación detrás, este punto suele estar controlado para así evitar una pérdida innecesaria de tiempo y dinero, pero aún así, nos sorprende encontrarnos algunos estudios relevantes donde no se ha tenido en cuenta los datos faltantes. Si hablamos de los estudios a nivel local de algunos centros de investigación, este problema se agrava aún más.

Los datos faltantes son uno de los puntos que se deberían de tener en cuenta en cualquier investigación. Creemos pues conveniente realizar una pequeña demostración de la importancia que tienen los datos perdidos en nuestras investigaciones y las posibles consecuencias que pueden darse si no se tratan adecuadamente.

Para ello, nos pondremos en una situación muy simple pero muy común que afecta a los estudios con datos longitudinales. Partiremos de una base de datos sencilla y bastante limitada tanto en tamaño como en variabilidad. Para ello nos basaremos en los cuestionarios de calidad de vida EQ5D, en los cuales hemos simulado una pérdida según diferentes mecanismos (MCAR, MAR, MNAR) y en diferentes porcentajes (10%, 20%, 30%).

Veremos que cuanto mayor porcentaje de pérdida tenemos, mayor desviación respecto a la realidad presentarán nuestros resultados. Para pérdidas de datos MCAR y MAR será mejor realizar la técnica de *Imputación Múltiple*, en cambio para MNAR es mejor optar por *Pairwise.*

Concluiremos que es muy importante conocer nuestra Base de Datos. Saber qué porcentaje de pérdida tenemos y qué tipo de pérdida está afectando a nuestros datos, así como conocer los diferentes métodos que existen para tratar los datos perdidos y poder aplicar el más idóneo a nuestro caso particular.

1. Introducción

La pérdida de datos en investigación es siempre una realidad y un problema a considerar. Es inevitable en todo estudio de investigación existan datos de los que no podamos disponer por debidos diversos motivos.

En el campo de los ensayos clínicos también vemos ésta problemática con más asiduidad de lo que nos gustaría.

En una revisión de lo publicado sobre el tema entre Julio y Diciembre de 2001 (Wood, White et al. 2004) [1] en revistas médicas de alto impacto, se detectan 63 ensayos con presencia de datos faltantes de un total de 71, para un 89%. En una revisión del 2014 de ensayos aleatorizados por grupos (Díaz-Ordaz, Kenward et al. 2014) [2], se detectan 95 de un total de 132, para un 72%; sin embargo aquí se destacan 41 ensayos (31%) en los cuales "no pudo verificarse la presencia de datos faltantes a partir del reporte publicado". Buuren (2012) critica la pobreza de los reportes de datos faltantes en las publicaciones. Ejemplifica, mediante una colección de pequeñas bases de datos (Hand, Daly et al. 1994) [3], la baja calidad de estos reportes: de las 510 bases de la colección, sólo 13 contenían un código para los datos faltantes.

En cualquier tipo de análisis estadístico se desea hacer inferencias válidas sobre una población de interés. La presencia de información faltante en una matriz de datos lleva consigo ciertos inconvenientes, dentro de los cuales, según Horton & Lipsitz (2001) [5], se encuentran: pérdida de eficiencia, complicaciones en el análisis de datos faltantes, y además estimadores sesgados que ponen en riesgo la validez del proceso.

Además, en muchos ensayos clínicos a pequeña escala que suelen hacer los investigadores en sus propios centros, no se suele tener en cuenta este tipo de situación. Bien por desconocimiento de la misma o porque sencillamente no le dan la importancia suficiente a su investigación. Solo en el caso de una posible publicación existen algunos investigadores que mandan a revisar su trabajo por un profesional.

En algunos diseños de experimentos es más común la presencia de información faltante que en otros, como es el caso de los relacionados con medidas repetidas; estos diseños en el campo de la experimentación son frecuentes, especialmente en las ciencias médicas, biológicas y agronómicas; no obstante, presentan ciertas dificultades con su manejo, debido no solo a la ocurrencia de observaciones faltantes, sino a la característica de dependencia entre las observaciones repetidas hechas sobre la misma unidad experimental, pues este tipo de diseños son raramente balanceados y completos. Estas dificultades llevan consigo complicaciones en el modelamiento, pérdida de precisión en estimaciones, y además la inferencia se ve afectada ya que se pueden presentar funciones no estimables.

Los datos longitudinales en los que cada sujeto o unidad experimental se mide u observa en ocasiones múltiples es probable que posean muchos datos perdidos debido a muy diversos motivos, como a la fatiga del informante, desconocimiento de la información solicitada, al rechazo de las personas a informar a cerca de temas sensibles, a la negativa de participar en una investigación, incluso a problemas asociados a la calidad del marco del muestreo. Además, también hemos de considerar la posibilidad de la existencia de observaciones "aberrantes" o poco probables.

Para evitar esa pérdida de datos, podríamos actuar en tres momentos diferentes del estudio:

1. Durante el diseño del experimento. Teniendo en cuenta el porcentaje de datos que podemos "perder" y aumentando el tamaño muestral para poder llegar al mínimo requerido.

2. Durante la etapa de recolección de datos mediante un adecuado monitoreo de la calidad de los datos.

3. Una vez recolectados los datos. Este momento corresponde al proceso de análisis de consistencia de los datos y valores perdidos, que debe formar parte de todo estudio, antes de la estimación de los estadísticos que permita dar respuesta a los objetivos definidos.

Referente al último apartado, existen varias formas de resolver esta problemática. Soluciones que pasan por "obviar" esos datos, o estimarlos en función de varios métodos (imputación).

Durante la década de los setenta, la imputación de datos significaba identificar y sustituir los registros sin información. En ese contexto, los procedimientos *hot-deck*, en sus distintas variantes, se aplicaban para suplir información en censos y encuestas, y los métodos de *ajuste por promedios* y *cold-deck* eran frecuentemente utilizados.

En Rubin (1976), se propuso un marco conceptual para el análisis de datos faltantes sustentado en métodos de inferencia estadística. Posteriormente, la aparición del algoritmo Expectation Maximization (EM) permitió generar estimadores robustos a partir de la aplicación del método de máxima verosimilitud (MV) (Dempster, Laird, y Rubin, 1977) [6], en donde las observaciones faltantes se asumen como variables aleatorias y los datos imputados se generan sin necesidad de ajustar modelos.

Más tarde, Rubin (1987), introdujo el concepto de imputación múltiple (IM), sustentado en la premisa de que cada dato faltante debe ser reemplazado a partir de m>1 simulaciones. La aplicación de esta técnica se facilitó con los avances computacionales y el desarrollo de los métodos bayesianos de simulación que aparecieron hacia finales de los ochenta (Schafer, 1997) [7]. A partir de ese momento, se propició el uso intensivo de estos algoritmos, debido a que distintas rutinas de de IM y MV se incluyeron en paquetes comerciales y de acceso gratuito.

Durante el decenio de los noventa se registraron avances notables. El método de reponderación, históricamente utilizado por los estadísticos de encuestas, se modificó con el propósito de utilizarlo en la imputación de datos en modelos de regresión (Ibrahim, 1990) [8]. También se desarrollaron técnicas en el campo de la bioestadística, con el propósito de resolver situaciones en las que los datos faltantes dependían de las características de la población.

El objetivo era disponer de algoritmos apropiados para analizar datos clínicos, ya que en este tipo de investigaciones es común que el número de pacientes se reduzca, y los estudios concluyan con menos personas de las que iniciaron (Little, 1995) [9].

En Little (1992) [10] y Little y Rubin (1987) [11], los métodos de imputación se clasifican como se muestra a continuación:

- Análisis de datos completos (*Listwise*)
- Análisis de datos disponibles (*Pairwise*)
- Imputación por medias no condicionadas
- Imputación por medias condicionadas mediante métodos de regresión
- Máxima verosimilitud (*MV*)
- Imputación múltiple (*IM*)

La literatura considera que el desarrollo del método IM ha zanjado el debate respecto a la mejor forma de imputar datos omitidos. Sin embargo, los estadísticos de encuestas han demostrado que las soluciones propuestas no tienen en cuenta el diseño de la muestra, ni las probabilidades de selección de las unidades de análisis, cuando los datos provienen de muestras complejas, situación que introduce sesgos en las estimaciones.

1.1.- Clasificación de los mecanismos de pérdidas de datos

Es importante determinar el patrón y mecanismo de pérdida de datos para poder escoger un método de imputación apropiado. Rubin (1976) [12] clasificó los mecanismos de pérdida de datos en tres tipos diferentes:

- Completamente al azar (**MCAR:** missing completely at random). Representa una situación para la cual la ausencia es independiente de las variables en su investigación. Es una ausencia de datos debida exclusivamente al azar.

  Dicho de otra manera, la probabilidad de que un sujeto presente un valor ausente en una variable no depende ni de otras variables del cuestionario ni de los valores de la propia variable con valores perdidos. Las observaciones con datos perdidos son una muestra aleatoria del conjunto de observaciones.

  Por ejemplo, un tubo que contiene una muestra de sangre de un individuo es roto por accidente o en el contexto de los calidades de vida, cuando la pérdida de la información no tiene nada que ver con el propio cuestionario de calidad de vida. Es decir, se pierde un calidad de vida o se olvida la paciente de rellenar cierto apartado.

- Perdidos al azar (**MAR**: missing at random). Se refiere cuando la ausencia de datos está ligada a las variables independientes del estudio, pero no a la variable dependiente.

  Cuando los sujetos con datos incompletos son diferentes significativamente de los que presentan datos completos en alguna variable, y el patrón de ausencia de datos puede ser predecible a partir de variables con datos observados en la base de datos del estudio que no muestran ausencia de datos.

  Por ejemplo, si se hace un test de aptitud a unos alumnos y a los que superan una nota de corte establecida se les hace otro más difícil mientras que a los demás no. Por tanto éstos tienen datos perdidos para la segunda variable y se debe a las observaciones de la primera.

- Un proceso que no es ni MCAR ni MAR se denomina no aleatorio (**MNAR**: missing non at random). Estos casos se dan cuando la pérdida de datos se debe a la variable dependiente, y posiblemente a alguna variable independiente.

  Por ejemplo, un caso MNAR es cuando en un cuestionario le preguntas a alguien por su renta anual y éste no contesta porque es muy alta.

Otra forma mas técnica de entender esta clasificación sería identificando a los registros sin información con una variable binaria (Z). Se asume el valor “1” si el dato existe y “0” en caso contrario. Rubin nos dice que la ausencia de datos debe analizarse como un fenómeno estocástico, por lo que Z debe considerarse una variable aleatoria con distribución de probabilidad conjunta, la cual da cuenta del porcentaje de omisión existente y de su relación con las observaciones completas.

Sea $Y_{com} = (Y_{obs}, Y_{mis})$ la variable de interés, donde Y_{obs} corresponde al vector de datos observados e Y_{mis} al vector de datos faltantes. Se considera que un proceso de datos omitidos se genera **MAR** si la distribución de los valores observados no depende del patrón de comportamiento de los registros sin información, es decir, depende de los valores observados.

$$Y_{mis} : P(Z|Y_{com}) = P(Z|Y_{obs})$$

En el caso de **MCAR**, ocurre cuando la omisión no depende de los datos observados, es decir, depende de los datos no observados.

$$P(R|Y_{com}) = P(Z)$$

MCAR y **MAR** se suelen clasificar como *ignorables*, mientras que **MNAR** se consideran como *no ignorables.*

1.2.- Clasificación de los patrones de pérdidas de datos

Además de esta clasificación según su mecanismo de pérdida, se pueden observar varios patrones de pérdida de datos. Ambas clasificaciones son muy útiles a la hora de saber qué mecanismo de imputación poder aplicar.

Los patrones se pueden clasificar en:

- Monótona (terminal): Los datos se pierden cuando un paciente no realiza más calidades de vida en un cierto momento del estudio

- Intermitente: La pérdida de datos suceden entre las diferentes evaluaciones del paciente.

- Mixto: Este patrón sucede cuando a un proceso de pérdida intermitente le sigue uno de pérdida monótona.

1.3.- Métodos de determinación de mecanismo de pérdida

En la actualidad existen muchos métodos para determinar el mecanismo de pérdida de datos. Métodos basados en test de hipótesis. Shona Fielding (2009) [13] nos ofrece un trabajo con la descripción de los métodos mas comúnmente usados y una comparación entre ellos.

Little (1986) [14] desarrolló un test basado en medias. Listing y Schlittgen (1998) [15-16] desarrollaron también un test basado en medias y además un procedimiento no-paramétrico que combina varios test de Wilcoxon. Schmitz y Franz (2002) [17] desarrollaron un versión no-paramétrica del primer test de los anteriores autores. Diggle (1998) [18] utilizó una aproximación para determinar si las pérdidas son una muestra aleatoria de la población total. Ridout (1991) [19] adoptó un enfoque parecido a Diggle, pero utilizando una regresión logística. Fairclough (2002) [20] describió un procedimiento muy parecido al método de Ridout.

1.4.- Principales métodos de imputación

Procedimientos tradicionales

- **Análisis de datos completos (***Listwise o case deletion LD***):** Esta análisis consiste en trabajar solo con la información completa para todas las variables. Es la práctica más utilizada, aunque se reconoce que no es la más apropiada porque genera sesgos en los coeficientes de asociación y de correlación (Kalton y Kaspryk, 1982) [21].

- **Análisis con los datos disponibles (***Pairwise deletion*): Esta posibilidad aboga por trabajar con toda la información disponible (*Available-case(AC)*). En este caso obtendremos distintos tamaños muestrales lo que limita la comparación de resultados. Este método también asume un patrón MCAR en los datos omitidos.

- **Reponderación**: De esta forma se intenta reducir el sesgo cuando no se satisface es supuesto MCAR en los datos omitidos. Es posible aplicar diferentes métodos para reponderar las observaciones. Los ponderadores (w_i) se interpretan como el número de unidades de la población que representa a cada elemento en la muestra, y es común que los algoritmos de reponderación se apliquen para compensar la falta de respuesta en subgrupos de interés.

  Cuando en una subclase se detecta ausencia de información, los ponderadores de las unidades que sí respondieron se utilizan para ajustar los factores de expansión, de tal forma que la submuestra observada genere estimaciones compatibles con los valores poblaciones de la subclase de interés. Para este proceso se pueden utilizar tanto datos propios de la misma muestra para ponderar como datos exógenos proveniente de otras bases de datos.

Imputación Simple

- **Imputación por media incondicional:** La imputación por media incondicional es una estrategia que consiste en calcular la media muestral para cada una de las variables que tiene datos faltantes, y luego utilizar este valor para sustituir todos los valores faltantes que tiene la variable correspondiente.

  Ejemplos donde se puede ver claramente la ineficiencia de este método son estudios donde queremos ver el salario de una familia. Si falta un 30% de los datos, estaremos obteniendo un 30% de registros con un valor fijo (la media) y varianza 0, lo cual subestimará la dispersión de la muestra total, además de incidir en el coeficiente de correlación con otras variables. El valor medio de la variable se mantiene, pero otros estadísticos que definen la forma de la distribución (varianza, covarianza, quantiles, sesgo, kurtosis…) se pueden ver afectados (Schafer y Graham 2002) [22]. A pesar de todos los inconvenientes señalados, es común el uso de esta metodología ya que se piensa que el promedio de los datos es un valor representativo de los valores que faltan.

- **Imputación por media condicional**: Una variante del modelo anterior que consiste en formar categorías a partir de covariables correlacionadas con la variable de interés, e imputar los datos faltantes con observaciones provenientes de la submuestra que comparte características comunes (Acock y Demo, 1994) [23].

  Al igual que la imputación de medias incondicionales, se asume que la falta de datos presenta un patrón MCAR y en este caso, existirán tantos promedios como categorías se formen, lo cual ayuda a disminuir los sesgos, pero en ningún caso los elimina.

- **Imputación con variables ficticias**: Este método consiste en crear una variable indicador para identificar las observaciones con datos omitidos (Cohen y Cohen 1983 y Cohen et al 2003) [24-25]. Se estima un modelo de regresión $y = \alpha + \beta Z$, donde suponemos que la variable predictora (y) es la condición de respuesta ("1" si tenemos respuesta y "0" si no la tenemos), entonces, a las personas con datos faltantes se les asigna la media de esta variable predictora.

  Debido a las confusiones a las que puede inducir este método, no se recomienda esta técnica de imputación.

- **Imputación mediante una distribución no condicionada (***Hot-deck)*: Este método tiene como objetivo llenar los registros vacíos (receptores) con información de campos con información completa (donantes), y los datos faltantes se reemplazan a partir de una selección aleatoria de valores observados, lo cual no introduce sesgos en la varianza del estimador (Madow, Nilssen y Olkin, 1983) [26].

  Este método ha sido utilizado desde hace mucho tiempo por organismos tan importantes como la oficina del censo de los EEUU. El algoritmo consiste en ubicar registros completos e incompletos, identificar características comunes de donantes y receptores, y decidir los valores que se utilizarán para imputar los datos omitidos. Para la aplicación del procedimiento es fundamental generar agrupaciones que garanticen que la imputación se llevará a cabo entre observaciones con características comunes, y la selección de los donantes se realiza en forma aleatoria evitando que se introduzcan sesgos en el estimador de la varianza.

  Existen variantes del procedimiento *hot-deck.* El "*algoritmo secuencial*", parte de un proceso de ordenación de los datos en cada subgrupo y selecciona donantes en la medida que recorre el archivo de datos. Su aplicación supone que la falta de respuesta se distribuye en forma aleatoria en cada una de las categorías, pero en caso de que la falta de respuesta se concentre en un estrato con pocas observaciones, es posible que se generen estimadores sesgados en la medida que el procedimiento seleccione varias veces el mismo donante.

  Por su parte, el "*método aleatorio*" identifica registros sin datos y elige en forma estocástica al donante. También existe la posibilidad de que el donante sea el "vecino más cercano" al registro sin datos, y la selección se efectúa a partir de la definición de criterios de distancia.

  El *hot-deck* y las variantes que se han comentado se consideran mejores opciones que los procedimientos *listwise deletion*, *pairwise deletion*, y es superior a los métodos de medias condicionadas y no condicionadas, ya que no introduce sesgos en el estimador y su error estándar. Si se desea preservar la distribución de probabilidad de las variables imputadas, conforme a la opinión de algunos autores se considera que el procedimiento *hot-deck* es más eficiente que el algoritmo la imputación múltiple y la regresión paramétrica (Durrant, 2005) [27].

- **Imputación por regresión**: Como su nombre indica, la imputación por regresión reemplaza valores faltantes con respuestas predichas de un modelo de regresión. En un análisis multivariante, los casos completos son utilizados para estimar un modelo de regresión donde la variable incompleta es la respuesta y las variables explicativas son algunas de las variables completas. El modelo de regresión estimado permite estimar respuestas predichas para los casos incompletos.

- **Imputación por regresión estocástica**: La imputación por regresión estocástica también utiliza ecuaciones de regresión para predecir las variables incompletas a partir de las variables completas, pero requiere un paso adicional que consiste en aumentar cada predicción con un término residual distribuido mediante la distribución del error, generalmente una distribución normal. Añadir residuos a los valores imputados reestablece la pérdida de variabilidad de los datos y efectivamente elimina el sesgo asociado con los esquemas de imputación de regresión estándar. Con este método de imputación obtenemos estimaciones insesgadas de los parámetros bajo datos MAR.

- **Imputación por máxima verosimilitud**: Los métodos de máxima verosimilitud se pueden aplicar en cualquier problema de estimación. En el análisis de datos omitidos, y asumiendo que los datos faltantes siguen un patrón MAR, se demuestra que la distribución marginal de los registros observados está asociada a una función de verosimilitud para un parámetro desconocido, bajo el supuesto de que el modelo es adecuado para el conjunto de datos completo.

  De acuerdo a Little y Rubin (1987), a esta función se le conoce como la función de verosimilitud, la cual ignora el mecanismo que generó los datos faltantes (verosimilitud de los datos observados conforme a Shafer y Graham (2002) [22]. El procedimiento para estimar los parámetros de un modelo utilizando una muestra con datos faltantes se resume a continuación:

  i. Estimar los parámetros del modelo con los datos completos con la función de máxima verosimilitud.
  ii. Utilizar los parámetros estimados para predecir los valores omitidos.
  iii. Sustituir los datos por las predicciones, y obtener nuevos valores de los parámetro maximizando la verosimilitud de la muestra completa.

  Un procedimiento eficiente para maximar la verosimilitud cuando existen datos faltantes es el algoritmo Expectarion-Maximization (Dempster, Laird y Rubin (1977) [28]

Imputación Múltiple

El método de imputación múltiple consiste en realizar varias imputaciones de las observaciones faltantes para luego analizar los conjuntos de datos completados y combinar los resultados obtenidos para obtener una estimación final. El análisis de imputación múltiple está dividido en tres fases: Fase de imputación, fase de análisis y fase de puesta en común.

- La *fase de imputación* crea múltiples copias de los conjuntos de datos (*m*), y cada una de ellas contiene diferentes estimaciones de los valores perdidos. Conceptualmente, este paso es una versión iterativa de la imputación por regresión estocástica, aunque sus fundamentos matemáticos se basan en muchas ocasiones en los principios de estimación bayesiana.

- El objetivo de la *fase de análisis*, como su nombre indica, es analizar los conjuntos de datos rellenados. Este paso aplica los mismos procedimientos estadísticos que un individuo hubiera utilizado si tuviera todos los datos. La única diferencia es que realizamos cada análisis *m* veces, una para cada conjunto de datos imputados.

- La *fase de análisis* nos lleva a *m* conjuntos de estimaciones de parámetros y errores estándar, con lo que el propósito de la fase de puesta en común es combinar todo en un conjunto simple de resultados. Rubin (1987) perfiló fórmulas relativamente sencillas para poner en común las estimaciones de los parámetros y los errores estándar. Por ejemplo, la estimación del parámetro puesto en común es simplemente la media aritmética de las *m* estimaciones de la fase de análisis. Combinar los errores estándar es ligeramente más complejo pero sigue la misma lógica. El proceso de analizar conjuntos de datos múltiples y poner en común los resultados parece laborioso, pero los paquetes de software de imputación múltiple automatizan completamente el procedimiento. Las *m* estimaciones son combinadas en una estimación en conjunto y una matriz de varianzas-covarianzas utilizando las reglas de Rubin, que están basadas en la teoría asintótica en un marco bayesiano. La matriz de varianzas-covarianzas combinada incorpora la variabilidad dentro de la imputación (incertidumbre sobre los resultados de unos conjuntos de datos imputados) y la variabilidad entre las imputaciones (relejando la incertidumbre debido a la información perdida).

1.5 Calidad de vida eq-5d[1]

EQ-5D es una medida estandarizada de la situación de la salud desarrollado por el "EUROQOL Group" con la finalidad de aportar una medida de evaluación simple y genérica.

Nos aporta de una manera fácil y sencilla una herramienta que nos permite evaluar la calidad de vida de los pacientes. Además de ser una herramienta estandarizada que nos otorga la posibilidad de reproducirlo, dándole así un gran valor en cualquier ensayo/estudio clínico.

El cuestionario EQ-5D está desarrollado para que el encuestado lo pueda rellenar por si mismo, lo que le confiere una gran versatilidad a la hora de administrarlo. Puede realizarse por medios tan variados como el correo postal, email e incluso en la sala de espera de las consultas. Además es un cuestionario breve que no lleva más de unos pocos minutos completarlo.

Existen diferentes tipos de cuestionarios en función de lo que se quiera analizar: EQ-5D-3L, EQ-5D-5L y EQ-5D-Y. Son cuestionarios bastante similares y las características descritas anteriormente es común a todos ellos. Para nuestro trabajo, nos vamos a centrar en el primero (EQ-5D-3L)

[1] http://www.euroqol.org/

EQ-5D-3L (ANEXO 1) consiste en dos páginas. Una página descriptiva y otra con una escala visual analógica (VAS).

- La *escala visual analógica* consta de una regla vertical que va desde el 0 (abajo) hasta el 100 (arriba). Donde el 0 se refiere al peor estado de salud posible y el 100 se refiere al mejor estado de salud.

La forma de determinar el estado de salud en este caso es más sencillo. Basta con apuntar el valor que va de 0-100 que marque el encuestado. Para valores perdidos se usa el 999.

- El sistema descriptivo consta de:

5 dimensiones (5D):

  - 1.- Movilidad
  - 2.- Cuidado Personal
  - 3.- Actividades de todos los días
  - 4.- Dolor/Malestar
  - 5.- Ansiedad/Depresión.

Cada dimensión tiene 3 niveles (3L):

  - 1.- Sin problemas
  - 2.- Algunos problemas
  - 3.- Verdaderos problemas.

Para determinar el estado de salud se combinan todas las posibilidades que ofrece el sistema descriptivo (5 dimensiones y 3 niveles), teniendo un total de 243 estados de salud posibles.

Nos referimos a cada estado de salud como un código de 5 dígitos que se obtiene de las respuestas dadas por el encuestado. Así pues, una encuesta que no tiene ningún problema en las cinco dimensiones se le asigna un valor de “11111”, mientras que estado de salud “11223” nos indica que no tiene ningún problema en la movilidad y en el cuidado personal, algunos problemas en las actividades diarias y en el dolor/malestar y verdaderos problemas con las ansiedad/depresión. Los valores perdidos se codifican como 9.

Estos estados de salud se pueden resumir en un único valor para poder trabajar de una forma mas cómoda. Este valor denominado “índice” se obtiene aplicando una fórmula que básicamente añade ciertos valores (también denominados “pesos”) a los diferentes niveles de las diferentes dimensiones.

Estos “pesos” se han obtenido exclusivamente para cada país. Es importante remarcar que cada país tiene unos valores específicos que no han de ser los mismos para otros países. Así pues podemos encontrarnos el caso de que para un país en concreto se le da mas importancia a la hora de tener problemas en la movilidad que en otro. Por tanto tendrá un valor “índice” más negativo en ese país un perfil “31111” que en otro (suponiendo que el resto de dimensiones se valoren igual).

Obtenemos de esta manera un valor índice que va desde 1 (mejor estado de salud posible) a 0 (peor estado de salud posible).

2. Objetivos del Trabajo

Como hemos visto hasta ahora, es muy común tener datos faltantes en nuestras investigaciones y está demostrado que esta pérdida de información afecta a los resultados. Un punto preocupante es ver como en la mayoría de las investigaciones no tienen en cuenta este hecho , obteniendo así resultados que podrían ser cuestionables.

El tratamiento de los datos faltantes debería ser tan importante como el planteamiento inicial de la investigación y el posterior control de la misma. Intentar evitar la pérdida de datos debería de ser prioritario para evitar sesgos importantes a la hora de obtener resultados.

Como hemos dicho anteriormente, las investigaciones importantes (como pueden ser los Ensayos Clínicos) suelen tener este punto en consideración, pero las investigaciones más pequeñas no suelen tener en cuenta este problema. Investigaciones humildes que pueden tener su importancia, puesto que pueden dar pie a ideas que pueden acabar desarrollando una gran investigación. Un ejemplo de ésto son los estudios que se realizan a nivel local en los centros de investigación.

Existe mucha literatura acerca de cómo tratar los datos perdidos en grandes investigaciones y con grandes bases de datos, pero no hemos encontrado muchos casos donde se hayan analizado estudios pequeños. Éstos, presentan unas limitaciones intrínsecas que los hacen un poco más especiales. Por ello, queremos realizar nuestro trabajo sobre un estudio pequeño.

Así pues, queremos concienciar a los investigadores sobre la importancia de los datos que perdemos durante nuestro estudio. Además, queremos proponer diferentes alternativas a realizar en el caso de que tengamos una pequeña y limitada base de datos inicial con la que pretendamos iniciar nuestra investigación.

2.1 Enfoque y método seguido

La forma más sencilla y evidente de ver la importancia de tratar los datos perdidos en cualquier investigación es partir de una base de datos completa e ir generando pérdidas. Posteriormente se vuelven a analizar los datos y se comparan los resultados obtenidos para ver su variación.

En nuestro caso vamos a generar una base de datos ficticia y vamos a generar pérdidas de datos en función de los diferentes mecanismos de pérdida (MCAR, MAR MNAR) y en función de diferentes porcentaje de pérdidas (10%, 20%, 30%).

De esta forma, a parte de ver el efecto de la pérdida de datos en nuestros resultados, podemos establecer a partir de qué porcentaje dicha pérdida empieza a ser significativa y si existe algún tipo de relación con los diferentes tipos de pérdidas.

Partiremos de una base de datos pequeña y básica como ejemplo de una investigación sencilla que se puede llevar a cabo en cualquier centro de investigación.

2.2 Planificación del Trabajo

Debido al campo tan amplio al que puede aplicar este trabajo, nos vamos a centrar en una de las áreas que mas se ve afectada por este proceso de pérdida de datos. Estamos hablando de los estudios con calidad de vida.

Un buen ensayo clínico presenta un diseño inicial validado y un posterior control exhaustivo con sus constantes monitorizaciones y revisiones pertinentes, pero aún así, parece ser que existen ciertos aspectos de ellos que no se toman tan en cuenta como otros objetivos a priori mas importantes como la supervivencia, tiempo a la recaída, etc.

Un ejemplo de estos procedimientos son los calidades de vida. Muchas veces no se le da tanta importancia a rellenar una encuesta sobre el estado de salud, como al echo de tener que recibir un tratamiento. Y esta falta de importancia viene tanto del investigador como del propio paciente. Es raro que un paciente se deje de tomar una dosis de su tratamiento, porque entiende que es lo que está mejorando su enfermedad, en cambio, no es tan raro que un paciente no rellene un calidad de vida porque no lo ve prioritario a la hora de su curación.

En la práctica habitual se ven muchos casos donde el mismo paciente es quien se niega a rellenar el calidad de vida, se le olvida cumplimentarlo, lo rellena al azar o incluso es el propio familiar del paciente quien contesta a las preguntas. Además, el hecho de tener que hacer el mismo cuestionario varias veces a lo largo del tratamiento lo hace más tedioso. Todo ésto, más la poca importancia que le suelen dar los pacientes, hace que sea uno de los procedimientos que más datos faltantes acumula.

Por ello vamos a utilizar una base de datos basándonos en los calidades de vida. Por ser uno de los ítems mas castigados por este problema. En concreto nos centraremos en la calidad de vida EQ5D que se suele usar en la mayoría de los ensayos clínicos de oncología.

Partiremos de una población donde TODOS los datos estarán recogidos y a partir de ella, obtendremos diferentes subgrupos donde existen pérdidas de datos. Se definirán previamente qué tipos de pérdidas de datos vamos a realizar y también, en qué proporción.

Posteriormente utilizaremos alternativas que existen para abordar el problema. Utilizaremos una de las técnicas más comunes y sencillas que existe y que consiste en eliminar los casos con datos faltantes *(Listwise),* trabajar con los datos que tenemos (*Pairwise)* y una técnica de imputación que se considera de las más correctas para utilizar en esta problemática, la *Imputación Múltiple.*

El trabajo se va a realizar mediante el programa estadístico R excepto la simulación de los resultados de las encuestas y la obtención de la variable "Índice", que se ha realizado mediante Excel. Ésto se debe a que para calcular la variable en cuestión se realiza mediante una fórmula que proporciona EUROQOL group en formato Excel[2].

La planificación del trabajo será tal y como se detalla en el ANEXO 2

[2] Se adjuntará junto con el trabajo dicha herramienta ("*EQ-5D-5L Crosswalk Index Value Calculator MAC.xls*").

2.3 Creación de la Base de datos[3]

Como es de esperar, no podemos hacer uso de los datos generados en un ensayo clínico debido a todas las clausulas de privacidad que poseen. Realizar una petición de uso, conllevaría una carga de trabajo extra, que para la finalidad de nuestro TFM no es necesaria. Por todo ello, no he utilizado las variables del ensayo clínico, sino que me he centrado en unas variables específicas que nos interesan, partiendo de una base de datos en blanco.

Las características principales de nuestra base de datos son:

- Tamaño de 25 pacientes que se analizan en 2 momentos de la investigación (50 cuestionarios en total).

- Variables ordinales, con 3 niveles cada una.

- Son datos longitudinales que se obtienen del mismo paciente en diferentes momentos. Lo que implica a la hora de realizar un test de significancia, que que son datos pareados.

Generamos estocásticamente mediante la función RANDOM() de Excel las respuestas de los cuestionarios EQ5D. Dichas respuestas han sido introducidas en la herramienta de Excel específica proporcionada por EUROQOL para obtener así las variables "Perfil" e "Índice". Guardamos la simulación en un archivo que se denomina "*Respuestas.xlxs".*

Una vez creada la base de datos inicial, la exportamos a R. La importación crea unas variables que no están bien clasificadas, por ello se crea una función que me ayudará a re-clasificarlas ('*reclass').* De esta forma, obtenemos una estructura correcta para trabajar.

Para agilizar el proceso realizado hasta ahora, he unificado todos los pasos anteriores en una única función. Llamamos a la función '*DBprepare (BD)*' y el argumento deberá de ser el nombre de la Base de datos entre comillas (para convertirlo en 'string'). La hoja del Excel donde se encuentran los datos, por defecto será la primera.

Para el posterior apartado de creación de pérdidas, vamos a necesitar una variable nueva para poder utilizarla a la hora de simular una pérdida ***MAR***. Por ello creamos la variable "*estudios*", que nos indica si el paciente tiene estudios superiores o no. Un valor *0* indica que no tiene y *1* nos está diciendo que el paciente tiene estudios superiores.

[3] Todo el código de R utilizado en este trabajo se adjuntará en dos archivo denominados "TFM.Rmd" y "TFM.html".

2.4 Simulación de pérdidas

Nuestra idea es comparar los diferentes tipos de pérdida de datos según propuso Rubin en 1987 así como diferentes grados. Para ello vamos a establecer previamente el porcentaje de pérdida de datos que vamos a tener:

50 cuestionarios x 5 variables = 250 ítems

- 10 % de pérdida supone 25 ítems menos
- 20 % de pérdida supone 50 ítems menos
- 30 % de pérdida supone 75 ítems menos

Mediante R, generamos un código que genere la simulación de pérdidas, obteniendo así las diferentes bases de datos con las pérdidas correspondientes

Para la generación de una pérdida de datos **MCAR** hemos simulado al azar las pérdidas de todos los 250 ítems que tenemos en nuestro estudio. Obtenemos así:

- MCAR_EQ5D10: Base de datos con pérdidas MCAR al 10%
- MCAR_EQ5D20: Base de datos con pérdidas MCAR al 20%
- MCAR_EQ5D30: Base de datos con pérdidas MCAR al 30%

Para la generación de una pérdida de datos **MAR**, hemos supuesto que los pacientes que no presentan estudios superiores no rellenan el cuestionario de calidad de vida correctamente y se dejan ítems sin rellenar. En cambio, los pacientes que presentan estudios superiores han rellenado los calidades de vida completamente y sin pérdidas. Obtenemos así las siguientes bases de datos:

- MAR_EQ5D10: Base de datos con pérdidas MAR al 10%
- MAR_EQ5D20: Base de datos con pérdidas MAR al 20%
- MAR_EQ5D30: Base de datos con pérdidas MAR al 30%

Por último, para la generación de una pérdida de datos **MNAR**, simulamos que los pacientes que menor índice tienen (peor se encuentran), no rellenan correctamente los calidades de vida. Obtenemos las siguientes bases de datos:

- MNAR_EQ5D10: Base de datos con pérdidas MNAR al 10%
- MNAR_EQ5D20: Base de datos con pérdidas MNAR al 20%
- MNAR_EQ5D30: Base de datos con pérdidas MNAR al 30%

En el ANEXO 3 podemos encontrar un esquema de los diferentes patrones de pérdida de datos para poder verlo de una forma mas visual

2.5 Evaluación calidad vida

Para evaluar la calidad de vida de los pacientes durante el tratamiento, vamos a analizarlo de dos formas. Una primera manera descriptiva, donde vamos a ver variación de porcentajes entre los pacientes que no tienen problemas frente a los que sí presentan algún tipo de molestia. También lo veremos de manera más global donde compararemos la evolución de la variable "índice" como variable que aglutina en un único valor todas las características que se consideran importantes a la hora de estimar la calidad de vida de un paciente.

2.5.1.- Comparación descriptiva

Para realizar una comparación descriptiva, vamos a reclasificar previamente los niveles del calidad de vida en dos grupos. Un primer nivel que denominaremos "*Sin problemas"* y otro denominado "*Con problemas"*. El nuevo nivel "*Sin problemas"* engloba las respuestas que tuvieron como valor "1", y el nivel "*Con problemas"* engloba las respuestas que fueron "2" o "3".

De esta sencilla forma podemos ver si los pacientes pasan de un estado con ciertos problemas iniciales a un estado en el que se acaban resolviendo. Analizaremos esta fase descriptiva para las cinco variables que recoge nuestro calidad de vida.

Para ello creamos una función en R denominada '*Descriptiva (BD)'* que nos calculará el porcentaje de pacientes que no presentan problemas en los diferentes aspectos que estudia el calidad de vida (Movilidad, Cuidado Personal, Actividades diarias, Dolor, Ansiedad)

2.5.2.- Comparación variable "Índice"

Gracias a la variable *"Índice"*, podemos realizar una comparación global de los calidades de vida. Como ya se ha comentado anteriormente, existe un algoritmo dependiente de cada país, que nos identifica el calidad de vida en un número que varía entre 0.5 y 1. A más calidad de vida del paciente, este valor se acercará mas al 1.

Así pues, vamos a comparar el valor índice promedio de la situación basal frente al valor índice promedio de la situación final. Posteriormente evaluaremos si las diferencias son clínicamente significativas al 95% de confianza mediante un test de significancia.

Hay que tener en cuenta que al utilizar una variable cuyo rango va de 0.5 a 1, dicha variable va a estar truncada por arriba y por abajo, por lo que no podremos tener una distribución normal diga lo que diga nuestro test de normalidad. Ahora bien, si tuviéramos muchas observaciones entonces si sería probable que se pueda aproximar por una normal, como pasa con las proporciones, en cuyo caso, el error de aproximación sería asumible.

Debido a esta disyuntiva, para valorar si ha habido diferencias significativas entre los diferentes *timepoints*, utilizaremos un test paramétrico (t-Student) y otro no paramétrico (Wilcoxon) y veremos si coinciden los resultados.

Creamos una función en R denominada '*Medias (BD)*' que nos proporcionará la media y las desviaciones estándar de manera global, en los cuestionarios basales y en los finales, y otra función denominada '*ComparaMedias(BD)*', que nos aplicará los dos test de significancia estadística al 5% (t-Student y Wilcoxon) sobre la media basal y final de la variable Índice.

Un asunto importante a tener en cuenta en este punto es que debido a las características de nuestra investigación, necesitamos realizar un test estadístico con datos pareados. Necesitamos que nuestro paciente tenga cuestionario de calidad de vida basal y final para poder hacer las comparaciones. Por ello será necesario eliminar de esta comparativa a los pacientes que no presenten los dos calidades de vida.

2.6 Tratamiento de los datos perdidos

Debido a las características de nuestra base de datos, vamos a trabajar con los métodos de *Listwise*, *Pairwise* e *Imputación Múltiple.*

La forma más fácil e intuitiva de actuar a la hora de manejar los datos perdidos consiste en obviarlos. Es decir, realizamos nuestras estadísticas simplemente no contando con los datos que faltan. En esta premisa se basan técnicas como *Listwise (análisis de datos completos)* o *Pairwise* (análisis de datos disponibles).

Mediante técnicas de datos disponibles tenemos una pérdida de muestra menor que en el análisis de datos completos, pero obtenemos subgrupos con distintos tamaños, lo que después nos puede limitar a la hora de ciertos análisis. Por ello, en nuestro caso, como procedimiento tradicional de imputación también utilizaremos el método de análisis de datos completos.

Finalmente, y para comparar los resultados con un método más actual y más completo, utilizaremos la *Imputación Múltiple* para imputar los datos faltantes y así poder seguir trabajando con toda la muestra.

2.6.1.- Técnicas Tradicionales

2.6.1.1- Listwise/Pairwise en la fase descriptiva

Para realizar este método vamos a analizar las bases de datos con la función '*Descriptiva' creada para tal fin*. Partimos de la base de datos que no presenta ninguna perdida y seguimos con las bases de datos con las pérdidas. De este modo, estaremos simulando la técnica *Pairwise*, pues estaremos analizando solo los datos disponibles.

Para realizar la técnica de *Listwise* crearemos una función denominada '*listwiseDB(DB)'* que nos aplicará la función '*Descriptiva(DB)*' solo en los casos completos de nuestra base de datos. Aplicamos dicha función para obtener los resultados pertinentes.

2.6.1.2- Listwise en la variable "Índice"

Mediante la función '*Medias (BD)*' obtenemos las medias y desviaciones estándar de la variable Índice de las diferentes bases de datos, y gracias al test '*ComparaMedias (BD)*' obtendremos la significancia clínica.

Antes de aplicar la función '*ComparaMedias (BD)'* prepararemos las bases de datos para trabajar solo con los pacientes que presentan los dos cuestionarios de calidad de vida. Crearemos así las siguientes mases de datos nuevas:

- MCAR_EQ5D10comp: Base de datos compensada con pérdidas MCAR al 10%
- MCAR_EQ5D20comp: Base de datos compensada con pérdidas MCAR al 20%
- MCAR_EQ5D30comp: Base de datos compensada con pérdidas MCAR al 30%
- MAR_EQ5D10comp: Base de datos compensada con pérdidas MAR al 10%
- MAR_EQ5D20comp: Base de datos compensada con pérdidas MAR al 20%
- MAR_EQ5D30comp: Base de datos compensada con pérdidas MAR al 30%
- MNAR_EQ5D10comp: Base de datos compensada con pérdidas MNAR al 10%
- MNAR_EQ5D20comp: Base de datos compensada con pérdidas MNAR al 20%
- MNAR_EQ5D30comp: Base de datos compensada con pérdidas MNAR al 30%

2.6.2.- Imputación de datos (Imputación múltiple)

La técnica de Imputación Múltiple la realizaremos mediante la librería MICE [29] de R. Este paquete realiza imputación múltiple utilizando Fully Conditionally Specification (FCS) implementado por el algoritmo MICE (Multiple Imputtion by Chained Equations).

El algoritmo MICE requiere una especificación de un método de imputación univariante separadamente para cada variante incompleta. Para nuestro trabajo, realizaremos la imputación mediante el método *"polr"*, que imputa valores de dos o más niveles ordenados por el modelo de probabilidades proporcionales (Proportional odds model (ordered, >=2 levels)).

Nos centramos solo en las variables que contiene el calidad de vida (Movilidad, Cuidado Personal, Actividades Diarias, Dolor y Ansiedad) y usamos todas ellas como predictoras. Una vez realizadas las imputaciones deberemos exportarlas a Excel para poder utilizar la herramienta proporcionada por EUROQOL ("*EQ-5D-5L Crosswalk Index Value Calculator MAC.xls*"). Una vez creadas las variables "perfil" e "Índice" las volvemos a importar a R[4].

[4] Podremos ver un resumen de la imputación en el archivo adjunto "*Resumen Imputación.xlsx"*

Obtenemos así las siguientes bases de datos:

- MCAR_EQ5D10_i: Base de datos imputada con pérdidas MCAR al 10%
- MCAR_EQ5D20_i: Base de datos imputada con pérdidas MCAR al 20%
- MCAR_EQ5D30_i: Base de datos imputada con pérdidas MCAR al 30%
- MAR_EQ5D10_i: Base de datos imputada con pérdidas MAR al 10%
- MAR_EQ5D20_i: Base de datos imputada con pérdidas MAR al 20%
- MAR_EQ5D30_i: Base de datos imputada con pérdidas MAR al 30%
- MNAR_EQ5D10_i: Base de datos imputada con pérdidas MNAR al 10%
- MNAR_EQ5D20_i: Base de datos imputada con pérdidas MNAR al 20%
- MNAR_EQ5D30_i: Base de datos imputada con pérdidas MNAR al 30%

- MCAR_EQ5D10_new: Base de datos imputada tras calcular las variables *perfil e índice* con la herramienta del EUROQOL*, con* pérdidas MCAR al 10%
- MCAR_EQ5D20_new: Base de datos imputada tras calcular las variables *perfil e índice* con la herramienta del EUROQOL*,* con pérdidas MCAR al 20%
- MCAR_EQ5D30_new: Base de datos imputada tras calcular las variables *perfil e índice* con la herramienta del EUROQOL*,* con pérdidas MCAR al 30%
- MAR_EQ5D10_new: Base de datos imputada tras calcular las variables *perfil e índice* con la herramienta del EUROQOL*,* con pérdidas MAR al 10%
- MAR_EQ5D20_new: Base de datos imputada tras calcular las variables *perfil e índice* con la herramienta del EUROQOL*,* con pérdidas MAR al 20%
- MAR_EQ5D30_new: Base de datos imputada tras calcular las variables *perfil e índice* con la herramienta del EUROQOL*,* con pérdidas MAR al 30%
- MNAR_EQ5D10_new: Base de datos imputada tras calcular las variables *perfil e índice* con la herramienta del EUROQOL*,* con pérdidas MNAR al 10%
- MNAR_EQ5D20_new: Base de datos imputada tras calcular las variables *perfil e índice* con la herramienta del EUROQOL*,* con pérdidas MNAR al 20%
- MNAR_EQ5D30_new: Base de datos imputada tras calcular las variables *perfil e índice* con la herramienta del EUROQOL*,* con pérdidas MNAR al 30%

2.6.2.1- Imputación en la fase descriptiva

Este apartado consistirá en aplicar la función creada previamente "Descriptiva(BD)" a las nuevas bases de datos imputadas.

2.6.2.2- Imputación en la comparación de la variable "Índice"

Este apartado consistirá en aplicar la función creada previamente "Medias(BD)" y ComparaMedias(BD)" a las nuevas bases de datos imputadas.

En la Tabla 1, 2 y 3 podemos ver un resumen de los resultados obtenidos

Listwise								
Mecanismo	Media global	n (Media global)	Media basal	n (Media basal)	Media final	n (Media final)	p_valor (t-test)	p_valor (Wilcoxon)
EQ5D	0.713 (0.075)	50	0.699 (0.071)	25	0.726 (0.079)	25	0.2159	0.2583
MCAR_EQ5D10	0.715 (0.074)	31	0.699 (0.081)	14	0.727 (0.067)	17		
MCAR_EQ5D20	0.724 (0.073)	18	0.706 (0.082)	8	0.738 (0.067)	10		
MCAR_EQ5D30	0.733 (0.078)	13	0.722 (0.089)	6	0.743 (0.072)	7		
MAR_EQ5D10	0.703 (0.078)	34	0.679 (0.064)	17	0.728 (0.084)	17		
MAR_EQ5D20	0.708 (0.079)	29	0.689 (0.063)	14	0.725 (0.089)	15		
MAR_EQ5D30	0.7 (0.068)	28	0.689 (0.063)	14	0.711 (0.074)	14		
MNAR_EQ5D10	0.744 (0.077)	33	0.733 (0.072)	14	0.753 (0.069)	19		
MNAR_EQ5D20	0.758 (0.062)	29	0.76 (0.052)	11	0.756 (0.069)	18		
MNAR_EQ5D30	0.765 (0.051)	28	0.76 (0.052)	11	0.768 (0.051)	17		

Eliminamos los pacientes que se han quedado sin pareja								
Mecanismo	Media global	n (Media global)	Media basal	n (Media basal)	Media final	n (Media final)	p_valor (t-test)	p_valor (Wilcoxon)
EQ5D	0.713 (0.075)	50	0.699 (0.071)	25	0.726 (0.079)	25	0.2159	0.2583
MCAR_EQ5D10	0.701 (0.074)	18	0.689 (0.071)	9	0.712 (0.08)	9	0.5471	0.6523
MCAR_EQ5D20	0.727 (0.06)	6	0.68 (0.044)	3	0.773 (0.023)	3	0.0572	0.25
MCAR_EQ5D30	0.735 (0.035)	2	0.71 (NA)	1	0.76 (NA)	1	NA	NA
MAR_EQ5D10	0.707 (0.077)	30	0.689 (0.061)	15	0.725 (0.089)	15	0.2328	0.2931
MAR_EQ5D20	0.7 (0.068)	28	0.689 (0.063)	14	0.711 (0.074)	14	0.4336	0.4895
MAR_EQ5D20	0.7 (0.068)	28	0.689 (0.063)	14	0.711 (0.074)	14	0.4336	0.4895
MNAR_EQ5D10	0.746 (0.054)	20	0.747 (0.075)	10	0.746 (0.025)	10	0.9734	0.9188
MNAR_EQ5D20	0.752 (0.044)	18	0.763 (0.058)	9	0.741 (0.02)	9	0.3504	0.4738
MNAR_EQ5D20	0.752 (0.044)	18	0.763 (0.058)	9	0.741 (0.02)	9	0.3504	0.4738

IM								
Mecanismo	Media global	n (Media global)	Media basal	n (Media basal)	Media final	n (Media final)	p_valor (t-test)	p_valor (Wilcoxon)
EQ5D	0.713 (0.075)	50	0.699 (0.071)	25	0.726 (0.079)	25	0.2159	0.2583
MCAR_EQ5D10	0.713 (0.08)	50	0.696 (0.078)	25	0.731 (0.079)	25	0.1437	0.1412
MCAR_EQ5D20	0.714 (0.072)	50	0.692 (0.066)	25	0.735 (0.073)	25	0.0746	0.08505
MCAR_EQ5D30	0.721 (0.077)	50	0.708 (0.079)	25	0.734 (0.075)	25	0.3125	0.4261
MAR_EQ5D10	0.720 (0.081)	50	0.698 (0.077)	25	0.741 (0.08)	25	0.04121	0.06263
MAR_EQ5D20	0.709 (0.075)	50	0.702 (0.066)	25	0.717 (0.085)	25	0.4516	0.4418
MAR_EQ5D30	0.712 (0.066)	50	0.715 (0.07)	25	0.710 (0.063)	25	0.7796	0.6528
MNAR_EQ5D10	0.732 (0.072)	50	0.716 (0.075)	25	0.749 (0.066)	25	0.03326	0.008214
MNAR_EQ5D20	0.729 (0.068)	50	0.725 (0.065)	25	0.732 (0.073)	25	0.5473	0.2651
MNAR_EQ5D30	0.752 (0.055)	50	0.754 (0.058)	25	0.750 (0.053)	25	0.7484	0.9772

Tabla 1.- Resumen resultados de la variable índice

	Porcentaje de pacientes sin problemas Basal									
	BASAL	MCAR_EQ5D10	MCAR_EQ5D20	MCAR_EQ5D30	MAR_EQ5D10	MAR_EQ5D20	MAR_EQ5D30	MNAR_EQ5D10	MNAR_EQ5D20	MNAR_EQ5D30
Movilidad	real	32	32	32	32	32	32	32	32	32
	Pairwaise	35	37	33	35	33	29	35	40	50
	Listwise	21	25	33	24	29	29	50	64	64
	Imputación Múltiple	36	32	36	36	32	40	36	32	52
Cuidado personal	real	20	20	20	20	20	20	20	20	20
	Pairwaise	17	20	22	21	24	26	24	24	27
	Listwise	21	12	0	24	21	21	21	27	27
	Imputación Múltiple	16	20	28	20	24	24	24	36	28
Actividades diarias	real	48	48	48	48	48	48	48	48	48
	Pairwaise	48	47	47	43	48	48	48	48	44
	Listwise	57	62	67	47	50	50	50	45	45
	Imputación Múltiple	44	48	44	44	48	44	48	16	40
Dolor	real	44	44	44	44	44	44	44	44	44
	Pairwaise	43	45	50	39	45	39	50	50	47
	Listwise	43	62	83	29	29	29	50	55	55
	Imputación Múltiple	44	44	44	40	40	44	48	20	64
Ansiedad	real	36	36	36	36	36	36	36	36	36
	Pairwaise	33	37	39	41	45	44	43	40	46
	Listwise	43	38	50	35	43	43	50	55	55
	Imputación Múltiple	36	28	36	40	40	48	44	36	56

Tabla 2.- Resumen resultados de la parte descriptiva Basal

	Porcentaje de pacientes sin problemas Final									
	FINAL	MCAR_EQ5D10	MCAR_EQ5D20	MCAR_EQ5D30	MAR_EQ5D10	MAR_EQ5D20	MAR_EQ5D30	MNAR_EQ5D10	MNAR_EQ5D20	MNAR_EQ5D30
Movilidad	real	44	44	44	44	44	44	44	44	44
	Pairwaise	48	45	53	43	45	53	44	45	48
	Listwise	47	50	43	47	53	50	47	50	53
	Imputación Múltiple	44	40	48	48	40	44	44	24	48
Cuidado personal	real	44	44	44	44	44	44	44	44	44
	Pairwaise	45	53	56	43	37	27	50	48	50
	Listwise	41	40	43	35	27	40	58	56	59
	Imputación Múltiple	48	56	48	48	32	56	56	24	52
Actividades diarias	real	16	16	16	16	16	16	16	16	16
	Pairwaise	16	18	14	18	20	21	17	14	15
	Listwise	24	30	29	18	20	21	16	17	18
	Imputación Múltiple	16	24	16	16	24	24	24	40	20
Dolor	real	44	44	44	44	44	44	44	44	44
	Pairwaise	45	44	36	48	45	56	45	45	50
	Listwise	41	40	43	53	53	50	53	56	59
	Imputación Múltiple	48	40	40	48	48	48	48	24	52
Ansiedad	real	40	40	40	40	40	40	40	40	40
	Pairwaise	36	36	39	43	42	33	39	43	42
	Listwise	29	30	43	41	40	36	42	44	47
	Imputación Múltiple	36	40	44	44	40	44	44	24	44

Tabla 3.- Resumen resultados de la parte descriptiva Final

3. Evaluación de Resultados[5]

3.1- Fase Descriptiva

Recordamos que la parte descriptiva consiste en ver el porcentaje de pacientes que no tienen ningún tipo de problema. Para analizar los resultados, hemos calculado las unidades de variación de manera absoluta entre el dato real y los datos obtenidos mediante los diferentes técnicas de dos maneras diferentes:

- Primero hemos estudiado la variación independiente del mecanismo de pérdida, es decir, solo fijándonos en el porcentaje de pérdida.
- Segundo, hemos analizado la variación en función del mecanismo de pérdida.

3.1.1- Independiente del mecanismo de pérdida

En este caso nos vamos a centrar en el porcentaje de pérdida y el método utilizado para imputar nuestros datos. La lógica en este tipo de comparación es que cuando mayor porcentaje de pérdida, mayor variación encontraremos, y eso es justamente lo que hemos encontrado.

En la Tabla 4 podemos ver que de forma general, la variación global, independiente del método usado, en una base de datos con un 10% de pérdida es de 4 unidades, frente a 6 unidades en el caso de un 20% y hasta 7 unidades en el caso de haber perdido un 30%.

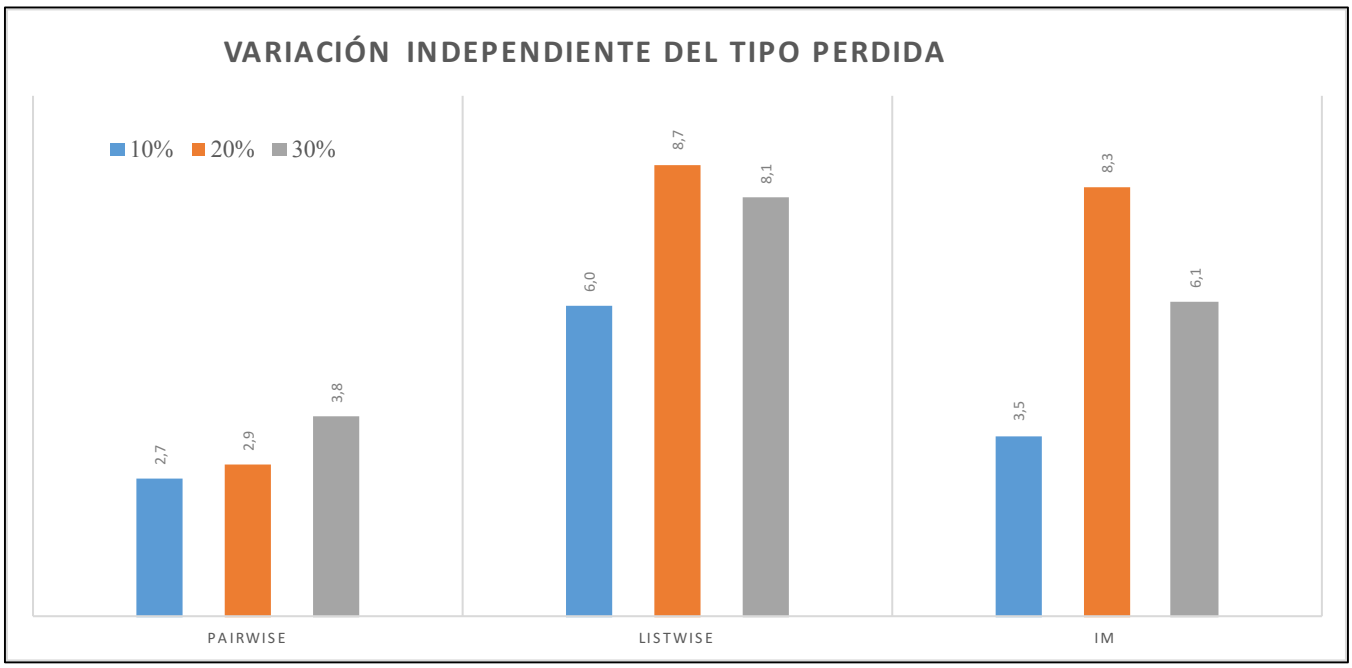
Centrándonos ahora en el mecanismo de pérdida y sin tener el cuenta el porcentaje, podemos observar que el método que más variación tiene es el de *Listwise* con 8.1 unidades, frente a 3.8 unidades del método *Pairwise* y 6.1 del método de *Imputación Múltiple.*

En la gráfica 1, podemos ver de manera más clara como se comporta la variación en conjunto. Podemos ver que el mejor método para este tipo de imputación es el *Pairwise*, que ofrece unas variaciones mucho menores que el resto de técnicas. A su vez, podemos ver que a mayor porcentaje de perdidas, mayor variación de resultados. Como dato curioso resaltar que la *Imputación Múltiple* y en Listwise, a partir del 20% de perdida empieza a ofrecer unos resultados mas dispares.

[5] Se adjunta archivo "Interpretación.Rmd" para ver el código utilizado en el cálculo

Item	Metodo	10%	20%	30%	Variación			
Movilidad	Pairwise	2,4	2,9	7,3	4,2			
Movilidad	Listwise	7,7	10,5	8,7	8,9			
Movilidad	IM	2,7	4,7	6,7	4,7			
Cuidado personal	Pairwise	2,7	4,7	8,4	5,3	Pairwise	Listwise	IM
Cuidado personal	Listwise	5,4	8,2	8,0	7,2	4,2	8,9	4,7
Cuidado personal	IM	4,7	10,7	7,4	7,6	5,3	7,2	7,6
Actividades diarias	Pairwise	1,4	1,5	2,2	1,7	1,7	5,8	6,5
Actividades diarias	Listwise	3,7	6,3	7,4	5,8	3,8	11,1	6,2
Actividades diarias	IM	2,7	12,0	4,7	6,5	4,3	7,3	5,3
Dolor	Pairwise	3,0	1,7	6,7	3,8	**3,8**	**8,1**	**6,1**
Dolor	Listwise	7,2	11,5	14,5	11,1			
Dolor	IM	3,4	9,3	6,0	6,2			
Ansiedad	Pairwise	3,9	3,9	5,2	4,3			
Ansiedad	Listwise	6,0	7,0	9,0	7,3			
Ansiedad	IM	4,0	4,7	7,4	5,3			
		4	**6**	**7**				
10%			20%			30%		
Pairwise	Listwise	IM	Pairwise	Listwise	IM	Pairwise	Listwise	IM
2,4	7,7	2,7	2,9	10,5	4,7	4,2	8,9	4,7
2,7	5,4	4,7	4,7	8,2	10,7	5,3	7,2	7,6
1,4	3,7	2,7	1,5	6,3	12,0	1,7	5,8	6,5
3,0	7,2	3,4	1,7	11,5	9,3	3,8	11,1	6,2
3,9	6,0	4,0	3,9	7,0	4,7	4,3	7,3	5,3
2,7	**6,0**	**3,5**	**2,9**	**8,7**	**8,3**	**3,8**	**8,1**	**6,1**

Tabla 4.- Variación sin tener en cuenta el tipo de pérdida



Gráfica 1.- Variación sin tener en cuenta el tipo de pérdida

3.1.2- Independiente del porcentaje de pérdida

En este apartado vamos a tener en cuenta el mecanismo de pérdida que estamos teniendo en cada caso a la hora de trabajar nuestros datos. En la Tabla 5 podemos ver el resumen de los resultados obtenidos

Podemos observar como el mecanismo de pérdida que más variación presenta es MNAR, 6.8 unidades, frente a las 4.4 unidades de MCAR y las 5.2 unidades de MAR.

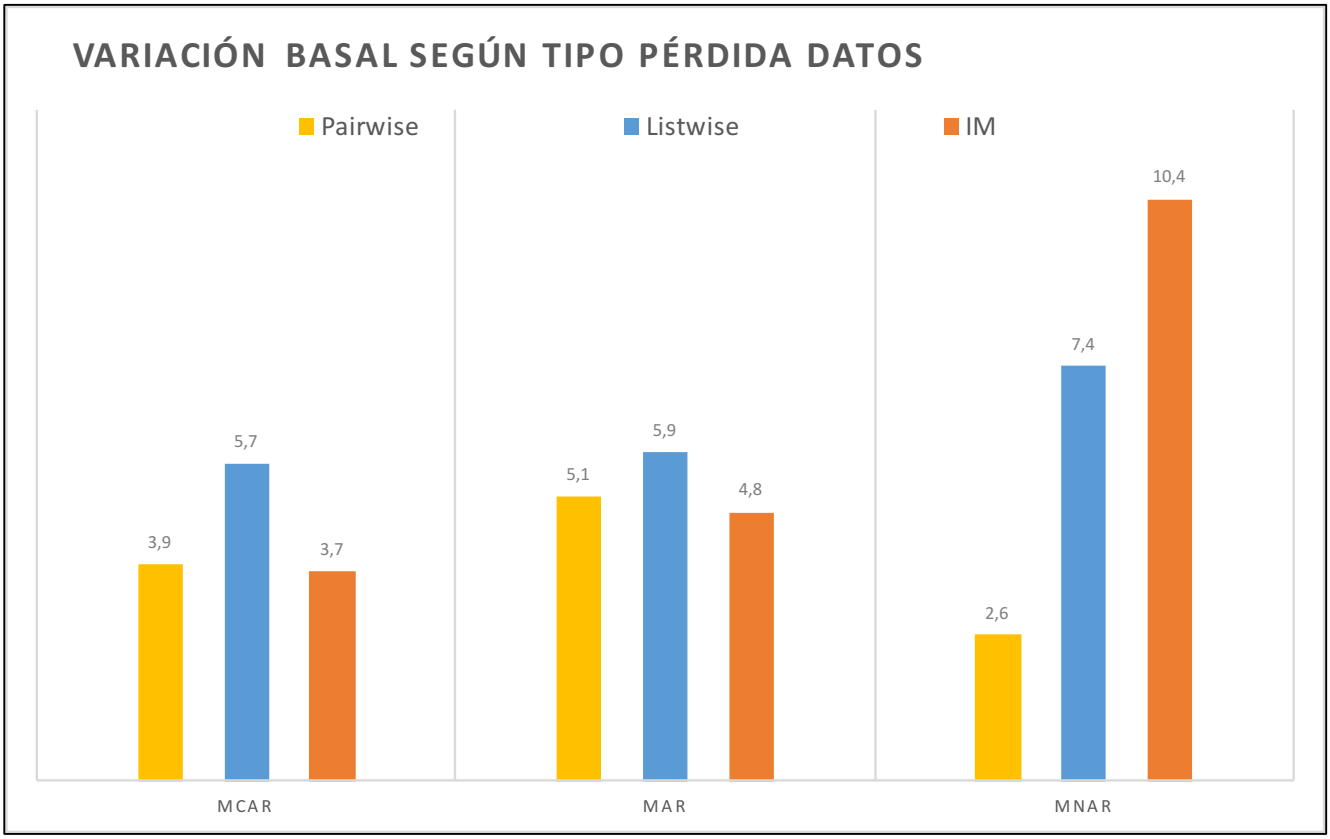
Si tenemos en cuenta el método usado, podemos ver que la *Imputación Múltiple* obtiene unos resultados mejores que el resto de métodos en situaciones MCAR y MAR. Como contraste, el caso concreto de MNAR, la *Imputación Múltiple* se dispara y es el método que más variación presenta, presentado el método de *Pairwise* la variación más pequeña de toda la serie (2.6 unidades).

Item	Metodo	MCAR_EQ5D10	MCAR_EQ5D20	MCAR_EQ5D30	MAR_EQ5D10	MAR_EQ5D20	MAR_EQ5D30	MNAR_EQ5D10	MNAR_EQ5D20	MNAR_EQ5D30	MCAR	MAR	MNAR
Movilidad	Pairwise	4	1	9	1	1	9	0	1	4	4,7	3,7	1,7
	Listwise	3	6	1	3	9	6	3	6	9	3,3	6,0	6,0
	IM	0	4	4	4	4	0	0	20	4	2,7	2,7	8,0
Cuidado personal	Pairwise	1	9	12	1	7	17	6	4	6	7,3	8,3	5,3
	Listwise	3	4	1	9	17	4	14	12	15	2,7	10,0	13,7
	IM	4	12	4	4	12	12	12	20	8	6,7	9,3	13,3
Actividades diarias	Pairwise	0	2	2	2	4	5	1	2	1	1,3	3,7	1,3
	Listwise	8	14	13	2	4	5	0	1	2	11,7	3,7	1,0
	IM	0	8	0	0	8	8	8	24	4	2,7	5,3	12,0
Dolor	Pairwise	1	0	8	4	1	12	1	1	6	3,0	5,7	2,7
	Listwise	3	4	1	9	9	6	9	12	15	2,7	8,0	12,0
	IM	4	4	4	4	4	4	4	20	8	4,0	4,0	10,7
Ansiedad	Pairwise	4	4	1	3	2	7	1	3	2	3,0	4,0	2,0
	Listwise	11	10	3	1	0	4	2	4	7	8,0	1,7	4,3
	IM	4	0	4	4	0	4	4	16	4	2,7	2,7	8,0
											4,4	5,2	6,8

MCAR			MNAR			MAR		
Pairwise	Listwise	IM	Pairwise	Listwise	IM	Pairwise	Listwise	IM
4,7	3,3	2,7	1,7	6,0	8,0	3,7	6,0	2,7
7,3	2,7	6,7	5,3	13,7	13,3	8,3	10,0	9,3
1,3	11,7	2,7	1,3	1,0	12,0	3,7	3,7	5,3
3,0	2,7	4,0	2,7	12,0	10,7	5,7	8,0	4,0
3,0	8,0	2,7	2,0	4,3	8,0	4,0	1,7	2,7
3,9	5,7	3,7	2,6	7,4	10,4	5,1	5,9	4,8

	MCAR	MAR	MNAR
Pairwise	3,9	5,1	2,6
Listwise	5,7	5,9	7,4
IM	3,7	4,8	10,4

Tabla 5.- Variación independiente del porcentaje de pérdida



Gráfica 2.- Variación independiente del porcentaje de perdida

3.2- Variable Índice

Hemos calculado la Media Global, Media Basal y Media Final de la variable junto con sus desviaciones estándar correspondiente. Además, lo hemos hecho de dos formas diferentes:

- Con todos los datos disponibles (*Listwise)*
- Realizando una *Imputación Múltiple*

En nuestra base de datos inicial no hay diferencias significativas entre el estado de calidad de vida Basal y el Final, y este es el resultado que nos gustaría encontrar con los diferentes métodos utilizados.

3.2.1- Listwise

Como podemos ver en la Tabla 6, obtenemos unos resultados de medias y desviaciones muy parecidos entre si. A grandes rasgos podemos ver como a medida que van aumentando las pérdidas, los resultados presentan más variación. Cabe destacar que en el mecanismo de pérdida MAR, parece ser que tenemos unas medias más estables y parece ser que MCAR presenta unas desviaciones menores.

El problema más obvio que tenemos es que al realizar el método de *Listwise*, estamos obteniendo tamaños muestrales diferentes. En una comparación como la que nos ocupa ahora, este problema nos impide poder realizar el test estadístico.

Recordar que estamos trabajando con muestras pareadas, por lo que necesitaremos tener los dos calidades de vida para poder analizar los datos. Por ello, el siguiente paso ha sido eliminar los pacientes que no tienen los dos cuestionarios completos. Obtenemos así los resultados que se muestran en la Tabla 7.

Los resultados referentes a la variación de las medias y las desviaciones son prácticamente iguales que antes. Pero lo importante aquí es ver que los resultados en todos los casos son iguales que el resultado real, es decir, sigue sin haber diferencias significativas entre el calidad de vida inicial y final.

Mención especial al mecanismo MCAR con un 30 % de pérdida de datos, que genera una muestra insuficiente para aplicar ningún test.

Listwise											
Mecanismo	Media global	sd Global	n (Media global)	Media basal	sd Basal	n (Media basal)	Media final	sd Final	n (Media final)	p_valor (t-test)	p_valor (Wilcoxon)
EQ5D	0,713	0,075	50	0,699	0,071	25	0,726	0,079	25	0.2159	0.2583
MCAR_EQ5D10	0,715	0,074	31	0,699	0,081	14	0,727	0,067	17		
MCAR_EQ5D20	0,724	0,073	18	0,706	0,082	8	0,738	0,067	10		
MCAR_EQ5D30	0,733	0,078	13	0,722	0,089	6	0,743	0,072	7		
MAR_EQ5D10	0,703	0,078	34	0,679	0,064	17	0,728	0,084	17		
MAR_EQ5D20	0,708	0,079	29	0,689	0,063	14	0,725	0,089	15		
MAR_EQ5D30	0,7	0,068	28	0,689	0,063	14	0,711	0,074	14		
MNAR_EQ5D10	0,744	0,077	33	0,733	0,072	14	0,753	0,069	19		
MNAR_EQ5D20	0,758	0,062	29	0,76	0,052	11	0,756	0,069	18		
MNAR_EQ5D30	0,765	0,051	28	0,76	0,052	11	0,768	0,051	17		
Listwise (Variación)											
Mecanismo	Media global	sd Global	n (Media global)	Media basal	sd Basal	n (Media basal)	Media final	sd Final	n (Media final)	p_valor (t-test)	p_valor (Wilcoxon)
EQ5D	0,713	0,075	50	0,699	0,071	25	0,726	0,079	25	0.2159	0.2583
MCAR_EQ5D10	0,002	0,001	31	0,000	0,010	14	0,001	0,012	17		
MCAR_EQ5D20	0,011	0,002	18	0,007	0,011	8	0,012	0,012	10		
MCAR_EQ5D30	0,020	0,003	13	0,023	0,018	6	0,017	0,007	7		
MAR_EQ5D10	0,010	0,003	34	0,020	0,007	17	0,002	0,005	17		
MAR_EQ5D20	0,005	0,004	29	0,010	0,008	14	0,001	0,010	15		
MAR_EQ5D30	0,013	0,007	28	0,010	0,008	14	0,015	0,005	14		
MNAR_EQ5D10	0,031	0,002	33	0,034	0,001	14	0,027	0,010	19		
MNAR_EQ5D20	0,045	0,013	29	0,061	0,019	11	0,030	0,010	18		
MNAR_EQ5D30	0,052	0,024	28	0,061	0,019	11	0,042	0,028	17		

Tabla 6.- Estudio de la variable Índice con todos los datos completos (*Listwise)*

Eliminamos los pacientes que se han quedado sin pareja											
Mecanismo	Media global	sd Global	n (Media global)	Media basal	sd Basal	n (Media basal)	Media final	sd Final	n (Media final)	p_valor (t-test)	p_valor (Wilcoxon)
EQ5D	0,713	0,075	50	0,699	0,071	25	0,726	0,079	25	0.2159	0.2583
MCAR_EQ5D10	0,701	0,074	18	0,689	0,071	9	0,712	0,08	9	0.5471	0.6523
MCAR_EQ5D20	0,727	0,06	6	0,68	0,044	3	0,773	0,023	3	0.0572	0.25
MCAR_EQ5D30	0,735	0,035	2	0,71	NA	1	0,76	NA	1	NA	NA
MAR_EQ5D10	0,707	0,077	30	0,689	0,061	15	0,725	0,089	15	0.2328	0.2931
MAR_EQ5D20	0,7	0,068	28	0,689	0,063	14	0,711	0,074	14	0.4336	0.4895
MAR_EQ5D30	0,7	0,068	28	0,689	0,063	14	0,711	0,074	14	0.4336	0.4895
MNAR_EQ5D10	0,746	0,054	20	0,747	0,075	10	0,746	0,025	10	0.9734	0.9188
MNAR_EQ5D20	0,752	0,044	18	0,763	0,058	9	0,741	0,02	9	0.3504	0.4738
MNAR_EQ5D30	0,752	0,044	18	0,763	0,058	9	0,741	0,02	9	0.3504	0.4738
VARIACIÓN											
Mecanismo	Media global	sd Global	n (Media global)	Media basal	sd Basal	n (Media basal)	Media final	sd Final	n (Media final)	p_valor (t-test)	p_valor (Wilcoxon)
EQ5D	0,713	0,075	50	0,699	0,071	25	0,726	0,079	25		
MCAR_EQ5D10	0,012	0,001	18	0,010	0,000	9	0,014	0,001	9		
MCAR_EQ5D20	0,014	0,015	6	0,019	0,027	3	0,047	0,056	3		
MCAR_EQ5D30	0,022	0,040	2	0,011	#¡VALOR!	1	0,034	#¡VALOR!	1		
MAR_EQ5D10	0,006	0,002	30	0,010	0,010	15	0,001	0,010	15		
MAR_EQ5D20	0,013	0,007	28	0,010	0,008	14	0,015	0,005	14		
MAR_EQ5D30	0,013	0,007	28	0,010	0,008	14	0,015	0,005	14		
MNAR_EQ5D10	0,033	0,021	20	0,048	0,004	10	0,020	0,054	10		
MNAR_EQ5D20	0,039	0,031	18	0,064	0,013	9	0,015	0,059	9		
MNAR_EQ5D30	0,039	0,031	18	0,064	0,013	9	0,015	0,059	9		

Tabla 7.- Eliminación de los pacientes sin los dos calidades de vida

3.2.2- Imputación Múltiple

Presentamos la Tabla 8 donde podemos ver que el comportamiento de las medias y las desviaciones son prácticamente los mismos que lo visto hasta ahora.

Pero aquí sí empezamos a ver resultados llamativos. Como podemos ver, tenemos dos resultados dispares en cuanto a la realidad de nuestra base de datos. Podemos observar que para un 10 % de pérdida de datos para el mecanismo MAR y MNAR los resultados son que sí hay diferencias significativas entre la calidad de vida Basal y Final.

IM											
Mecanismo	Media global	sd Global	n (Media global)	Media basal	sd Basal	n (Media basal)	Media final	sd Final	n (Media final)	p_valor (t-test)	p_valor (Wilcoxon)
EQ5D	0,713	0,075	50	0,699	0,071	25	0,726	0,079	25	0.2159	0.2583
MCAR_EQ5D10	0,713	0,08	50	0,696	0,078	25	0,731	0,079	25	0.1437	0.1412
MCAR_EQ5D20	0,714	0,072	50	0,692	0,066	25	0,735	0,073	25	0.0746	0.08505
MCAR_EQ5D30	0,721	0,077	50	0,708	0,079	25	0,734	0,075	25	0.3125	0.4261
MAR_EQ5D10	0,72	0,081	50	0,698	0,077	25	0,741	0,08	25	**0.04121**	**0.06263**
MAR_EQ5D20	0,709	0,075	50	0,702	0,066	25	0,717	0,085	25	0.4516	0.4418
MAR_EQ5D30	0,712	0,066	50	0,715	0,07	25	0,71	0,063	25	0.7796	0.6528
MNAR_EQ5D10	0,732	0,072	50	0,716	0,075	25	0,749	0,066	25	**0.03326**	**0.008214**
MNAR_EQ5D20	0,729	0,068	50	0,725	0,065	25	0,732	0,073	25	0.5473	0.2651
MNAR_EQ5D30	0,752	0,055	50	0,754	0,058	25	0,75	0,053	25	0.7484	0.9772
VARIACIÓN											
Mecanismo	Media global	sd Global	n (Media global)	Media basal	sd Basal	n (Media basal)	Media final	sd Final	n (Media final)	p_valor (t-test)	p_valor (Wilcoxon)
EQ5D	0,713	0,075	50	0,699	0,071	25	0,726	0,079	25		
MCAR_EQ5D10	0,000	0,005	50	0,003	0,007	25	0,005	0,000	25		
MCAR_EQ5D20	0,001	0,003	50	0,007	0,005	25	0,009	0,006	25		
MCAR_EQ5D30	0,008	0,002	50	0,009	0,008	25	0,008	0,004	25		
MAR_EQ5D10	0,007	0,006	50	0,001	0,006	25	0,015	0,001	25		
MAR_EQ5D20	0,004	0,000	50	0,003	0,005	25	0,009	0,006	25		
MAR_EQ5D30	0,001	0,009	50	0,016	0,001	25	0,016	0,016	25		
MNAR_EQ5D10	0,019	0,003	50	0,017	0,004	25	0,023	0,013	25		
MNAR_EQ5D20	0,016	0,007	50	0,026	0,006	25	0,006	0,006	25		
MNAR_EQ5D30	0,039	0,020	50	0,055	0,013	25	0,024	0,026	25		

Tabla 8.- Resultados del análisis de la variable índice mediante Imputación Múltiple

4. Interpretación de Resultados

4.1- Fase Descriptiva

Como hemos podido observar, el porcentaje de pérdida es crítico a la hora de trabajar con bases de datos que contienen datos faltantes. Este es un hecho obvio, a mayor falta de información, más difícil será obtener valores más reales. Un tarea importante en este punto, sería poder estimar a partir de qué porcentaje de pérdida nos debemos preocupar. En una fase puramente descriptiva, deberíamos definir qué variación estamos dispuestos a tener, y en función de ello, podemos fijar que porcentaje de pérdida podemos asumir.

Si nos centramos en los diferentes mecanismos que hemos usado para trabajar nuestras pérdidas, podemos concluir que otro punto importante es conocer qué mecanismo está causando nuestra pérdida de datos. Como se puede observar, el método que menos variación presenta es el de la *Imputación Múltiple*, siempre y cuando no se trata de un patrón MCAR, en cuyo caso la mejor opción seria *Pairwise*.

En ningún caso sería buena opción *Listwise* , pues se reduce mucho el tamaño de la muestra y en un caso como el nuestro que ya partimos de una base de datos pequeña, es una penalización que no podemos asumir.

Si se nos plantea una situación donde tenemos los tres mecanismos de pérdida en nuestra base de datos, nos inclinaríamos por el método de *Pairwise*, puesto que de manera general, sin tener en cuenta el mecanismo, presenta unas variaciones menores. En este caso, la *Imputación Múltiple* pierde potencia. Parece ser que el hecho de contener pérdidas MNAR es un lastre bastante grande. Hay que tener en cuenta que para que funcione correctamente la *Imputación Múltiple* se han de cumplir una serie de requisitos, como que los datos faltantes sean MAR (MCAR también funciona bien con unos ajustes) y además no siempre es fácil encontrar un modelo adecuado para la variable de interés.

Otro aspecto importante que subyace de todo lo dicho es que debemos intentar, dentro de la medida de lo posible, evitar los mecanismos de pérdida MNAR. Alteran al mejor método que hemos encontrado en nuestro estudio y además, es el mecanismo que más valores dispares ofrece. Si tuviéramos este tipo de mecanismo actuando sobre la pérdida de nuestros datos, se podría optar por realizar la técnica de *Pairwise*, que es donde hemos visto menos variación.

4.2- Variable Índice

Para este tipo de estudio, ya no son tan importante las variaciones entre los resultados, puesto que vamos a ver diferencias significativas entre dos poblaciones. Variaciones pueden existir, ahora hay que ver si son lo suficientemente importantes como para afectar a nuestros resultados. Así pues, analizar los resultados mediante la variable Índice es un método más robusto que el anterior.

Como hemos podido ver, el principal problema que tenemos al realizar el análisis sin imputar los datos es que obtenemos unos tamaños muestrales diferentes, cosa que nos va a perjudicar porque debilita la significancia de las pruebas. Además, en nuestro caso en concreto donde nuestros datos son pareados, debemos reducir más aún la muestra eliminando a los pacientes que no dispongan de los dos calidades de vida. Si ya de partida venimos de una base de datos con una n pequeña, esto puede ser catastrófico, como se puede ver en el caso MCAR_EQ5D30, donde solo tenemos un caso en cada grupo y no se puede realizar el test. Este punto se puede solucionar fácilmente teniendo una muestra lo suficientemente grande como para poder soportar una perdida de datos.

Por lo dicho anteriormente, el imputar nuestros datos cuando nuestra base de datos no es lo suficientemente grande como para soportar tal pérdida, debería de ser la mejor opción a valorar. De esta forma, volveríamos a tener la misma *n* que al inicio del estudio. Ahora bien, también hay que ser conscientes de que no es una técnica exenta de restricciones. Restricciones que si no se cumplen pueden dar lugar a resultados no deseados como ha sido nuestro caso.

Podemos ver como tenemos dos situaciones donde sí se están viendo diferencias significativas entre la calidad de vida Basal y la Final. Nos referimos a la situación MAR_EQ5D10 y MNAR_EQ5D10.

El hecho de que aparezcan cuando tenemos un porcentaje de pérdida más pequeño (10%) nos podría estar indicando simplemente que al tener menos pérdida, tenemos mas datos para comparar, y eso hace que podamos ver más diferencias entre los grupos (ésto va en consonancia con lo visto en los resultados de las bases de datos que han quedado al eliminar los pacientes que no disponían de sus dos calidades de vida).

Pero es obvio que algo no funciona como debería, pues nos está mostrando diferencias cuando en realidad no las hay. Para intentar explicarlo se pueden plantear dos propuesta que deberían ser estudiadas:

- Cuando realizamos el test de Wilcoxon podemos observar como en algunas comparaciones nos hace un aviso[6] donde nos está indicando que tenemos muestras repetidas. Ésto hace que perdamos variabilidad en nuestra muestra y por lo tanto no podremos calcular unos *p_valores* exactos.

  Habría que tener en cuenta que al realizar un test con muestras pareadas se está realizando un test de simetría y se está asumiendo que nuestras distribuciones no son sesgadas, y en este tipo de base de datos, no es tan raro encontrar el mismo resultado varias veces. Ésto acompañado de una muestra inicial pequeña puede hacer un problema a la hora de tener resultados repetidos.

- Debemos de conocer bien la técnica de la *Imputación Múltiple*. Partimos de una base de datos pequeña, donde estamos imputando variables ordinales y con pocos niveles. Hemos de saber que existen situaciones en las que los supuestos del método no se cumplen y que no siempre es fácil encontrar un modelo adecuado para la variable de interés.

NOTA: Tras los comentarios de nuestra directora del TFM, realizamos la imputación de la variable "Índice" mediante el método de "pmm", el cual está indicado especialmente para imputar variables cuantitativas que no están normalmente distribuidas.

Recordemos que la variable en cuestión es cuantitativa, no ordinal, y a demás, teníamos duda en cuanto a su distribución. El resultado lo podemos ver en la Tabla 9

IM (pmm)								
Mecanismo	Media global	n (Media global)	Media basal	n (Media basal)	Media final	n (Media final)	p_valor (t-test)	p_valor (Wilcoxon)
EQ5D	0.713 (0.075)	50	0.699 (0.071)	25	0.726 (0.079)	25	0.2159	0.2583
MCAR_EQ5D10	0.709 (0.07)	50	0.701 (0.075)	25	0.716 (0.066)	25	0.5069	0.4418
MCAR_EQ5D20	0.706 (0.067)	50	0.691 (0.058)	25	0.72 (0.073)	25	0.1871	0.173
MCAR_EQ5D30	0.738 (0.09)	50	0.722 (0.104)	25	0.754 (0.073)	25	0.2438	0.2259
MAR_EQ5D10	0.712 (0.073)	50	0.7 (0.069)	25	0.723 (0.076)	25	0.2404	0.3957
MAR_EQ5D20	0.721 (0.069)	50	0.711 (0.062)	25	0.731 (0.078)	25	0.3015	0.4042
MAR_EQ5D30	0.71 (0.069)	50	0.704 (0.061)	25	0.716 (0.078)	25	0.5365	0.4839
MNAR_EQ5D10	0.73 (0.07)	50	0.712 (0.067)	25	0.748 (0.069)	25	**0.01211**	**0.009312**
MNAR_EQ5D20	0.735 (0.07)	50	0.734 (0.062)	25	0.735 (0.077)	25	0.9612	0.5677
MNAR_EQ5D30	0.748 (0.068)	50	0.741 (0.076)	25	0.756 (0.059)	25	0.381	0.3104

Tabla 9.- Resultados del análisis de la variable índice mediante Imputación Múltiple (método "pmm")

Como podemos ver ahora, todas las imputaciones excepto las realizadas con una base de datos que posee un mecanismo de pérdida MNAR al 10%, son no clínicamente significativas, lo que coincide con el resultado real.

El hecho de que no funcione bien el método de la *Imputación Múltiple* cuando el mecanismo de pérdida es MNAR ya lo conocíamos y puede que sea la explicación mas plausible para este resultado.

[6] *Warning in wilcox.test.default(x = c(0.68, 0.6, 0.69, 0.83, 0.6, 0.76, 0.69, : cannot compute exact p-value with ties*

5. Conclusiones

Las conclusiones que obtenemos de nuestro trabajo son las siguientes:

- Debemos de conocer la estructura de nuestra base de datos y las limitaciones de nuestro estudio.
- A mayor porcentaje de pérdida, mayor variación obtenemos. Es importante estimar qué porcentaje de pérdida de datos estamos dispuestos a asumir en nuestra investigación.
- Hemos de conocer el mecanismo de pérdida de datos que está afectando a nuestra investigación.
- Finalmente, hemos de conocer los diferentes métodos que existen para tratar los datos perdidos.

Además, podemos concluir que para un caso como el nuestro, donde partimos de un estudio longitudinal pequeño, con muestras pareadas y variables ordinales con pocos niveles:

- El método de *Imputación Múltiple* es el más efectivo para mecanismos de pérdida MCAR y MAR.
- La *Imputación Múltiple* nos da la ventaja de no perder tamaño muestral.
- Para mecanismo MNAR, la técnica de *Pairwise* sería la mas apropiada.
- En el caso de presentar los tres mecanismos de pérdida en nuestro estudio, deberíamos utilizar *Pairwise.*
- En ningún caso es aconsejable utilizar el método *Listwise,* pues produce una pérdida del tamaño de la muestra.
- Hemos de evitar las pérdidas MNAR

6. Observaciones finales

Inicialmente este estudio pretendía ser más ambicioso. En principio se quería trabajar con una base de datos real y trabajar la pérdida de datos con más técnicas de las propuestas. Debido a la falta de tiempo se decidió partir de una base de datos en banco e ir creándola desde cero. Ésto nos ha limitado de manera considerable el posterior tratamiento de los valores perdidos.

Pero este contratiempo nos benefició en el sentido de haber podido orientar de una forma mas concreta nuestra investigación hacia una necesidad existente y que no se ha tratado tan a fondo en la literatura.

Las conclusiones generales son las mismas que se plantean en todos los estudios encontrados sobre valores perdidos y las conclusiones particulares de nuestra base de datos tampoco difieren tanto.

Este trabajo no deja de ser una aproximación a una situación tan compleja como la que implican los datos faltantes, y como tal, aún queda mucho trabajo por hacer. A nivel de este estudio el siguiente paso sería encontrar el motivo real por el que obtenemos unos resultados diferentes mediante la técnica de *Imputación Múltiple* cuando analizamos la variable índice. Hemos propuesto varias alternativas, pero convendría confirmarlas.

Pensando en un futuro, se podría ampliar este estudio de diferentes formas:

- Aplicando más técnicas, e incluso alguna más especifica
- Ver los resultados con un mayor porcentaje de pérdida y también con un tamaño muestral más grande.
- Encontrar nuevas covariables que pudieran ayudarnos

Por último, encontrar un mecanismo de imputación específico o crear funciones en R específicas para realizar un tratamiento básico de las bases de datos, facilitando así la integración de los investigadores mas noveles en este campo sería una buena meta final.

7. Glosario

- **TFM:** Trabajo Fin de Master
- **Fully Conditionally Specification** (FCS)
- **MICE** (Multiple Imputtion by Chained Equations)
- **reclass (BD)**: Función creada en R para re-clasificar las variables importadas de Excel.
- **DBprepare (BD)**: Función creada en R para preparar automáticamente la base de datos importada de Excel
- **MCAR_EQ5D10**: Base de datos con pérdidas MCAR al 10%
- **MCAR_EQ5D20**: Base de datos con pérdidas MCAR al 20%
- **MCAR_EQ5D30**: Base de datos con pérdidas MCAR al 30%
- **MAR_EQ5D10:** Base de datos con pérdidas MAR al 10%
- **MAR_EQ5D20**: Base de datos con pérdidas MAR al 20%
- **MAR_EQ5D30**: Base de datos con pérdidas MAR al 30%
- **MNAR_EQ5D10**: Base de datos con pérdidas MNAR al 10%
- **MNAR_EQ5D20**: Base de datos con pérdidas MNAR al 20%
- **MNAR_EQ5D30**: Base de datos con pérdidas MNAR al 30%
- **Descriptiva (BD)**: Función creada en R para obtener los resultados de la parte descriptiva del estudio.
- **Medias (BD)**: Función de R creada para obtener la media y la desviación estándar en los cuestionarios de manera global, basal y final.
- **ComparaMedias (BD):** Función de R creada para realizar el test de t-Student y Wilcoxon.
- **listwiseDB (DB):** Función creada en R para aplicar la función '*Descriptiva(DB)*' solo en los casos completos de nuestra base de datos.
- **MCAR_EQ5D10comp:** Base de datos compensadas con pérdidas MCAR al 10%
- **MCAR_EQ5D20comp**: Base de datos compensadas con pérdidas MCAR al 20%
- **MCAR_EQ5D30comp:** Base de datos compensadas con pérdidas MCAR al 30%
- **MAR_EQ5D10comp:** Base de datos compensadas con pérdidas MAR al 10%
- **MAR_EQ5D20comp:** Base de datos compensadas con pérdidas MAR al 20%
- **MAR_EQ5D30comp:** Base de datos compensadas con pérdidas MAR al 30%
- **MNAR_EQ5D10comp:** Base de datos compensadas con pérdidas MNAR al 10%
- **MNAR_EQ5D20comp:** Base de datos compensadas con pérdidas MNAR al 20%
- **MNAR_EQ5D30comp:** Base de datos compensadas con pérdidas MNAR al 30%
- **MICE():** Librería de R que nos permite realizar la Imputación Múltiple
- **MCAR_EQ5D10_i:** Base de datos imputada con pérdidas MCAR al 10%
- **MCAR_EQ5D20_i:** Base de datos imputada con pérdidas MCAR al 20%
- **MCAR_EQ5D30_i:** Base de datos imputada con pérdidas MCAR al 30%
- **MAR_EQ5D10_i:** Base de datos imputada con pérdidas MAR al 10%
- **MAR_EQ5D20_i:** Base de datos imputada con pérdidas MAR al 20%
- **MAR_EQ5D30_i:** Base de datos imputada con pérdidas MAR al 30%
- **MNAR_EQ5D10_i:** Base de datos imputada con pérdidas MNAR al 10%
- **MNAR_EQ5D20_i:** Base de datos imputada con pérdidas MNAR al 20%
- **MNAR_EQ5D30_i:** Base de datos imputada con pérdidas MNAR al 30%

- **MCAR_EQ5D10_new:** Base de datos imputada tras calcular las variables *perfil e índice* con la herramienta del EUROQOL*, con* pérdidas MCAR al 10%
- **MCAR_EQ5D20_new:** Base de datos imputada tras calcular las variables *perfil e índice* con la herramienta del EUROQOL*,* con pérdidas MCAR al 20%
- **MCAR_EQ5D30_new:** Base de datos imputada tras calcular las variables *perfil e índice* con la herramienta del EUROQOL*,* con pérdidas MCAR al 30%
- **MAR_EQ5D10_new:** Base de datos imputada tras calcular las variables *perfil e índice* con la herramienta del EUROQOL*,* con pérdidas MAR al 10%
- **MAR_EQ5D20_new:** Base de datos imputada tras calcular las variables *perfil e índice* con la herramienta del EUROQOL*,* con pérdidas MAR al 20%
- **MAR_EQ5D30_new:** Base de datos imputada tras calcular las variables *perfil e índice* con la herramienta del EUROQOL*,* con pérdidas MAR al 30%
- **MNAR_EQ5D10_new:** Base de datos imputada tras calcular las variables *perfil e índice* con la herramienta del EUROQOL*,* con pérdidas MNAR al 10%
- **MNAR_EQ5D20_new:** Base de datos imputada tras calcular las variables *perfil e índice* con la herramienta del EUROQOL*,* con pérdidas MNAR al 20%
- **MNAR_EQ5D30_new:** Base de datos imputada tras calcular las variables *perfil e índice* con la herramienta del EUROQOL*,* con pérdidas MNAR al 30%

8. Bibliografia

[1] WOOD, A. M., I. R. WHITE and S. G. THOMPSON (2004): Are missing outcome data adequately handled? A review of published randomized controlled trials in major medical journals. **Clinical trials,** 1, 368-376.

[2] DÍAZ-ORDAZ, K., M. G. KENWARD, A. COHEN, C. L. COLEMAN and S. ELDRIDGE (2014): Are missing data adequately handled in cluster randomised trials? A systematic review and guidelines. **Clinical Trials,** II(5), 590-600.

[3] BUUREN, S. (2012): **Flexible imputation of missing data.** Chapman & Hall/CRC Press, London-New York.

[4] HAND, D. J., F. DALY, A. D. LUNN, K. J. MCCONWAY and E. OSTROWSKI (1994): **A handbook of small data sets.** Chapman & Hall, London.

[5] Horton, N. & Lipsitz, S. (2001), 'Multiple Imputation in Practice: Comparison of Software Packages for Regression Models With Missing Variables', American Statistical Association 55 (3), 244–254.

[6] Dempster, A. P., Laird, N. M. & Rubin, D. B. (1977), 'Maximum Likelihood from Incomplete Data Via the EM Algorithm', Journal of the Royal Statistical 39 , 1–38.

[7] Schafer,J.L (1997), Analysis of incomplete multivariate data, London, Chapman y Hall.

[8] Ibrahim, J. G., (1990), Incomplete data in generalized linear models, Journal of the American Statistical Association .

[9] Little, R. J. A. (1995), Modeling the dropout mechanism in repeated-measures studies, Journal of the American Statistical Association .

[10] Little, R. J. A. (1992), Regression with missing X's, A review, Journal of the American Statistical Association 87.

[11] Little, R. J., y D. Rubin (1987). Statistical analysis with missing data, New York, Wiley.

[12] Rubin, D. B. (1976), Inference and missing data, Biometrika 63.

[13] Shona Fielding, Peter M Fayers and Craig R Ramsay, Investigating the missing data mechanism in quality of life outcomes: a comparison of approaches, Health and Quality of Life Outcomes 2009, 7:57 doi:10.1186/1477-7525-7-57

[14] Little, R. J. A. (1986), A test of Missing Completely at Random for multivariate data with missing values, Sociological Methods and Research 18.

[15] Listing J, Schlittgen R: Tests if dropouts are missed at random. Biometrical Journal 1998, 40:929-935.

[16] Listing J, Schlittgen R: nonparametric test for random dropouts. Biometrical Journal 2003, 45:113-127.

[17] Schmitz N, Franz M: A bootstrap method to test if study dropouts are missing randomly. Quality & Quantity 2002, 36:1-16.

[18] Diggle PJ: Testing for random dropouts in repeated measurements data. Biometrics 1989, 45:1255-1258.

[19] Ridout MS: Testing for random dropouts in repeated measurement data. Biometrics 1991, 47:1617-1619.

[20] Fairclough DL: and Analysis of Quality of Life Studies in Clinical Trials Chapman and Hall; 2002.

[21] Kalton, G. y D. Kasprzyk (1982), Imputing for Missing Surveys Responses, Proceedings of the Section on Survey Research Methods, American Statistical Association.

[22] Schafer, J. L. y J. W. Graham (2002), Missing Data, Our View of the State of the Art, Psychological Methods .vol. 7, No. 2.

[23] Acock, C. A., y D. Demo (1994), Family diversity and well-being, Thousand Oaks, C. A. Sage.

[24] Cohen, J., y P. Cohen (1983), Applied multiple regression/correlation analysis for the behavioral sciences (Sec. ed.), Hillsdale, N. J., Erlbaum.

[25] Cohen, J., P. Cohen, S. West, y L. Aiken, (2003), Applied multiple regression/correlation analysis for the behavioral sciences (Third ed.), Mahwah, N. J. Erlbaum.

[26] Madow, W. G., J. Nisselson y I. Olkin (Eds.) (1983), Incomplete data in sample surveys, vol. 1, Report and case studies, New York, Academic Press.

[27] Durrant, B. G. (2005), Imputation Methods for Handling Item-Non-response in the Social Science: A Methodological Review, ESRC National Centre for Research Methods and Southampton Statistical Science Research Institute, University of Southampton, NCRM Methods Review Papers , NCMR/002.

[28] Dempster, A. P., N. M. Lair, y D. B. Rubin (1977), Maximum likelihood estimation from incomplete data via the EM algorithm (with discussion). Journal of the Royal Statistical Society, Series B , 39.

[29] Stef van Buuren and Karin Groothuis-Oudshoorn, MICE: Multivariate Imputation by Chained Equations in R. Journal of Statistical Software

9. Anexos

ANEXO 1.- Modelo calidad de vida EQ-5D-3L

Figure 1: EQ-5D-3L (UK English sample version)

By placing a tick in one box in each group below, please indicate which statements best describe your own health state today.

Mobility

I have no problems in walking about	❑
I have some problems in walking about	❑
I am confined to bed	❑

Self-Care

I have no problems with self-care	❑
I have some problems washing or dressing myself	❑
I am unable to wash or dress myself	❑

Usual Activities (*e.g. work, study, housework, family or leisure activities)*

I have no problems with performing my usual activities	❑
I have some problems with performing my usual activities	❑
I am unable to perform my usual activities	❑

Pain/Discomfort

I have no pain or discomfort	❑
I have moderate pain or discomfort	❑
I have extreme pain or discomfort	❑

Anxiety/Depression

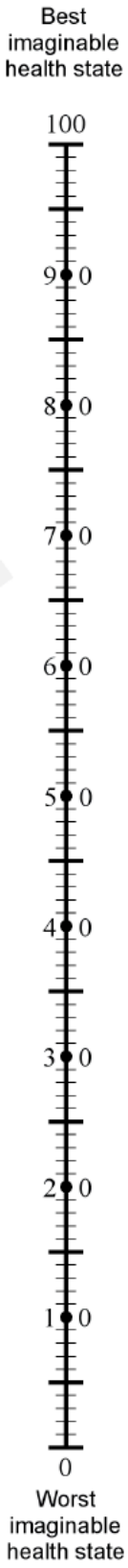
I am not anxious or depressed	❑
I am moderately anxious or depressed	❑
I am extremely anxious or depressed	❑

SAMPLE

To help people say how good or bad a health state is, we have drawn a scale (rather like a thermometer) on which the best state you can imagine is marked **100** and the worst state you can imagine is marked **0**.

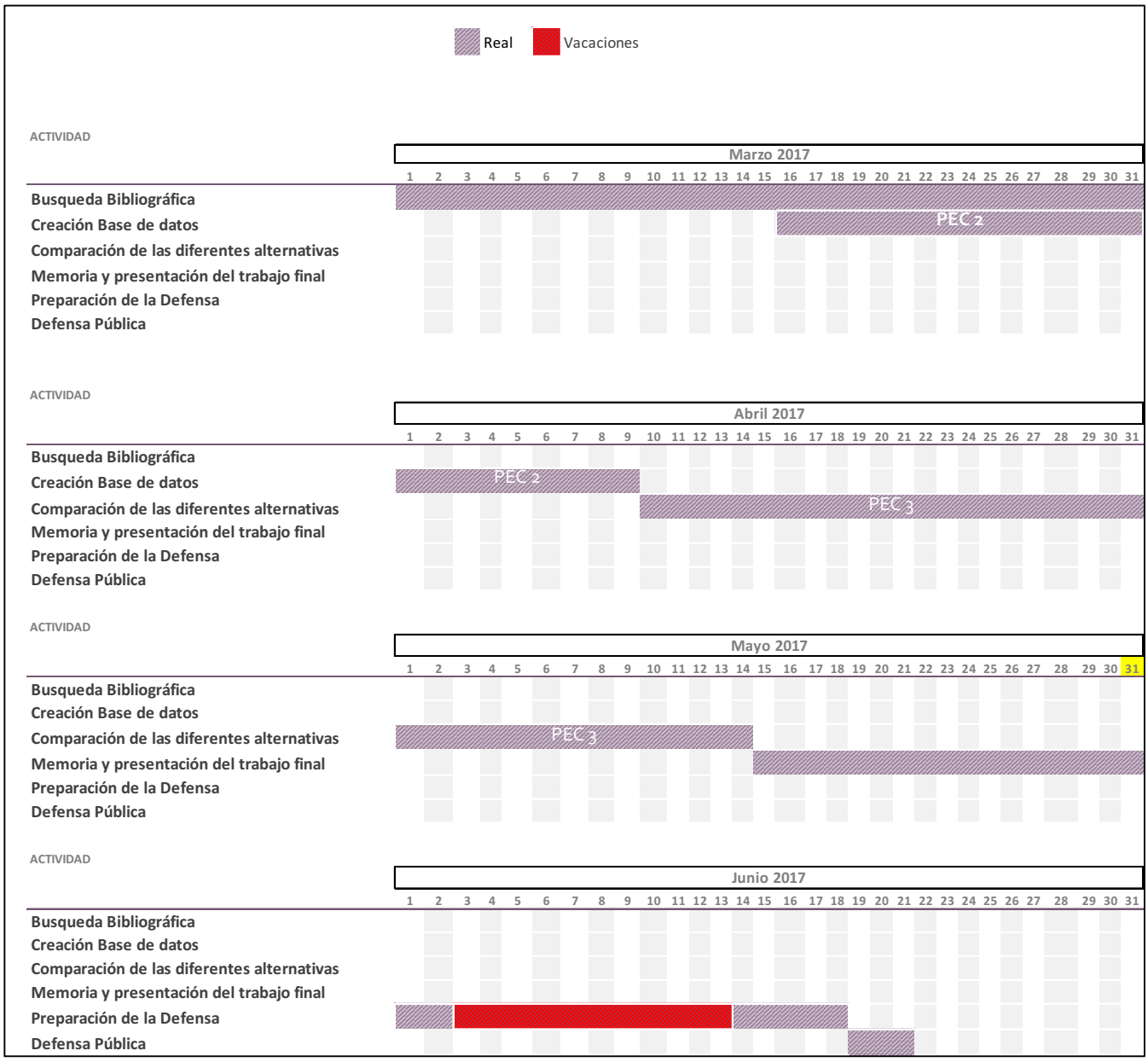
We would like you to indicate on this scale how good or bad your own health is today, in your opinion. Please do this by drawing a line from the box below to whichever point on the scale indicates how good or bad your health state is today.

Your own health state today

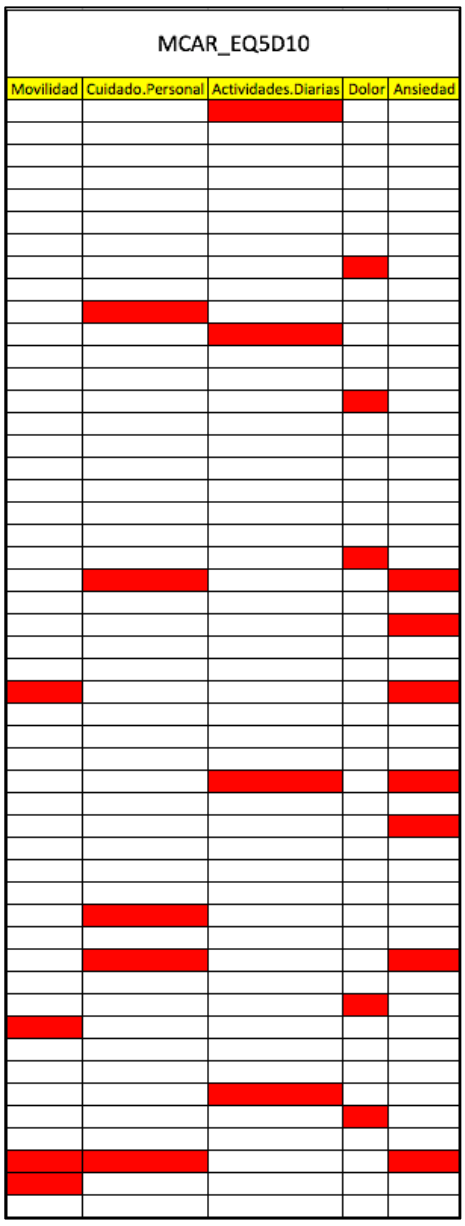


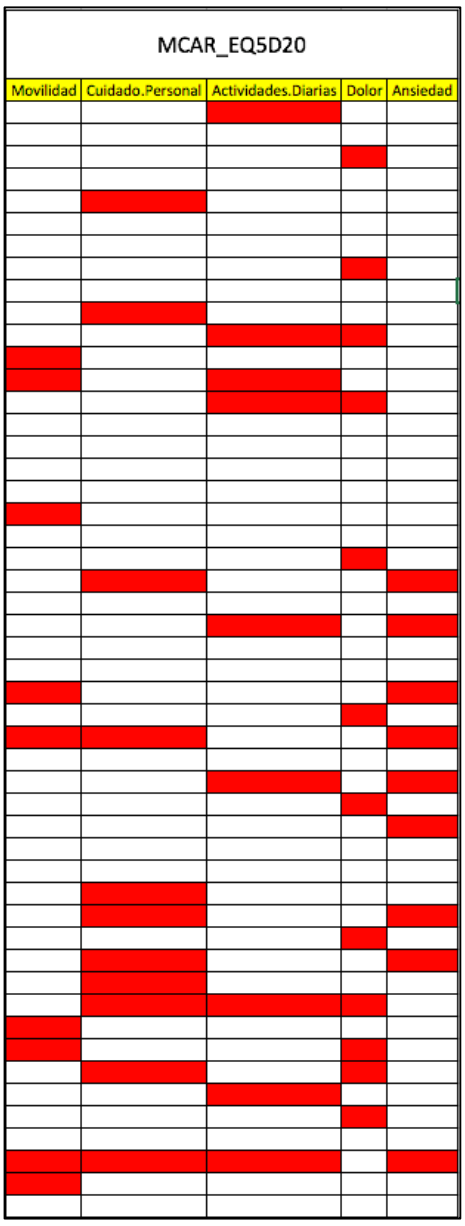
SAMPLE

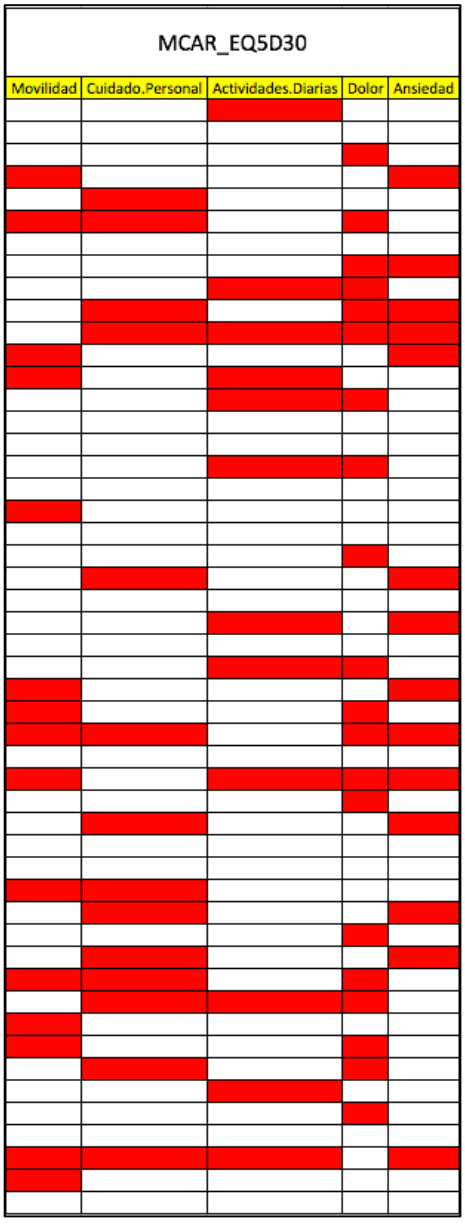
ANEXO 2.- Planificación del trabajo

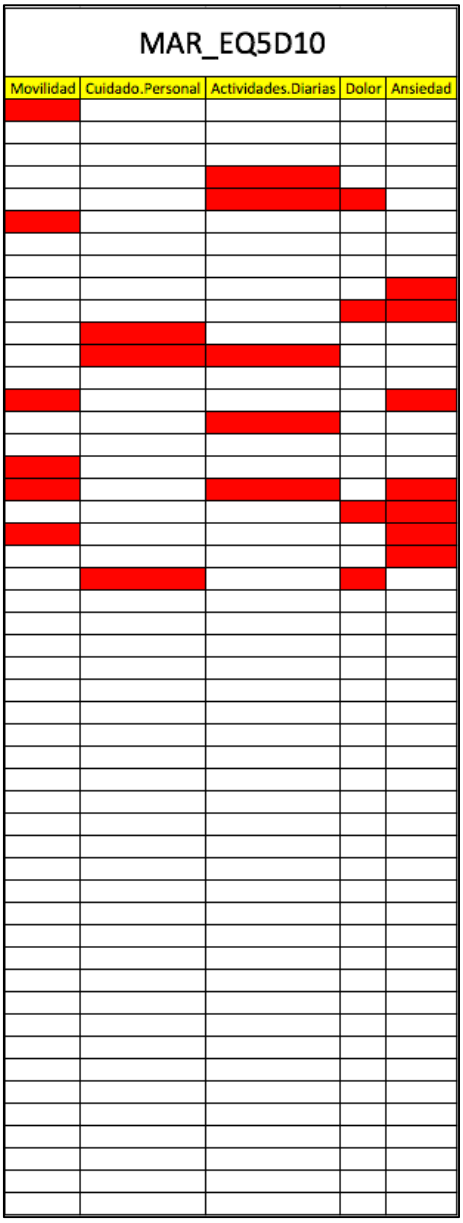


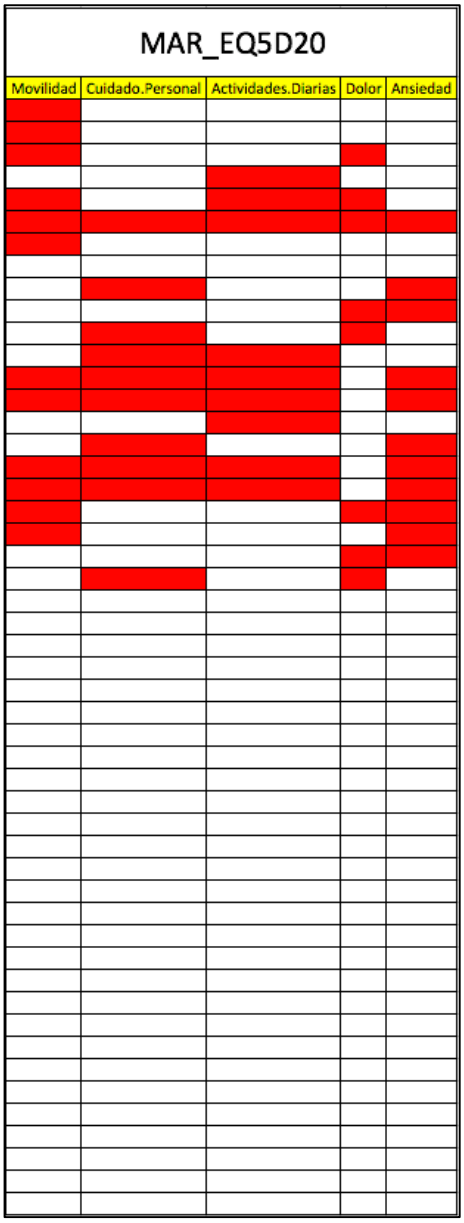
ANEXO 3.- Observación de los patrones de pérdida de datos

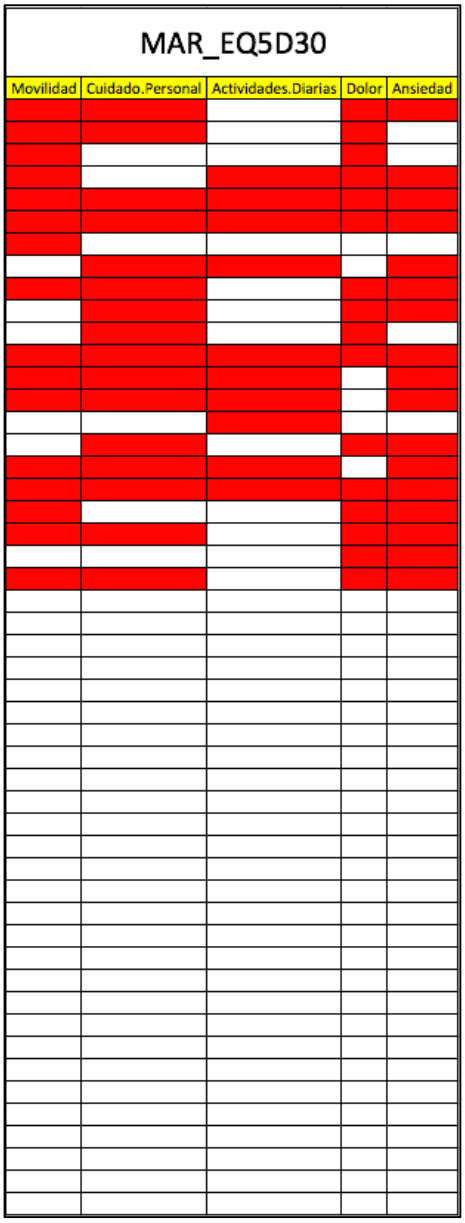




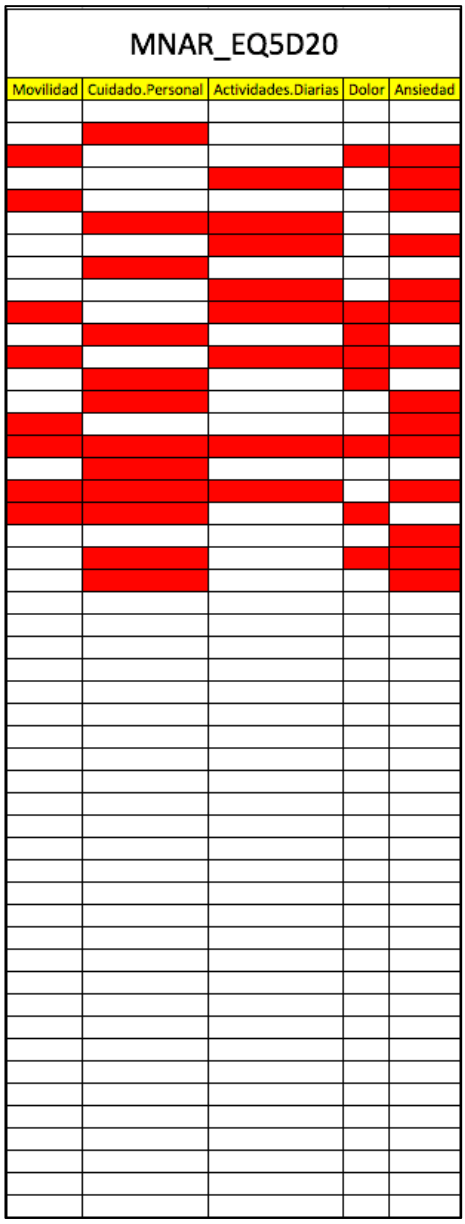


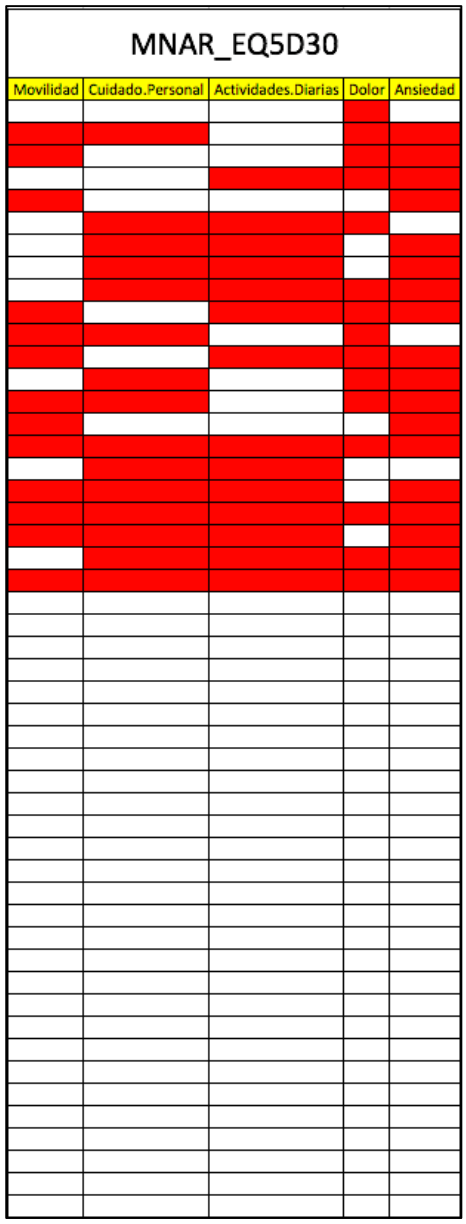

MAR_EQ5D10
Movilidad
Cuidado.Personal
Actividades.Diarias
Dolor
Ansiedad


MAR_EQ5D20
Movilidad
Cuidado.Personal
Actividades.Diarias
Dolor
Ansiedad


MAR_EQ5D30
Movilidad
Cuidado.Personal
Actividades.Diarias
Dolor
Ansiedad


MNAR_EQ5D10
Movilidad
Cuidado.Personal
Actividades.Diarias
Dolor
Ansiedad


MNAR_EQ5D20
Movilidad
Cuidado.Personal
Actividades.Diarias
Dolor
Ansiedad


MNAR_EQ5D30
Movilidad
Cuidado.Personal
Actividades.Diarias
Dolor
Ansiedad

ANEXO 4.- Código R usado

```
# Función para llamar y preprar nuestra base de datos

reclass <- function(BD){
  for (i in 1:length(BD))
    ifelse(class(BD[,i])!="numeric",
      BD[,i]<-as.factor(BD[,i]),
      BD[,i]<-as.numeric(BD[,i]))
  return(BD)}

DBprepare<-function(DB,sheet=1){
  library(gdata)
  setwd("~/Dropbox/UOC/ASIGNATURAS/TFM")
  archivo<-paste(getwd(),"/",DB,".xlsx",sep = "")
  EQ5D<-read.xls(archivo, sheet)
  EQ5D<-as.data.frame(reclass(EQ5D))
  EQ5D$Movilidad<-ordered(EQ5D$Movilidad,levels=c("3","2","1")) #Definimos como variable ordenada co
n el orden deseado (por defecto R toma orden alfabetico)
  EQ5D$Cuidado.Personal<-ordered(EQ5D$Cuidado.Personal,levels=c("3","2","1"))
  EQ5D$Actividades.Diarias<-ordered(EQ5D$Actividades.Diarias,levels=c("3","2","1"))
  EQ5D$Dolor<-ordered(EQ5D$Dolor,levels=c("3","2","1"))
  EQ5D$Ansiedad<-ordered(EQ5D$Ansiedad,levels=c("3","2","1"))
  save(EQ5D, file = "DB.rda");print(str(EQ5D))
  load<-load("DB.rda",envir = .GlobalEnv)
}
```

```
# Ahora crearemos una nueva columna con la variable "estudios" que usaremos posteriormente para pode
r gestionar el tipo de mecanismo de pérdidas

set.seed(19)
estudios<-sample(0:1,nrow(EQ5D)/2, replace = T) # Los pacientes que tienen estudios superiores serán
 codificados como "1" y los que no, como "0"
EQ5D<-EQ5D[order(EQ5D$Timepoint),] # Ordenamos por timepoint para asegurarnos que asignamos al mimo
 paciente el mismo resultado
EQ5D$estudios<-as.factor(estudios) # Añadimos localización
EQ5D<-EQ5D[order(EQ5D$Paciente),] # Volvemos a ordenar por paciente
summary(EQ5D)
```

```
# Creamos el 10 % de pérdidas MCAR

set.seed(17)
rows<-sample(1:50, 100, replace = T)
columns<-sample(3:7, 100,replace = T)
x<-EQ5D

for(i in 1:25){
  ifelse(is.na(x[rows[i],columns[i]]),
         x[rows[i+1],columns[i]]<-NA,
         x[rows[i],columns[i]]<-NA)
}

sum(is.na(x[,3:7]))
```

```
## [1] 25
```

```
MCAR_EQ5D10<-x
for(i in which(!complete.cases(MCAR_EQ5D10))){
  MCAR_EQ5D10[i,8:9]<-NA
}
```

```
# Creamos el 20 %

for(i in 26:51){
  ifelse(is.na(x[rows[i],columns[i]]),
         x[rows[i+1],columns[i]]<-NA,
         x[rows[i],columns[i]]<-NA)
}

sum(is.na(x[,3:7]))
```

```
## [1] 50
```

```
MCAR_EQ5D20<-x
for(i in which(!complete.cases(MCAR_EQ5D20))){
  MCAR_EQ5D20[i,8:9]<-NA
}
```

```
# Creamos el 30 %

for(i in 51:76){
  ifelse(is.na(x[rows[i],columns[i]]),
         x[rows[i+1],columns[i]]<-NA,
         x[rows[i],columns[i]]<-NA)
}

sum(is.na(x[,3:7]))
```

```
## [1] 75
```

```
MCAR_EQ5D30<-x
for(i in which(!complete.cases(MCAR_EQ5D30))){
  MCAR_EQ5D30[i,8:9]<-NA
}
```

```r
# Seleccionamos solo los pacientes sin estudios superiores "0" y los vamos eliminando. Creamos el 10
 % de pérdidas MAR

ordered<-EQ5D[order(EQ5D$estudios),] # Ordenamos primero para mayor comodidad (Sin estudios primero)

# Creamos unos nuevos valores de selección
set.seed(17)
rows2<-sample(1:22, 200, replace = T) # Los primero 22 son de fuera de la CV
columns2<-sample(3:7, 200,replace = T)

# Creamos el 10 % de MAR

for(i in 1:25){
  ifelse(is.na(ordered[rows2[i],columns2[i]]),
         ordered[rows2[i+1],columns2[i]]<-NA,
         ordered[rows2[i],columns2[i]]<-NA)
}

sum(is.na(ordered))
```

```
## [1] 25
```

```r
MAR_EQ5D10<-ordered
for(i in which(!complete.cases(MAR_EQ5D10))){
  MAR_EQ5D10[i,8:9]<-NA
}
```

```r
# Creamos el 20 % de MAR

for(i in 26:59){
  ifelse(is.na(ordered[rows2[i],columns2[i]]),
         ordered[rows2[i+1],columns2[i]]<-NA,
         ordered[rows2[i],columns2[i]]<-NA)
}

sum(is.na(ordered))
```

```
## [1] 50
```

```r
MAR_EQ5D20<-ordered
for(i in which(!complete.cases(MAR_EQ5D20))){
  MAR_EQ5D20[i,8:9]<-NA
}
```

```r
# Creamos el 30 % de MAR

for(i in 60:99){
  ifelse(is.na(ordered[rows2[i],columns2[i]]),
         ordered[rows2[i+1],columns2[i]]<-NA,
         ordered[rows2[i],columns2[i]]<-NA)
}

sum(is.na(ordered))
```

```
## [1] 75
```

```r
MAR_EQ5D30<-ordered
for(i in which(!complete.cases(MAR_EQ5D30))){
  MAR_EQ5D30[i,8:9]<-NA
}
```

```r
ordered2<-EQ5D[order(EQ5D$Indice),] # Los ordenamos de menor a mayor índice

# Creamos unos nuevos valores de selección
set.seed(19)
rows3<-sample(1:22, 200, replace = T) # Los primero 22 son de fuera de la CV
columns3<-sample(3:7, 200,replace = T)

# Creamos el 10 % de MNAR

for(i in 1:25){
  ifelse(is.na(ordered2[rows3[i],columns3[i]]),
         ordered2[rows3[i+1],columns3[i]]<-NA,
         ordered2[rows3[i],columns3[i]]<-NA)
}

sum(is.na(ordered2))
```

```
## [1] 25
```

```r
MNAR_EQ5D10<-ordered2
for(i in which(!complete.cases(MNAR_EQ5D10))){
  MNAR_EQ5D10[i,8:9]<-NA
}
```

```r
# Creamos el 20 % de MNAR

for(i in 26:54){
  ifelse(is.na(ordered2[rows3[i],columns3[i]]),
         ordered2[rows3[i+1],columns3[i]]<-NA,
         ordered2[rows3[i],columns3[i]]<-NA)
}

sum(is.na(ordered2))
```

```
## [1] 50
```

```r
MNAR_EQ5D20<-ordered2
for(i in which(!complete.cases(MNAR_EQ5D20))){
  MNAR_EQ5D20[i,8:9]<-NA
}
```

```r
# Creamos el 30 % de MNAR

for(i in 55:90){
  ifelse(is.na(ordered2[rows3[i],columns3[i]]),
         ordered2[rows3[i+1],columns3[i]]<-NA,
         ordered2[rows3[i],columns3[i]]<-NA)
}

sum(is.na(ordered2))
```

```
## [1] 75
```

```r
MNAR_EQ5D30<-ordered2
for(i in which(!complete.cases(MNAR_EQ5D30))){
  MNAR_EQ5D30[i,8:9]<-NA
}
```

```
Descriptiva<-function(DB){
  cero<-
subset(DB,DB$Timepoint==0,c("Timepoint","Movilidad","Cuidado.Personal","Actividades.Diarias","Dolor",
nsiedad"))
  uno<-
subset(DB,DB$Timepoint==1,c("Timepoint","Movilidad","Cuidado.Personal","Actividades.Diarias","Dolor",
nsiedad"))

  # Vemos el porcentaje de movilidad de pacientes basales que no tenian problemas
  mov_basal<-summary(cero$Movilidad)[3]/(summary(cero$Movilidad)[1]+summary(cero$Movilidad)[2]+summa
ry(cero$Movilidad)[3])
  # Vemos el porcentaje de movilidad de pacientes finales que no tenian problemas
  mov_final<-summary(uno$Movilidad)[3]/(summary(uno$Movilidad)[1]+summary(uno$Movilidad)
[2]+summary(uno$Movilidad)[3])

  # Vemos el porcentaje de Cuidado.Personal de pacientes basales que no tenian problemas
  cp_basal<-summary(cero$Cuidado.Personal)[3]/(summary(cero$Cuidado.Personal)[1]+summary(cero$Cuidad
o.Personal)[2]+summary(cero$Cuidado.Personal)[3])
  # Vemos el porcentaje de Cuidado.Personal de pacientes finales que no tenian problemas
  cp_final<-summary(uno$Cuidado.Personal)[3]/(summary(uno$Cuidado.Personal)[1]+summary(uno$Cuidado.P
ersonal)[2]+summary(uno$Cuidado.Personal)[3])

  # Vemos el porcentaje de Actividades.Diarias de pacientes basales que no tenian problemas
  ad_basal<-summary(cero$Actividades.Diarias)[3]/(summary(cero$Actividades.Diarias)[1]+summary(cero
$Actividades.Diarias)[2]+summary(cero$Actividades.Diarias)[3])
  # Vemos el porcentaje de Actividades.Diarias de pacientes finales que no tenian problemas
  ad_final<-summary(uno$Actividades.Diarias)[3]/(summary(uno$Actividades.Diarias)[1]+summary(uno$Act
ividades.Diarias)[2]+summary(uno$Actividades.Diarias)[3])

  # Vemos el porcentaje de Dolor de pacientes basales que no tenian problemas
  dol_basal<-summary(cero$Dolor)[3]/(summary(cero$Dolor)[1]+summary(cero$Dolor)[2]+summary(cero$Dolo
r)[3])
  # Vemos el porcentaje de Dolor de pacientes finales que no tenian problemas
  dol_final<-summary(uno$Dolor)[3]/(summary(uno$Dolor)[1]+summary(uno$Dolor)[2]+summary(uno$Dolor)
[3])

  # Vemos el porcentaje de Ansiedad de pacientes basales que no tenian problemas
  ans_basal<-summary(cero$Ansiedad)[3]/(summary(cero$Ansiedad)[1]+summary(cero$Ansiedad)
[2]+summary(cero$Ansiedad)[3])
  # Vemos el porcentaje de Ansiedad de pacientes finales que no tenian problemas
  ans_final<-summary(uno$Ansiedad)[3]/(summary(uno$Ansiedad)[1]+summary(uno$Ansiedad)[2]+summary(uno
$Ansiedad)[3])
cat("El porcentaje de MOVILIDAD de pacientes basales que no tenian problemas es
de",round(mov_basal*100,0),"%\n")
cat("El porcentaje de MOVILIDAD de pacientes finales que no tenian problemas es
de",round(mov_final*100,0),"%\n")
# plot(DB$Movilidad~DB$Timepoint,ylab = "Movilidad",xlab="Timepoint")
cat("\n")
cat("El porcentaje de CUIDADO PERSONAL de pacientes basales que no tenian problemas es de",round(cp_
basal*100,0),"%\n")
cat("El porcentaje de CUIDADO PERSONAL de pacientes finales que no tenian problemas es de",round(cp_
final*100,0),"%\n")
# plot(DB$Cuidado.Personal~DB$Timepoint,ylab = "Cuidado Personal",xlab="Timepoint")
cat("\n")
cat("El porcentaje de ACTIVIDADES DIARIAS de pacientes basales que no tenian problemas es
de",round(ad_basal*100,0),"%\n")
cat("El porcentaje de ACTIVIDADES DIARIAS de pacientes finales que no tenian problemas es
de",round(ad_final*100,0),"%\n")
# plot(DB$Actividades.Diarias~DB$Timepoint,ylab = "Actividades Diarias",xlab="Timepoint")
cat("\n")
cat("El porcentaje de DOLOR de pacientes basales que no tenian problemas es
de",round(dol_basal*100,0),"%\n")
cat("El porcentaje de DOLOR de pacientes finales que no tenian problemas es
de",round(dol_final*100,0),"%\n")
# plot(DB$Dolor~DB$Timepoint,ylab = "Dolor",xlab="Timepoint")
cat("\n")
cat("El porcentaje de ANSIEDAD de pacientes basales que no tenian problemas es de",round(ans_basal*1
00,0),"%\n")
cat("El porcentaje de ANSIEDAD de pacientes finales que no tenian problemas es de",round(ans_final*1
00,0),"%\n")
# plot(DB$Ansiedad~DB$Timepoint,ylab = "Ansiedad",xlab="Timepoint")
}
```

```r
# Función para calcular las medias
Medias<-function(BD) {

  indice_m_G<-round(mean(BD$Indice,na.rm = T),3) # Índice medio global
  indice_sd_G<-round(sd(BD$Indice,na.rm = T),3) # Desviación media global
  indice_m_0<-round(mean(BD[BD$Timepoint==0,]$Indice,na.rm = T),3) # Índice medio basal
  indice_sd_0<-round(sd(BD[BD$Timepoint==0,]$Indice,na.rm = T),3) # Desviación media basal
  indice_m_1<-round(mean(BD[BD$Timepoint==1,]$Indice,na.rm = T),3) # Índice medio final
  indice_sd_1<-round(sd(BD[BD$Timepoint==1,]$Indice,na.rm = T),3) # Desviación media final

  cat("Índice medio global es ", indice_m_G,"(",indice_sd_G,")","n =",sum(complete.cases(BD)),"\n")
  cat("Índice medio basal es ", indice_m_0,"(",indice_sd_0,")","n =",sum(complete.cases(BD[BD$Timepo
int==0,])),"\n")
  cat("Índice medio final es ", indice_m_1,"(",indice_sd_1,")","n =",sum(complete.cases(BD[BD$Timepo
int==1,])),"\n")
}
```

```r
# Creamos una función para eliminar los registros con datos faltantes

listwiseDB<-function(DB){
  completo<-DB[which(complete.cases(DB)),]
  Descriptiva(completo)
}
```

```r
# Función para comparar las medias

ComparaMedias<-function(BD) {

  print(t.test(Indice~Timepoint, data = BD,paired=T))
  print(wilcox.test(Indice~Timepoint, data = BD,paired=T))
}
```

```r
# Función para la imputación de datos

library(mice)
library(VIM) # Para las gráficas de distribución NAs

# Con esta función, podemos comprobar los NAs que tenemos en nuestra Base de Datos
NA_resume<-function(BD){
  print(md.pattern(BD[3:7]))
  print(sapply(BD[,3:7], function(x) sum(is.na(x))))
  aggr_plot <- aggr(BD[,3:7], col=c('navyblue','red'), numbers=TRUE, sortVars=TRUE, labels=names(dat
a), cex.axis=.7, gap=3, ylab=c("Histogram of missing data","Pattern"))
  plot(aggr_plot)
}

# Comprobamos así que nuestros NAs son correctos

NA_resume(EQ5D)
```

```r
imp<-mice(MCAR_EQ5D10[3:7],seed = 17)
```

Exportamos las imputaciones para poder usar el algoritmo

```
ExportDB(MCAR_EQ5D10_i,sheetName = "MCAR_EQ5D10_i")
ExportDB(MCAR_EQ5D20_i,sheetName = "MCAR_EQ5D20_i")
ExportDB(MCAR_EQ5D30_i,sheetName = "MCAR_EQ5D30_i")

ExportDB(MAR_EQ5D10_i,sheetName = "MAR_EQ5D10_i")
ExportDB(MAR_EQ5D20_i,sheetName = "MAR_EQ5D20_i")
ExportDB(MAR_EQ5D30_i,sheetName = "MAR_EQ5D30_i")

ExportDB(MNAR_EQ5D10_i,sheetName = "MNAR_EQ5D10_i")
ExportDB(MNAR_EQ5D20_i,sheetName = "MNAR_EQ5D20_i")
ExportDB(MNAR_EQ5D30_i,sheetName = "MNAR_EQ5D30_i")
```

Vamos a comparar ahora las bases de datos imputadas. Para ello, primero las vamos a importar

```
# Importamos las BDs imputadas

library(xlsx)

MCAR_EQ5D10_new<- read.xlsx("Respuestas_i.xlsx", sheetName= "MCAR_EQ5D10",header = T)
MCAR_EQ5D10_new$Movilidad<-ordered(MCAR_EQ5D10_new$Movilidad,levels=c("3","2","1")) #Definimos como
 variable ordenada con el orden deseado (por defecto R toma orden alfabetico)
MCAR_EQ5D10_new$Cuidado.Personal<-ordered(MCAR_EQ5D10_new$Cuidado.Personal,levels=c("3","2","1"))
MCAR_EQ5D10_new$Actividades.Diarias<-
ordered(MCAR_EQ5D10_new$Actividades.Diarias,levels=c("3","2","1"))
MCAR_EQ5D10_new$Dolor<-ordered(MCAR_EQ5D10_new$Dolor,levels=c("3","2","1"))
MCAR_EQ5D10_new$Ansiedad<-ordered(MCAR_EQ5D10_new$Ansiedad,levels=c("3","2","1"))
MCAR_EQ5D10_new$Perfil<-as.factor(MCAR_EQ5D10_new$Perfil)
str(MCAR_EQ5D10_new)
```

```
## 'data.frame':    50 obs. of  10 variables:
##  $ Paciente           : Factor w/ 25 levels "1","10","11",..: 1 1 12 12 19 19 20 20 21 21 ...
##  $ Timepoint          : Factor w/ 2 levels "0","1": 1 2 1 2 1 2 1 2 1 2 ...
##  $ Movilidad          : Ord.factor w/ 3 levels "3"<"2"<"1": 2 2 1 2 2 3 3 1 2 3 ...
##  $ Cuidado.Personal   : Ord.factor w/ 3 levels "3"<"2"<"1": 1 3 1 3 2 2 2 3 3 3 ...
##  $ Actividades.Diarias: Ord.factor w/ 3 levels "3"<"2"<"1": 1 1 3 1 3 2 3 1 2 2 ...
##  $ Dolor              : Ord.factor w/ 3 levels "3"<"2"<"1": 3 3 1 2 1 2 3 2 1 3 ...
##  $ Ansiedad           : Ord.factor w/ 3 levels "3"<"2"<"1": 2 1 2 1 1 2 3 1 2 3 ...
##  $ Perfil             : Factor w/ 45 levels "11211","11212",..: 32 20 43 22 26 10 6 35 19 1 ...
##  $ Indice             : num  0.635 0.756 0.598 0.685 0.626 0.687 0.868 0.666 0.694 0.924 ...
##  $ estudios           : Factor w/ 2 levels "0","1": 1 1 1 1 2 2 1 1 1 1 ...
```

10. Material Complementario

- **EQ-5D-5L Crosswalk Index Value Calculator MAC**: Herramienta en Excel que se ha utilizado para calcular la variable *índice* y *perfil* de nuestra base de datos.

- **Resumen Imputación**: Resumen de los datos imputados

- **Resumen Resultados:** Resumen de los resultados del TFM

- **Código R:** Carpeta donde se adjunta el código utilizado para el TFM y la interpretación, junto con lo necesario para reproducirlo de nuevo.

- **Interpretación:** Archivo html con el código utilizado ejecutado del archivo Interpretación.Rmd.

- **TFM:** Archivo html con el código utilizado ejecutado del archivo TFM.Rmd.